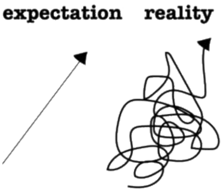
Expectation Maximization Algorithm

Sergej Dogadov
s.dogadov@tu-berlin.de
Technische Universität Berlin - Machine Learning Group
Machine Learning 2 Tutorial SoSe 2019

The EM Algorithm in General



Theory

The *expectation maximization* algorithm (EM algorithm) is a general technique for finding maximum likelihood solutions for probabilistic models having latent variables.

- Consider a probabilistic model with all observed variables $\boldsymbol{X}$ and the hidden variables $\boldsymbol{Z}$.
- The joint distribution $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ is parameterized by the model parameters $\boldsymbol{\theta}$.

Goal: to maximize the log-likelihood which can be expressed as:

$$\ln p(\boldsymbol{X}|\theta) = \ln \int_Z p(\boldsymbol{X}, \boldsymbol{Z}|\theta) d\boldsymbol{Z} = \ln \int_Z p(\boldsymbol{Z}|\boldsymbol{X}, \theta) p(\boldsymbol{X}|\theta) d\boldsymbol{Z}$$

Theory

Multiply and divide by an arbitrary distribution over $\boldsymbol{Z}$ called $q(\boldsymbol{Z})$.

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln \int_{\boldsymbol{Z}} \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})q(\boldsymbol{Z})}{q(\boldsymbol{Z})} d\boldsymbol{Z} = \ln \mathbb{E}_{z\sim q}\left[\frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})}\right]$$

By Jensen's inequality:

$$\ln \mathbb{E}_{z\sim q}\left[\frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})}\right] \geq \mathbb{E}_{z\sim q}\left[\ln \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})}\right] = \underbrace{\mathcal{L}(q,\boldsymbol{\theta})}_{\text{lower bound}}$$

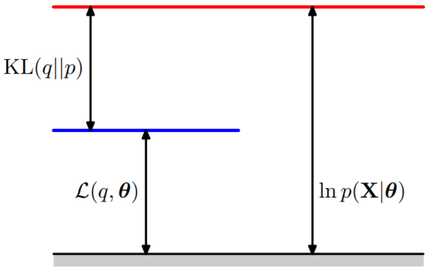
The lower bound can be expressed with the log-likelihood term as:

$$\mathcal{L}(q,\boldsymbol{\theta}) = \mathbb{E}_{z\sim q}\left[\ln \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})p(\boldsymbol{X}|\boldsymbol{\theta})}{q(\boldsymbol{Z})}\right] = -\mathbb{D}_{KL}(q||p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}) + \ln p(\boldsymbol{X}|\boldsymbol{\theta})$$

$$\boxed{\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{L}(q,\boldsymbol{\theta}) + \mathbb{D}_{KL}(q||p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}$$

Lower bound on the log likelihood function $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$

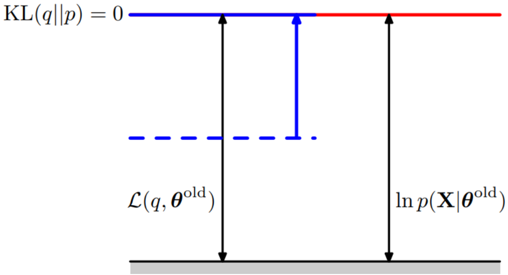
$$\boxed{\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathbb{D}_{KL}(q||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})}$$



For any arbitrary distribution $q(\boldsymbol{Z})$ the KL divergence $\mathbb{D}_{KL}(q||p) \geq 0$.

E-step

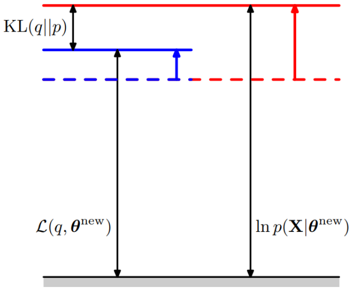
The EM algorithm is a two-stage iterative optimization technique for finding maximum likelihood solutions.



E-step: the lower bound $\mathcal{L}(q, \boldsymbol{\theta}^{old})$ is maximized w.r.t. $q(\boldsymbol{Z})$.
It's reached when KL term vanishes, in other words when $q(\boldsymbol{Z})$ is equal the posterior distribution $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\boldsymbol{old}})$ for the fixed model parameters $\boldsymbol{\theta}^{old}$.

M-step

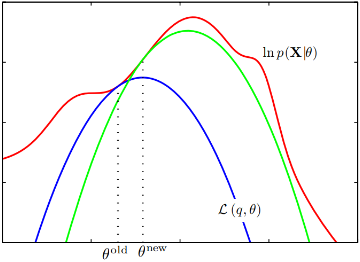
In the M-step the distribution $q(\boldsymbol{Z})$ is fixed and the lower bound $\mathcal{L}(\boldsymbol{Z}, \boldsymbol{\theta})$ is optimized w.r.t. $\boldsymbol{\theta}$ to give some new value $\boldsymbol{\theta}^{new}$.



Because the KL divergence is non negative, the log likelihood $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ increases at least as much as the lower bound does.

EM algorithm in the parameter space

The red curve is the log likelihood function whose value we wish to maximize. In the first **E-step** we evaluate the posterior distribution over latent variables. The blue line is a lower bound $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$ equals the log likelihood at $\boldsymbol{\theta}^{old}$.



In the **M-step**, the bound is maximized giving the value $\boldsymbol{\theta}^{new}$, which gives a larger value of log likelihood than $\boldsymbol{\theta}^{old}$. The subsequent E step then constructs a bound that is tangential at $\boldsymbol{\theta}^{new}$ as shown by the green curve.

The EM Algorithm overview

Given a joint distribution $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ over observed variables $\boldsymbol{X}$ and latent variables $\boldsymbol{Z}$, governed by model parameters $\boldsymbol{\theta}$.

Goal: to maximize the joint likelihood function $p(\boldsymbol{X}|\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$

1. Choose an initial values for model parameters $\boldsymbol{\theta}_{old}$
2. **E-step**: Evaluate posterior $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}_{old})$
3. **M-step**: Evaluate $\boldsymbol{\theta}_{new}$ given by

$$\boldsymbol{\theta}_{new} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}_{old}), \text{ where } \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}_{old}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}_{old}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$$

4. Iterate till convergence either of the log likelihood or the parameter values by setting $\boldsymbol{\theta}_{old} \leftarrow \boldsymbol{\theta}_{old}$