Section 6: EM Algorithm

Ruofan Ma

Gov2018 2024 Spring

March 2, 2024

Motivation: What happens when ML estimation is hard?

- Sometimes log-likelihood difficult to maximize directly numerically

Motivation: What happens when ML estimation is hard?

- Sometimes log-likelihood difficult to maximize directly numerically
- MLE assumes we have complete dataset (variables all present), from which we can think about selecting parameters that result in the best fit of the joint probability of the data

Motivation: What happens when ML estimation is hard?

- Sometimes log-likelihood difficult to maximize directly numerically
- MLE assumes we have complete dataset (variables all present), from which we can think about selecting parameters that result in the best fit of the joint probability of the data
- Sometimes we have missing data!

Motivation: What happens when ML estimation is hard?

- Sometimes log-likelihood difficult to maximize directly numerically
- MLE assumes we have complete dataset (variables all present), from which we can think about selecting parameters that result in the best fit of the joint probability of the data
- Sometimes we have missing data! Sometimes we have latent variables!

Motivation: What happens when ML estimation is hard?

- Sometimes log-likelihood difficult to maximize directly numerically
- MLE assumes we have complete dataset (variables all present), from which we can think about selecting parameters that result in the best fit of the joint probability of the data
- Sometimes we have missing data! Sometimes we have latent variables! Sometimes we have clustering!

Motivation: What happens when ML estimation is hard?

- Sometimes log-likelihood difficult to maximize directly numerically
- MLE assumes we have complete dataset (variables all present), from which we can think about selecting parameters that result in the best fit of the joint probability of the data
- Sometimes we have missing data! Sometimes we have latent variables! Sometimes we have clustering!
- EM algorithm is a procedure for algorithm construction, not a specific algorithm. Each problem is different, only the structure of the Expectation and Maximization steps are common

Motivation: What happens when ML estimation is hard?

- Sometimes log-likelihood difficult to maximize directly numerically
- MLE assumes we have complete dataset (variables all present), from which we can think about selecting parameters that result in the best fit of the joint probability of the data
- Sometimes we have missing data! Sometimes we have latent variables! Sometimes we have clustering!
- EM algorithm is a procedure for algorithm construction, not a specific algorithm. Each problem is different, only the structure of the Expectation and Maximization steps are common
- How exactly they are programmed is problem-dependent.

Motivation: What happens when ML estimation is hard?

- Sometimes log-likelihood difficult to maximize directly numerically
- MLE assumes we have complete dataset (variables all present), from which we can think about selecting parameters that result in the best fit of the joint probability of the data
- Sometimes we have missing data! Sometimes we have latent variables! Sometimes we have clustering!
- EM algorithm is a procedure for algorithm construction, not a specific algorithm. Each problem is different, only the structure of the Expectation and Maximization steps are common
- How exactly they are programmed is problem-dependent.
- Widely applied in machine learning for density estimation and clustering

Roadmap for today

- Big picture of EM
  - What is EM? Why is it called EM? When do we want to use EM? What are the properties of EM?
- How to set up an EM algorithm
  - E-step: inferring the missing values given the parameters
  - M-step: optimizing the parameters given the "filled in" data
- Lab exercise (Rmd available on course website)

Big Picture

Setup

- Data X observed, set of (possibly made up) latent variables Z, model parameters θ

Big Picture

Setup

- Data X observed, set of (possibly made up) latent variables Z, model parameters θ
- Goal of **EM** is to find a maximization to the likelihood function $p(X \mid \theta)$ with respect to parameter θ when this expression or its log cannot be discovered by typical MLE methods

Big Picture

Setup

- Data X observed, set of (possibly made up) latent variables Z, model parameters θ
- Goal of **EM** is to find a maximization to the likelihood function $p(X \mid \theta)$ with respect to parameter θ when this expression or its log cannot be discovered by typical MLE methods
- Suppose for each observation $x_i \in X$, we get a corresponding value $z_i \in Z$.

Big Picture

Setup

- Data X observed, set of (possibly made up) latent variables Z, model parameters θ
- Goal of **EM** is to find a maximization to the likelihood function $p(X \mid \theta)$ with respect to parameter θ when this expression or its log cannot be discovered by typical MLE methods
- Suppose for each observation $x_i \in X$, we get a corresponding value $z_i \in Z$.
- $\{X, Z\}$ is called the complete dataset and the likelihood of the complete set is $p(X, Z \mid \theta)$. However, we don't know the complete set, only our observed X. To proceed we need to construct the posterior $p(Z \mid X, \theta)$.

Big Picture

Setup

- If we have $p(Z \mid X, \theta)$ we can compute the likelihood for the complete set:

$$p(X, Z \mid \theta) = p(Z \mid X, \theta)p(X \mid \theta)$$

Big Picture

Setup

- If we have $p(Z \mid X, \theta)$ we can compute the likelihood for the complete set:

$$p(X, Z \mid \theta) = p(Z \mid X, \theta)p(X \mid \theta)$$

- Assume now we also know an estimate θ_i for θ - this allows us to compute the posterior $p(Z \mid X, \theta_i)$

Big Picture

Setup

- If we have $p(Z \mid X, \theta)$ we can compute the likelihood for the complete set:
$$p(X, Z \mid \theta) = p(Z \mid X, \theta) p(X \mid \theta)$$
- Assume now we also know an estimate θ_i for θ - this allows us to compute the posterior $p(Z \mid X, \theta_i)$
- Log-likelihood: $\log p(X \mid \theta) = \log \{\sum_z p(X, Z = z \mid \theta)\} = \log \{\sum_z p(X \mid Z = z, \theta) \cdot p(Z = z \mid \theta)\}$
- EM algorithm will maximize $\log p(X \mid \theta)$ but since log is strict monotonous function, it will also maximize $p(X \mid \theta)$

Usage of EM in Social Science

Examples

- Image segmentation Carson et al. 2002 IEEE
- Latent data models Cappe & Moulines 2009 JRSS
- Missing data Honaker et al. 2011 JSS
- Political attention in text Quinn et al. 2010 AJPS
- Mixture models and network data Newman & Leicht 2007 PNAS

What/How to Optimize?

Goal: Optimize $\ell(\boldsymbol{\theta})$ when there is missing data/latent variable

- Consider an arbitrary distribution $q(\mathbf{Z})$ over the hidden variables. Then the observed data log-likelihood function can be written as:

What/How to Optimize?

Goal: Optimize $\ell(\boldsymbol{\theta})$ when there is missing data/latent variable

- Consider an arbitrary distribution $q(\mathbf{Z})$ over the hidden variables. Then the observed data log-likelihood function can be written as:

$$\ell(\mathbf{X}|\boldsymbol{\theta}) := \log \left[\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) \right] = \log \left[\sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q(\mathbf{Z})} \right]$$

What/How to Optimize?

Goal: Optimize $\ell(\boldsymbol{\theta})$ when there is missing data/latent variable

- Consider an arbitrary distribution $q(\mathbf{Z})$ over the hidden variables. Then the observed data log-likelihood function can be written as:

$$\ell(\mathbf{X}|\boldsymbol{\theta}) := \log \left[\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) \right] = \log \left[\sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q(\mathbf{Z})} \right]$$

- Notice that $\log(\cdot)$ is a concave function. Therefore, by Jensen's inequality, $\log \sum_i a_i x_i \geq \sum_i a_i \log x_i$:

$$\ell(\mathbf{X}|\boldsymbol{\theta}) \geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[\frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q(\mathbf{Z})} \right]$$

What/How to Optimize?

Goal: Optimize $\ell(\boldsymbol{\theta})$ when there is missing data/latent variable

- Consider an arbitrary distribution $q(\mathbf{Z})$ over the hidden variables. Then the observed data log-likelihood function can be written as:

$$\ell(\mathbf{X}|\boldsymbol{\theta}) := \log \left[\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) \right] = \log \left[\sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q(\mathbf{Z})} \right]$$

- Notice that $\log(\cdot)$ is a concave function. Therefore, by Jensen's inequality, $\log \sum_i a_i x_i \geq \sum_i a_i \log x_i$:

$$\ell(\mathbf{X}|\boldsymbol{\theta}) \geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[\frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q(\mathbf{Z})} \right]$$

- Let's denote this lower bound on the RHS as $Q(\boldsymbol{\theta}, q) \Rightarrow$ The larger the Q, the larger the ℓ

Decomposing the Target Function

Goal: Pick the q that yields the tightest lower bound

- We can further manipulate Q:

$$Q(\boldsymbol{\theta}, q) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[\frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q(\mathbf{Z})} \right]$$

Decomposing the Target Function

Goal: Pick the q that yields the tightest lower bound

- We can further manipulate Q:

$$
\begin{aligned}
Q(\boldsymbol{\theta}, q) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[\frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q(\mathbf{Z})} \right] \\
&= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[\frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) p(\mathbf{X}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right]
\end{aligned}
$$

Decomposing the Target Function

Goal: Pick the q that yields the tightest lower bound

- We can further manipulate Q:

$$
\begin{aligned}
Q(\boldsymbol{\theta}, q) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[\frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q(\mathbf{Z})} \right] \\
&= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[\frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) p(\mathbf{X}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right] \\
&= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[\frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right] + \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}|\boldsymbol{\theta})
\end{aligned}
$$

Decomposing the Target Function

Goal: Pick the q that yields the tightest lower bound

- We can further manipulate Q:

$$
\begin{aligned}
Q(\boldsymbol{\theta}, q) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[\frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q(\mathbf{Z})} \right] \\
&= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[\frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) p(\mathbf{X}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right] \\
&= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[\frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right] + \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}|\boldsymbol{\theta}) \\
&= -\mathbb{KL}(q(\mathbf{Z}) || p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})) + \log p(\mathbf{X}|\boldsymbol{\theta})
\end{aligned}
$$

Decomposing the Target Function

Goal: Pick the q that yields the tightest lower bound

- We can further manipulate Q:

$$
\begin{aligned}
Q(\boldsymbol{\theta}, q) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[\frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q(\mathbf{Z})} \right] \\
&= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[\frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) p(\mathbf{X}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right] \\
&= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[\frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right] + \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}|\boldsymbol{\theta}) \\
&= -\mathbb{KL}(q(\mathbf{Z}) || p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})) + \log p(\mathbf{X}|\boldsymbol{\theta})
\end{aligned}
$$

Decomposing the Target Function

Goal: Pick the q that yields the tightest lower bound

- We can further manipulate Q:

$$
\begin{aligned}
Q(\boldsymbol{\theta}, q) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[\frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q(\mathbf{Z})} \right] \\
&= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[\frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) p(\mathbf{X}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right] \\
&= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[\frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right] + \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}|\boldsymbol{\theta}) \\
&= -\mathbb{KL}(q(\mathbf{Z}) || p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})) + \log p(\mathbf{X}|\boldsymbol{\theta})
\end{aligned}
$$

- Since $\mathbb{KL}(\cdot) \geq 0$, and $\log p(\mathbf{X}|\boldsymbol{\theta})$ does not depend on q, Q can be maximized by setting $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$

E-Step: Update the target function/missing value

- Key Insight: Expected complete data log-likelihood is a lower bound!

E-Step: Update the target function/missing value

- Key Insight: Expected complete data log-likelihood is a lower bound!

$$Q(\theta, q) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[\frac{p(\mathbf{X}, \mathbf{Z} \mid \theta)}{q(\mathbf{Z})} \right]$$

E-Step: Update the target function/missing value

- Key Insight: Expected complete data log-likelihood is a lower bound!

$$
\begin{aligned}
Q(\boldsymbol{\theta}, q) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[\frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q(\mathbf{Z})} \right] \\
&= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) - \sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})
\end{aligned}
$$

E-Step: Update the target function/missing value

- Key Insight: Expected complete data log-likelihood is a lower bound!

$$
\begin{aligned}
Q(\boldsymbol{\theta}, q) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[\frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q(\mathbf{Z})} \right] \\
&= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) - \sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z}) \\
&= \underbrace{\mathbb{E}_q \left[\log p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) \right]}_{\text{expected complete data log-likelihood}} + \underbrace{\mathbb{H}(q)}_{\text{nuisance function w.r.t. } \theta}
\end{aligned}
$$

E-Step: Update the target function/missing value

- Key Insight: Expected complete data log-likelihood is a lower bound!

$$
\begin{aligned}
Q(\boldsymbol{\theta}, q) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[\frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q(\mathbf{Z})} \right] \\
&= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) - \sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z}) \\
&= \underbrace{\mathbb{E}_q \left[\log p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) \right]}_{\text{expected complete data log-likelihood}} + \underbrace{\mathbb{H}(q)}_{\text{nuisance function w.r.t. } \theta}
\end{aligned}
$$

E-Step: Update the target function/missing value

- Key Insight: Expected complete data log-likelihood is a lower bound!

$$
\begin{aligned}
Q(\theta, q) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left[\frac{p(\mathbf{X}, \mathbf{Z} \mid \theta)}{q(\mathbf{Z})} \right] \\
&= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z} \mid \theta) - \sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z}) \\
&= \underbrace{\mathbb{E}_q \left[\log p(\mathbf{X}, \mathbf{Z} \mid \theta) \right]}_{\text{expected complete data log-likelihood}} + \underbrace{\mathbb{H}(q)}_{\text{nuisance function w.r.t. } \theta}
\end{aligned}
$$

- Plugging in the condition $q^t(\mathbf{Z}) = p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^t)$, where $\boldsymbol{\theta}^t$ is our estimate of the parameters at iteration t:

$$
Q(\boldsymbol{\theta}, q^t) = \mathbb{E}_{q^t} \left[\log p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta}) \right] + \text{const.}
$$

M-Step: Update the parameter

- In M-step, we obtain an update of the parameter $\boldsymbol{\theta}$ by solving the following optimization problem:

M-Step: Update the parameter

- In M-step, we obtain an update of the parameter $\boldsymbol{\theta}$ by solving the following optimization problem:

$$\boldsymbol{\theta}^{t+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t) = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_q \left[\log p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})\right]$$

M-Step: Update the parameter

- In M-step, we obtain an update of the parameter $\boldsymbol{\theta}$ by solving the following optimization problem:

$$\boldsymbol{\theta}^{t+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t) = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_q \left[\log p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})\right]$$

- We then feed the value of $\boldsymbol{\theta}^{t+1}$ to obtain an update for the target function, thereby iterating between E-step and M-step until a stopping criterion is met.

M-Step: Update the parameter

- In M-step, we obtain an update of the parameter $\boldsymbol{\theta}$ by solving the following optimization problem:

$$\boldsymbol{\theta}^{t+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t) = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_q \left[\log p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})\right]$$

- We then feed the value of $\boldsymbol{\theta}^{t+1}$ to obtain an update for the target function, thereby iterating between E-step and M-step until a stopping criterion is met.
- The dark magic of EM: There is a theoretical guarantee that the EM algorithm monotonically increases the log-likelihood of the observed data.

EM monotonically increases the observed data log-likelihood

EM monotonically increases the observed data log-likelihood

Lemma (E-Step produces tight lower bound)

By setting $q^t(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^t), \mathbb{KL}(q||p) = 0$, hence:

$$Q(\boldsymbol{\theta}^t, \boldsymbol{\theta}^t) = \log p(\mathbf{X} \mid \boldsymbol{\theta}^t) = \ell(\boldsymbol{\theta}^t)$$

EM monotonically increases the observed data log-likelihood

Lemma (E-Step produces tight lower bound)

By setting $q^t(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^t), \mathbb{KL}(q||p) = 0$, hence:

$$Q(\boldsymbol{\theta}^t, \boldsymbol{\theta}^t) = \log p(\mathbf{X} \mid \boldsymbol{\theta}^t) = \ell(\boldsymbol{\theta}^t)$$

Theorem (Monotonicity of EM)

$$\ell(\boldsymbol{\theta}^{t+1}) \geq Q(\boldsymbol{\theta}^{t+1}, \boldsymbol{\theta}^t) \geq Q(\boldsymbol{\theta}^t, \boldsymbol{\theta}^t) = \ell(\boldsymbol{\theta}^t)$$

, where the first inequality follows since $Q(\boldsymbol{\theta}^t, \cdot)$ is a lower bound for $\ell(\boldsymbol{\theta}^t)$; the second inequality follows from the M-step: $Q(\boldsymbol{\theta}^{t+1}, \boldsymbol{\theta}^t) = \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t) \geq Q(\boldsymbol{\theta}^t, \boldsymbol{\theta}^t)$; and the third equality follows from the above lemma.

Visualization of EM

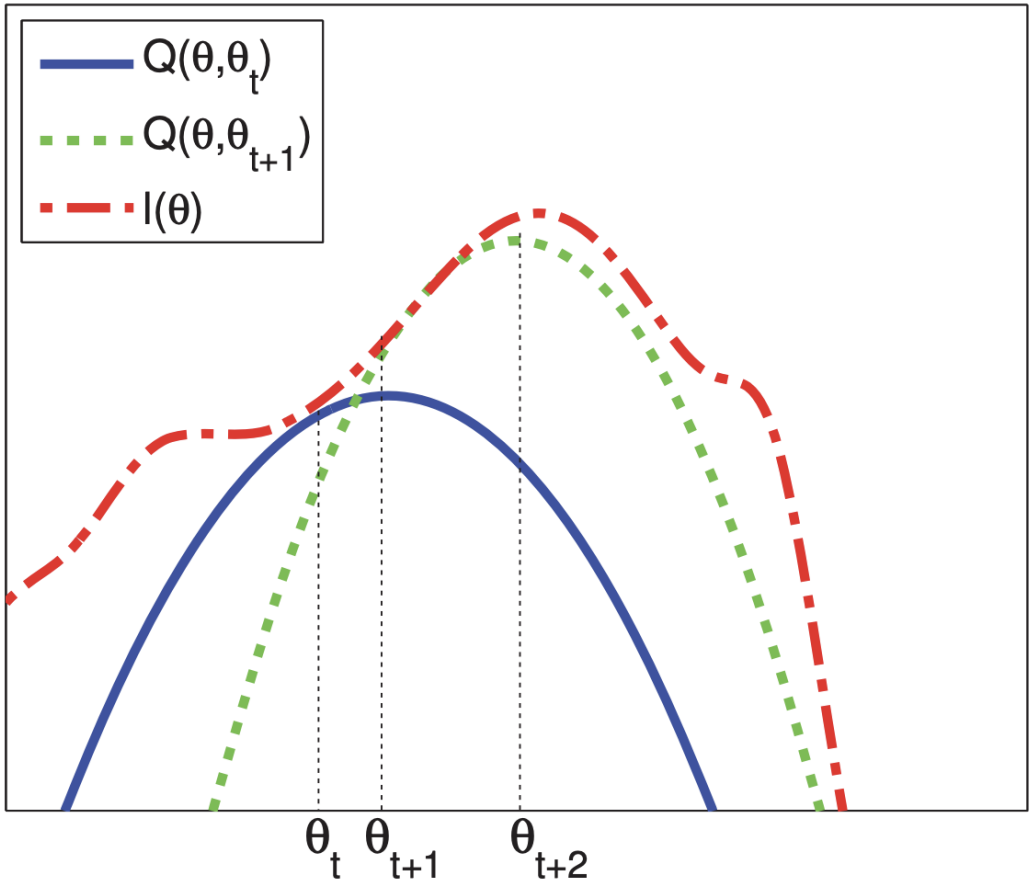

Figure: Source: Murphy, p.367