Bayesian Methods

Paolo Favaro

Contents

- Bayesian Decision Theory
- Expectation-Maximization
- Majorization-Minimization

Task

- Observe an X-ray image of a patient and decide whether the patient has a tumor or not

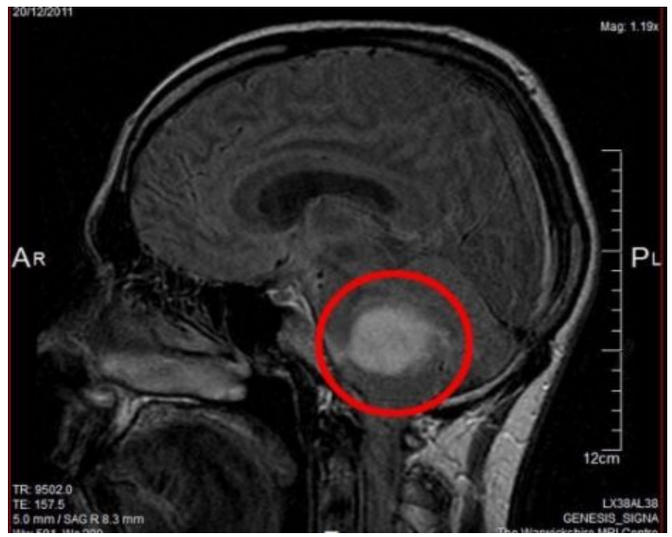
- Input/data = image
- Output/target = yes/no

Task

- Observe an X-ray image of a patient and decide whether the patient has a tumor or not

- Input/data = image

- Output/target = yes/no

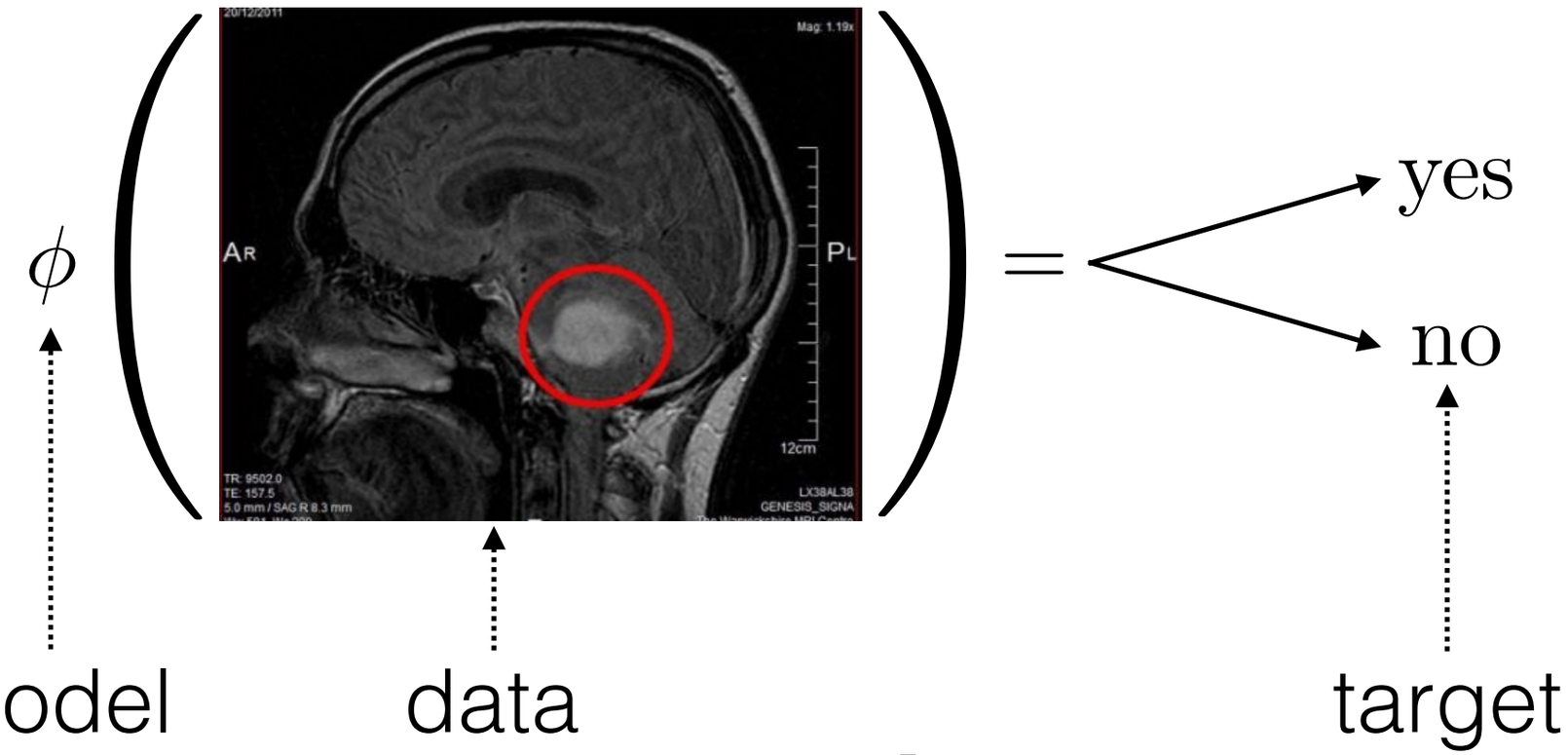


- **How do we pose this as a numerical problem?**

Solving a task

- **Define a model to do the task**

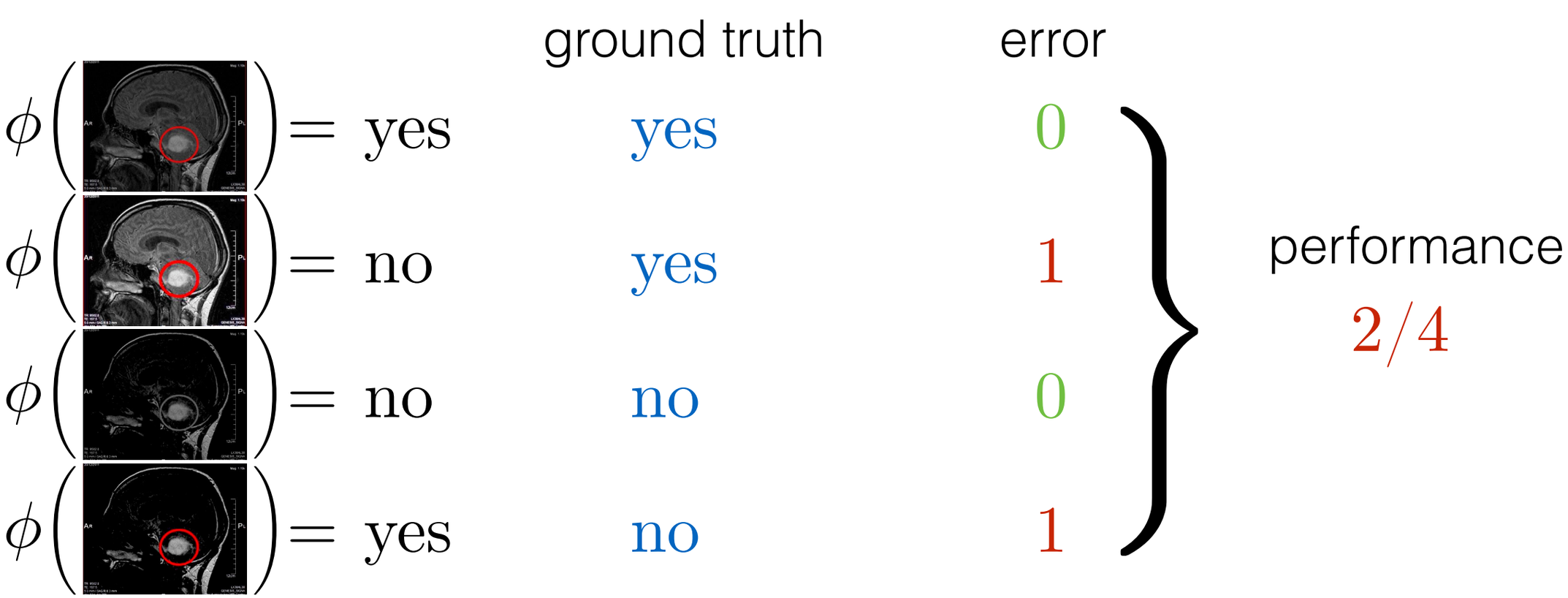
- The model is a function that maps given inputs to desired outputs



Solving a task

- **Measure how well the model works on the task**

- Count the mistakes or how close we are to the desired output (**performance**)



Basic notation

- Suppose that we have an observation vector $x \in \mathcal{X}$ together with a target vector $y \in \mathcal{Y}$

- Our goal is to predict y given x

- The space $\mathcal{Y}$ where y lives is continuous for a **regression** problem and discrete for a **classification** problem

- The joint probability $p(x, y)$ captures all the knowledge about x and y

Decision rule

- Given m observations $x_1, \ldots, x_m$
- Obtain an estimate ϕ for each $y_1, \ldots, y_m$ that best describes them

- ϕ is a **decision rule** (the model) and maps x to $\phi(x)$

Decision rule

- Examples of decision rules for **classification**

$$\phi(x) = \begin{cases} 1 & \text{if } w^\top x + b > 0 \\ 0 & \text{if } w^\top x + b \leq 0 \end{cases}$$ hyperplane

$$\phi(x) = \frac{1}{1 + e^{-(w^\top x + b)}}$$ logistic

$$\phi(x) = \begin{bmatrix} \frac{e^{a_1 x_1}}{\sum_{i=1}^{n} e^{a_i x_i}} \\ \frac{e^{a_2 x_2}}{\sum_{i=1}^{n} e^{a_i x_i}} \\ \cdots \\ \frac{e^{a_n x_n}}{\sum_{i=1}^{n} e^{a_i x_i}} \end{bmatrix}$$ softmax

Decision rule

- Examples of decision rules for **regression**

$$\phi(x) = w^\top x + b$$ hyperplane

$$\phi(x) = \sum_{i=0}^{n} w_i x^i$$ polynomial

$$\phi(x) = \sum_{i=1}^{n} w_i e^{-\frac{|w_i^\top x + b_i|^2}{\tau_i^2}}$$ radial basis function (RBF)

Loss function

- To choose the decision rule, we define a **loss function** L, which is a measure of how well ϕ describes the target variables

- L is a function of y and ϕ and defines their similarity

- Examples

  - $L(y, \phi, x) = |y - \phi(x)|^2$ **quadratic loss**

  - $L(y, \phi, x) = \mathbf{1}\{y \neq \phi(x)\}$ **0-1 loss**

Bayes risk

- **Bayes risk** is a measure of the performance across the whole distribution of observed and target variables of a decision rule given a certain loss function

$$E_{X,Y}[L(y, \phi, x)] = \int L(y, \phi, x) p(x, y) dx dy$$

Bayes risk

- **Bayes risk** is a measure of the performance across the whole distribution of observed and target variables of a decision rule given a certain loss function

$$
\begin{aligned}
E_{X,Y}[L(y,\phi,x)] &= \int L(y,\phi,x)p(x,y)dxdy \\
&= \int L(y,\phi,x)p(y|x)p(x)dxdy \\
&= E_X[E_{Y|X}[L(y,\phi,x)]]
\end{aligned}
$$

Bayes risk

- We define the optimal decision rule by solving

$$\hat{\phi} = \arg\min_{\phi} E_X[E_{Y|X}[L(y, \phi, x)]]$$

- Thus we can solve the problem element-wise via

$$\hat{\phi}(x) = \arg\min_{\phi(x)} E_{Y|X}[L(y, \phi(x), x)]$$

- The **posterior expected loss** is

$$E_{Y|X}[L(y, \phi, x)] = \int L(y, \phi, x) p(y|x) dx$$

Example #1

- Quadratic loss function

$$L(y, \phi, x) = |y - \phi(x)|^2$$

- Bayes risk minimization yields

$$\hat{\phi} = \arg\min_{\phi} \int |y - \phi(x)|^2 p(x, y) dx dy$$

Example #1

- Compute derivatives with respect to ϕ and set to 0

$$2\int(\phi(x)-y)p(x,y)dy=0$$

we separate the two terms

$$\phi(x)\int p(x,y)dy=\int yp(x,y)dy$$

and use marginalization

$$\phi(x)p(x)=\int yp(x,y)dy$$

Example #1

- We finally obtain the **conditional mean**

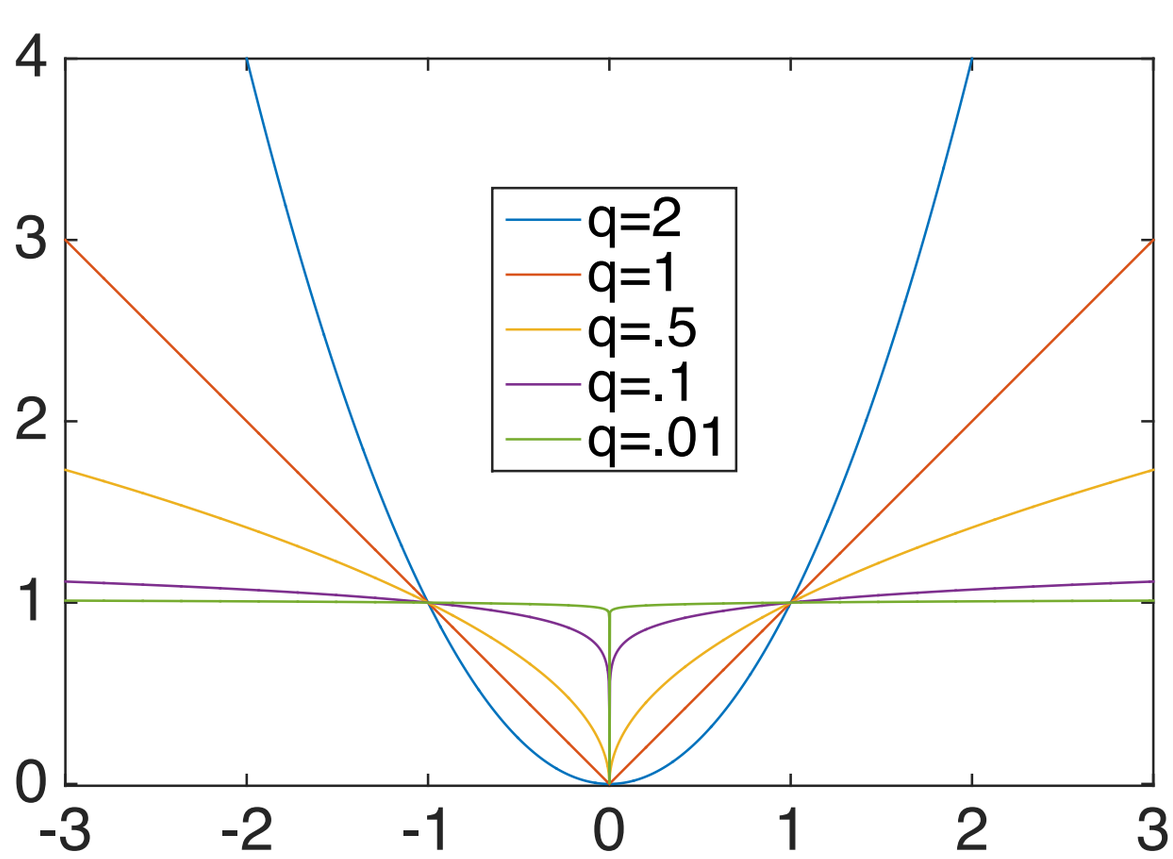
$$\phi(x) = \int yp(y|x)dy = E_{Y|X}[y]$$

and Bayes risk becomes

$$E_X[E_{Y|X}[|Y - \phi(X)|^2]] = E_X[E_{Y|X}[|Y - E_{Y|X}[y]|^2]]$$
$$= E_X[\mathrm{var}(Y|X)]$$

Example #2

- Consider Minkowski's loss $L_q(y, \phi, x) = |y - \phi(x)|^q$



Example #2

- Consider Minkowski's loss

$$L_q(y, \phi, x) = |y - \phi(x)|^q$$

- Let $q=1$, then Bayes risk minimization gives

$$\hat{\phi} = \arg\min_{\phi} \int |y - \phi(x)| p(x, y) dx dy$$

Example #2

- Let us rewrite Bayes risk in a simpler form

$$E_{X,Y}[L_1(Y,\phi,X)] = \int |y-\phi(x)|p(x,y)dxdy$$

$$= \int \left(\int |y-\phi(x)|p(y|x)dy \right) p(x)dx$$

$$= \int \left(\int_{y|y\succ\phi(x)} (y-\phi(x))p(y|x)dy + \int_{y|y\prec\phi(x)} (\phi(x)-y)p(y|x)dy \right) p(x)dx$$

Example #2

- Take derivatives with respect to ϕ and set to 0

$$\frac{\delta E_{X,Y}[L_1(Y, \phi, X)]}{\delta \phi} = 0$$

Example #2

- Take derivatives with respect to ϕ and set to 0

$$\left(\int_{y|y \succ \phi(x)} p(y|x)dy - \int_{y|y \prec \phi(x)} p(y|x)dy \right) p(x) = 0$$

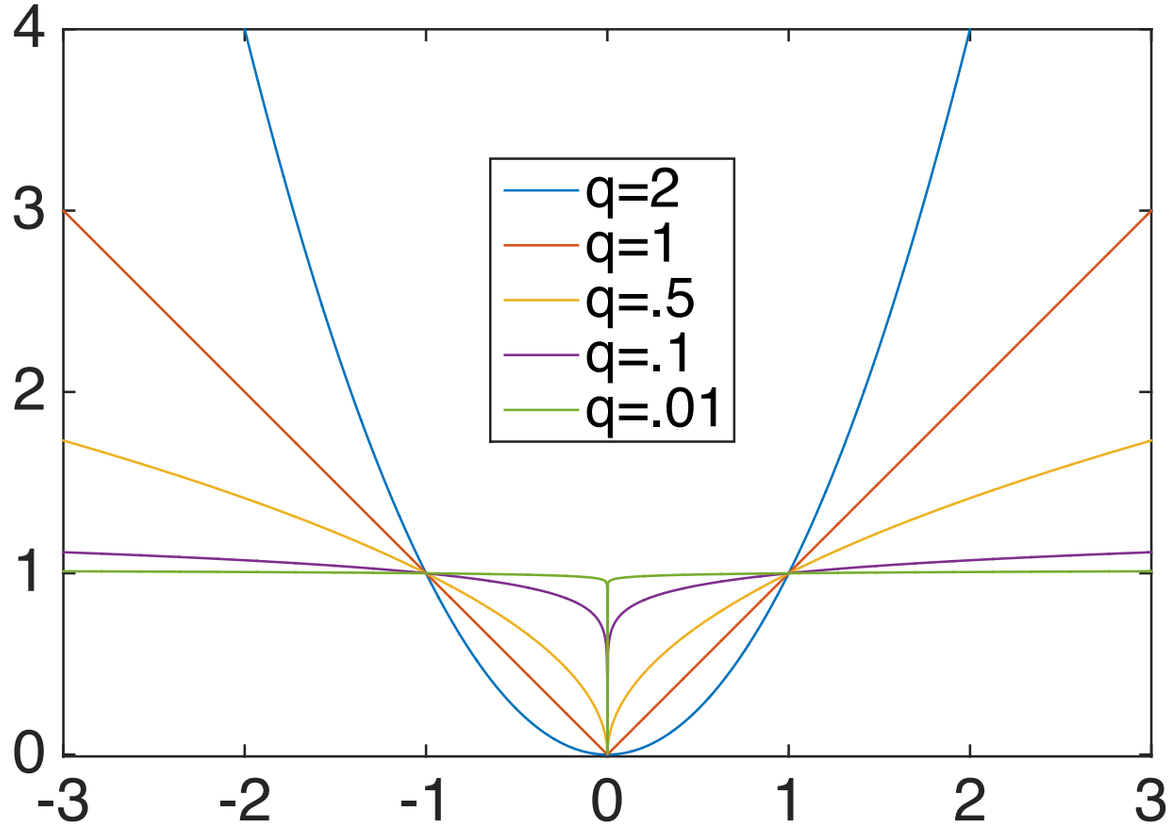
- That is, ϕ is the **conditional median**

$$\int_{y|y \succ \phi(x)} p(y|x)dy = \int_{y|y \prec \phi(x)} p(y|x)dy = \frac{1}{2}$$

Example #3

- Recall Minkowski's loss $L_q(y, \phi, x) = |y - \phi(x)|^q$

$$q \to 0$$



Example #3

- Recall Minkowski's loss $L_q(y, \phi, x) = |y - \phi(x)|^q$
- When $q \to 0$ the loss converges to

$$\lim_{q \to 0} |y - \phi(x)|^q = \begin{cases} 1 & \text{if } y \neq \phi(x) \\ 0 & \text{if } y = \phi(x) \end{cases}$$

Example #3

- Recall Minkowski's loss $L_q(y, \phi, x) = |y - \phi(x)|^q$

- Let $q \to 0$, then Bayes risk minimization leads to

$$\begin{aligned}
\hat{\phi} &= \arg\min_{\phi} \int L_{q\to 0}(y, \phi, x) p(x, y) dx dy \\
&= \arg\min_{\phi} \int \left(\int L_{q\to 0}(y, \phi, x) p(y|x) dy \right) p(x) dx \\
&= \arg\min_{\phi} 1 - \int p(\phi(x)|x) p(x) dx
\end{aligned}$$

Maximum a Posteriori

- Recall Minkowski's loss $L_q(y, \phi, x) = |y - \phi(x)|^q$

- Let $q \to 0$, then Bayes risk minimization leads to

Maximum a Posteriori

$$\hat{\phi}(x) = \arg\max_{\phi(x)} p(\phi(x)|x)$$

Maximum a Posteriori

- Can be rewritten as

$$\hat{\phi}(x) = \arg\max_{\phi(x)} p(\phi(x)|x)$$

$$= \arg\max_{\phi(x)} \frac{p(x, \phi(x))}{p(x)}$$

$$= \arg\max_{\phi(x)} \frac{p(x|\phi(x))p_Y(\phi(x))}{p(x)}$$

$$= \arg\max_{\phi(x)} p(x|\phi(x))p_Y(\phi(x))$$

$$= \arg\max_{\phi(x)} \log p(x|\phi(x)) + \log p_Y(\phi(x))$$

data

prior

Example #4

- From the Maximum a Posteriori formulation

$$\hat{\phi}(x) = \arg\max_{\phi(x)} \log p(x|\phi(x)) + \log p_Y(\phi(x))$$

we choose the data model

$$p(x|\phi(x)) \propto e^{-\frac{|x-\phi(x)|^2}{2\sigma^2}}$$

and the prior

$$p_Y(\phi(x)) \propto e^{-\frac{|\phi(x)|^2}{2\sigma_Y^2}}$$

Example #4

- We obtain

$$\hat{\phi}(x) = \arg\max_{\phi(x)} \log p(x|\phi(x)) + \log p_Y(\phi(x))$$

$$= \arg\min_{\phi(x)} \frac{|x - \phi(x)|^2}{2\sigma^2} + \frac{|\phi(x)|^2}{2\sigma_Y^2}$$

which gives the closed-form solution

$$\hat{\phi}(x) = \frac{\sigma_Y^2}{\sigma^2 + \sigma_Y^2} x$$

Example #5

- From the Maximum a Posteriori formulation

$$\hat{\phi}(x) = \arg\max_{\phi(x)} \log p(x|\phi(x)) + \log p_Y(\phi(x))$$

we choose the data model

$$p(x|\phi(x)) \propto e^{-\frac{1}{2}(x-A\phi(x))^\top \Sigma^{-1}(x-A\phi(x))}$$

and the prior

$$p_Y(\phi(x)) \propto e^{-\frac{1}{2}|\Delta\phi(x)|^2}$$

Example #5

- We obtain

$$\hat{\phi}(x) = \arg\max_{\phi(x)} \log p(x|\phi(x)) + \log p_Y(\phi(x))$$

$$= \arg\min_{\phi(x)} \frac{1}{2}(x - A\phi(x))^\top \Sigma^{-1}(x - A\phi(x)) + \frac{1}{2}|\Delta\phi(x)|^2$$

which gives the closed-form solution

$$\hat{\phi}(x) = \left(A^\top \Sigma^{-1} A + \Delta^\top \Delta\right)^{-1} A^\top \Sigma^{-1} A x$$

Example #6

- From the Maximum a Posteriori formulation

$$\hat{\phi}(x) = \arg\max_{\phi(x)} \log p(x|\phi(x)) + \log p_Y(\phi(x))$$

we choose the data model

$$p(x \mid \phi(x)) \propto e^{-\frac{|x - A\phi(x)|^2}{2\sigma^2}}$$

and the prior

$$p_Y(\phi(x)) \propto e^{-|\nabla\phi(x)|_{TV}}$$

Example #6

- We obtain

$$\hat{\phi}(x) = \arg\max_{\phi(x)} \log p(x|\phi(x)) + \log p_Y(\phi(x))$$

$$= \arg\min_{\phi(x)} \frac{1}{2\sigma^2}|x - A\phi(x)|^2 + |\nabla\phi(x)|_{TV}$$

which has no known closed-form solution

$$\hat{\phi}(x) = ?$$

Example #6

- How do we solve

$$\hat{\phi}(x) = \arg\min_{\phi(x)} \frac{1}{2\sigma^2}|x - A\phi(x)|^2 + |\nabla\phi(x)|_{TV}$$

- Recall the techniques in the previous lectures: Discretize the energy, compute the energy gradient, and solve the gradient equation $\nabla_\phi E = 0$ with gradient descent or linearization

Example #6

- If we use gradient descent we iterate

$$\phi^{t+1}(x) = \phi^t(x) - \epsilon \nabla_\phi E\left[\phi^t\right]$$

where

$$E[\phi] = \frac{1}{2\sigma^2}|x - A\phi(x)|^2 + |\nabla\phi(x)|_{TV}$$

and then let

$$\hat{\phi}(x) = \phi^\tau(x)$$

Example #6

- Issues with the original energy
  - Computation of the gradient $\nabla_\phi E[\phi]$ at each iteration might be computationally intensive (e.g., inversion of large matrices)
  - Gradient might be undefined (e.g., absolute value)
  - Difficult to incorporate additional constraints

Example #6

- An approach to minimize these energies is to use Majorization Minimization

- We describe this method in the next part

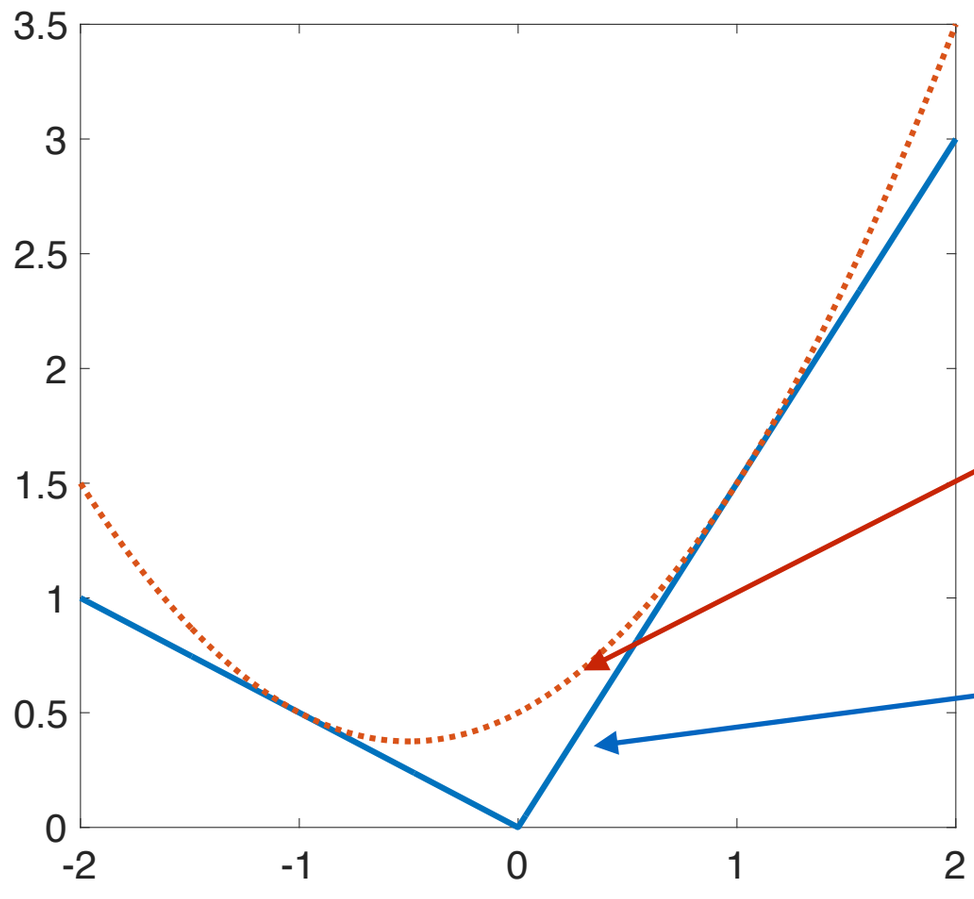
Majorization Minimization

- MM includes EM as a special case
- Easy to apply
- Probabilistic framework - free

Majorization Minimization

- A method to build an **optimization procedure**

    1. Can deal with non-differentiable problems



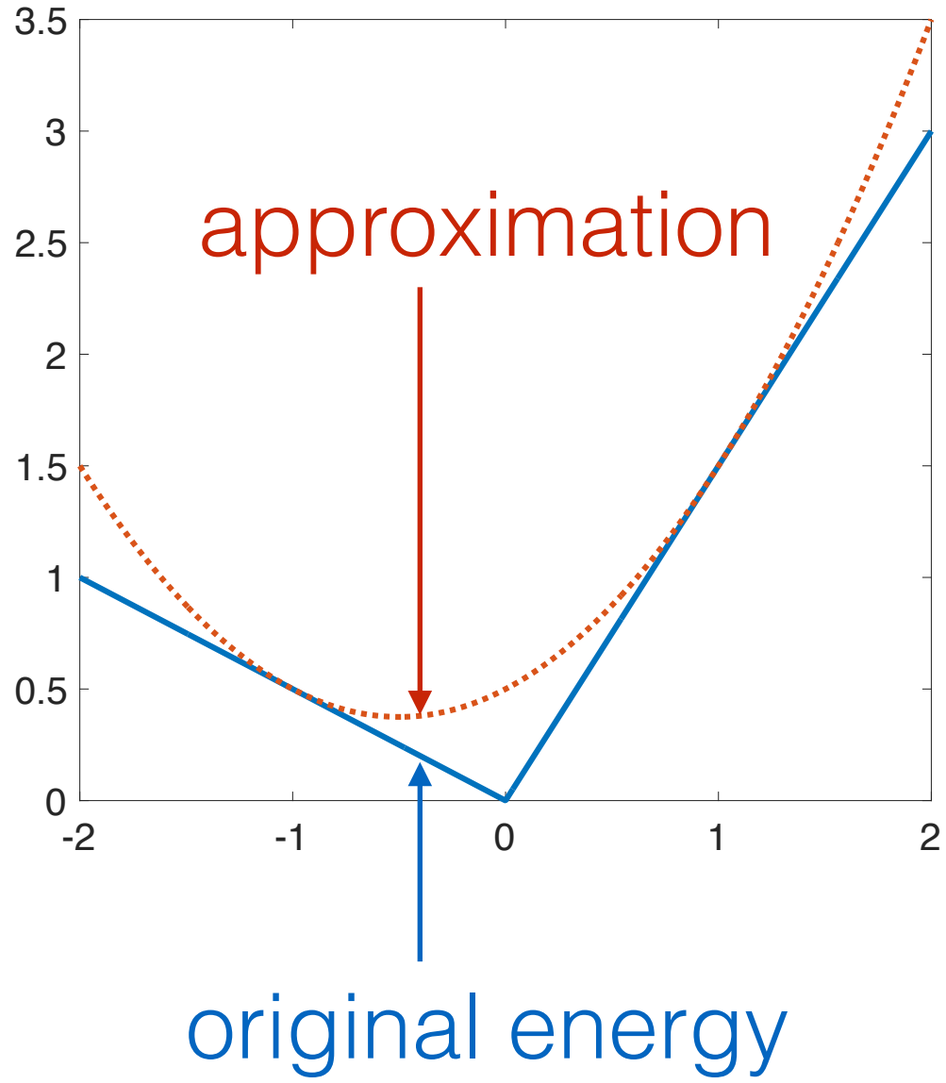
Majorization Minimization

- A method to build an optimization procedure

1. Can deal with non-differentiable problems
2. Can reduce computations (eg, matrix inversion, variables separation)
3. Can linearize problem
4. Can deal with constraints (equalities/ inequalities)

Majorization Minimization

- Key ideas
  - **At each iteration** introduce a “nice” approximation of the original energy
  - Solve the approximation efficiently/easily
  - Solving the approximation leads to a solution of the original energy



Majorization Minimization

- Let $E(\phi, x)$ be a function to be optimized with respect to ϕ, the model parameters, given some observation x

$$\hat{\phi}(x) = \arg\min_{\phi} E(\phi, x)$$

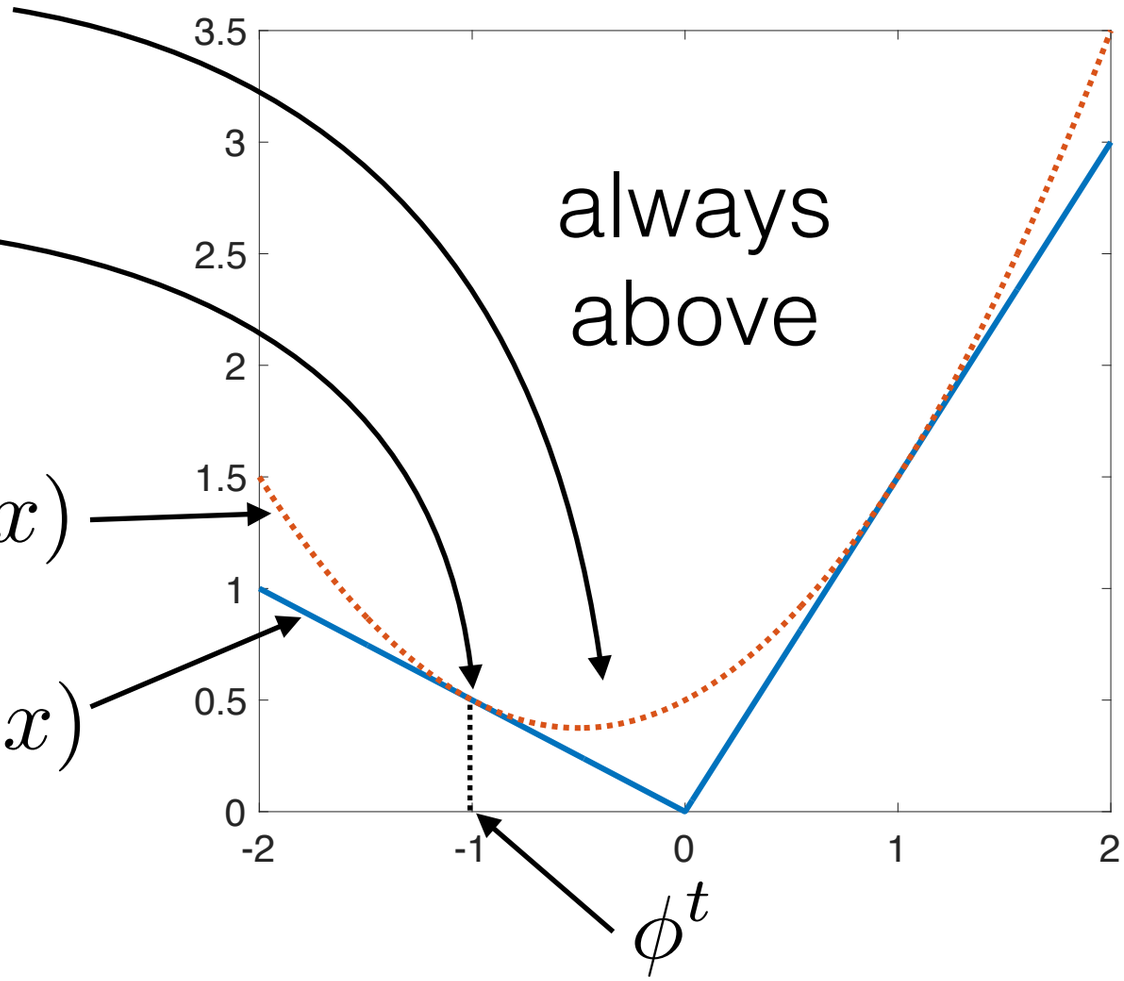
- Suppose that it is not "easy" to optimize $E(\phi, x)$ with standard gradient descent

Majorization Minimization

- Define the **surrogate function** $g(\phi|\phi^t, x)$ such that

$$g(\phi|\phi^t, x) \geq E(\phi, x) \quad \forall \phi$$

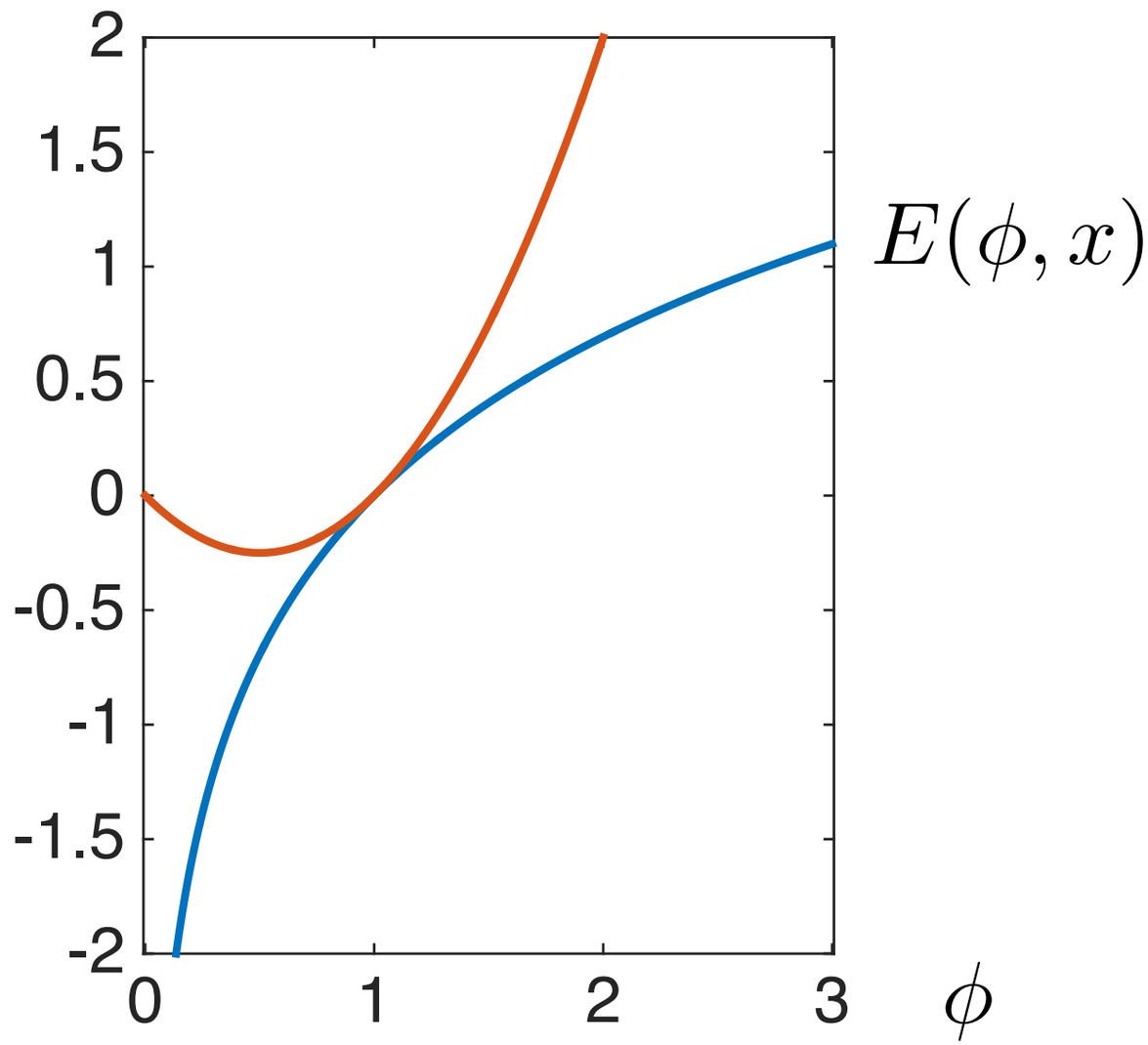
$$g(\phi^t|\phi^t, x) = E(\phi^t, x)$$



$g(\phi|\phi^t, x)$ **changes** at each iteration

Example

- Consider $E(\phi, x) = \log(\phi) - \log(x)$

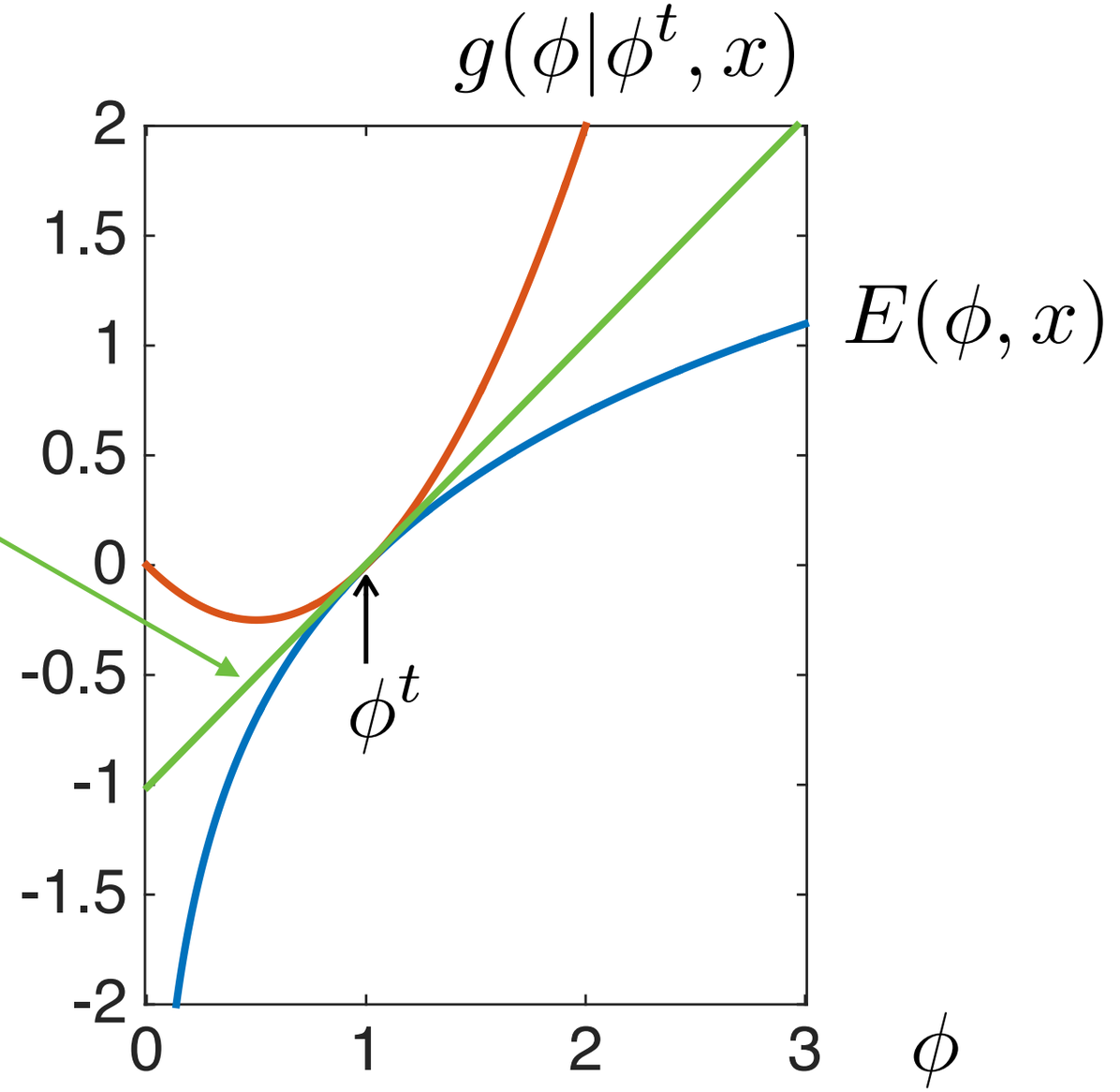


Example

- Consider $E(\phi, x) = \log(\phi) - \log(x)$

- Define

$$g\left(\phi \mid \phi^t, x\right) = \log \phi^t - \log x + \frac{1}{\phi^t}\left(\phi - \phi^t\right) + \frac{1}{2\left(\phi^t\right)^2}\left(\phi - \phi^t\right)^2$$



Majorization Minimization

- The surrogate function is said to **majorize** the objective function
- The following (MM) algorithm

$$\phi^{t+1} = \arg\min_{\phi} g(\phi|\phi^t, x)$$

minimizes the original objective function

Majorization Minimization

- Minimization of the original objective function

$$g(\phi|\phi^t, x) \geq E(\phi, x) \quad \forall\phi$$

$$\begin{aligned} E(\phi^{t+1}, x) &\leq g(\phi^{t+1}|\phi^t, x) \\ &\leq g(\phi^t|\phi^t, x) \quad \leftarrow \phi^{t+1} = \arg\min_{\phi} g(\phi|\phi^t, x) \\ &= E(\phi^t, x) \end{aligned}$$

$$g(\phi^t|\phi^t, x) = E(\phi^t, x)$$

- The MM solution minimizes the original energy

Building surrogate functions

- MM boils down to defining the surrogate function
- Here are some common techniques that can be tried

    a. Jensen's inequality

    b. Supporting hyperplanes

    c. Quadratic upper bound of a convex function

    d. Arithmetic-Geometric mean inequality

    e. Cauchy-Schwarz inequality

Jensen's inequality

- For a convex function φ we have

$$\varphi(\mathbf{E}[X]) \leq \mathbf{E}[\varphi(X)]$$

random variable

expectation

- This is a useful upper bound (see the EM derivation)

Jensen's inequality

- For a convex function φ we have

$$\varphi(\mathbf{E}[X]) \leq \mathbf{E}[\varphi(X)]$$

- If φ is strictly convex then we have equality if and only if

$$X = \text{const}$$

- The opposite inequality applies to concave functions

Example

- Let $E(\phi, x) = p(\phi|x)$ and $\varphi(z) = -\log(z)$
- $\varphi(z)$ is convex

Example

- Let $E(\phi, x) = p(\phi|x)$ and $\varphi(z) = -\log(z)$

$$\log p(\phi|x) = \log \int q(z) \frac{p(\phi, z|x)}{q(z)} dz$$

Example

- Let $E(\phi, x) = p(\phi|x)$ and $\varphi(z) = -\log(z)$

$$\varphi\left(\mathbf{E}\left[\frac{p(\phi, z|x)}{q(z)}\right]\right)$$

$$\log p(\phi|x) = \log \int q(z) \frac{p(\phi, z|x)}{q(z)} dz$$

VI

Jensen's inequality

$$\geq \int q(z) \log \frac{p(\phi, z|x)}{q(z)} dz$$

$$\mathbf{E}\left[\varphi\left(\frac{p(\phi, z|x)}{q(z)}\right)\right]$$

Example

- Let $E(\phi, x) = p(\phi|x)$ and $\varphi(z) = -\log(z)$

$$\begin{aligned}
\log p(\phi|x) &= \log \int q(z) \frac{p(\phi, z|x)}{q(z)} dz \\
&\geq \int q(z) \log \frac{p(\phi, z|x)}{q(z)} dz \\
&= -\mathrm{KL}(q(z) \| p(\phi, z|x))
\end{aligned}$$

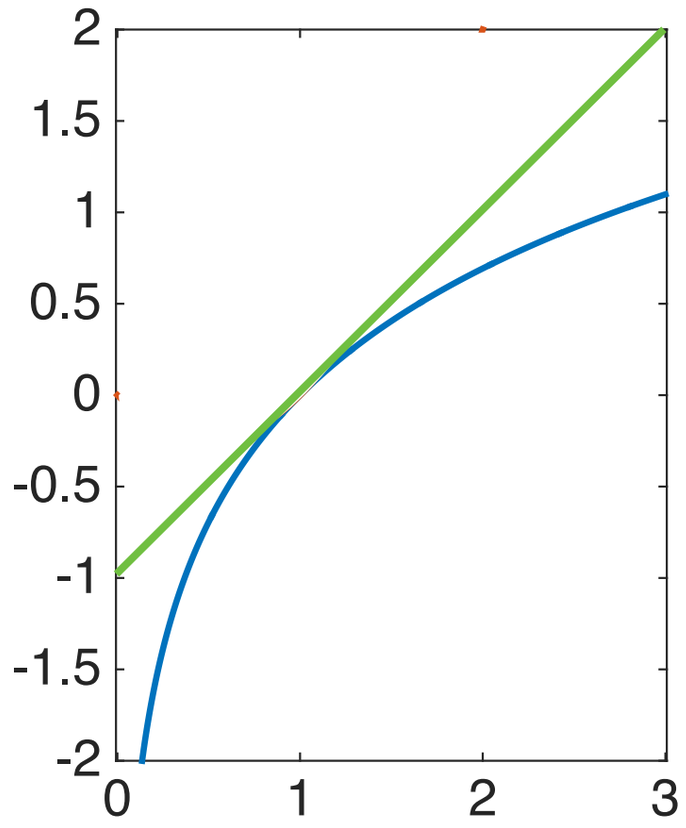
Kullback-Leibler divergence

with equality if $q(z) = p(z|\phi, x)$

Supporting hyperplanes

- If the objective function is concave then

$$E(\phi, x) \leq E(\phi^t, x) + \nabla E(\phi^t, x)^\top (\phi - \phi^t) = g(\phi | \phi^t, x)$$



Example

- Consider the concave energy

$$E(\phi, x) = (\phi - 1)^2 + \lambda \log(|\phi| + 1)$$

- Apply the supporting hyperplane to the logarithm

$$g(\phi|\phi^t, x) = (\phi - 1)^2 + \lambda \log(|\phi^t| + 1) + \frac{\lambda}{|\phi^t| + 1}(\phi - |\phi^t| - 1)$$

Example

- With the surrogate function

$$g(\phi|\phi^t, x) = (\phi - 1)^2 + \lambda \log(|\phi^t| + 1) + \frac{\lambda}{|\phi^t| + 1}(\phi - |\phi^t| - 1)$$

the MM algorithm $\phi^{t+1} = \arg\min_{\phi} g(\phi|\phi^t, x)$ becomes

$$\phi^{t+1} = 1 - \frac{\lambda}{2(|\phi^t| + 1)}$$

Example

- Simple application: This can be used to compute $\frac{1}{\sqrt{2}}$
- Let $\lambda = 1$ and set the derivatives of the original function to 0

$$\nabla E(\phi, x) = 2(\phi - 1) + \frac{1}{\phi + 1} = 0$$

- The solution (minimum) is $\phi = \frac{1}{\sqrt{2}}$

  and MM gives us a simple* algorithm to compute it

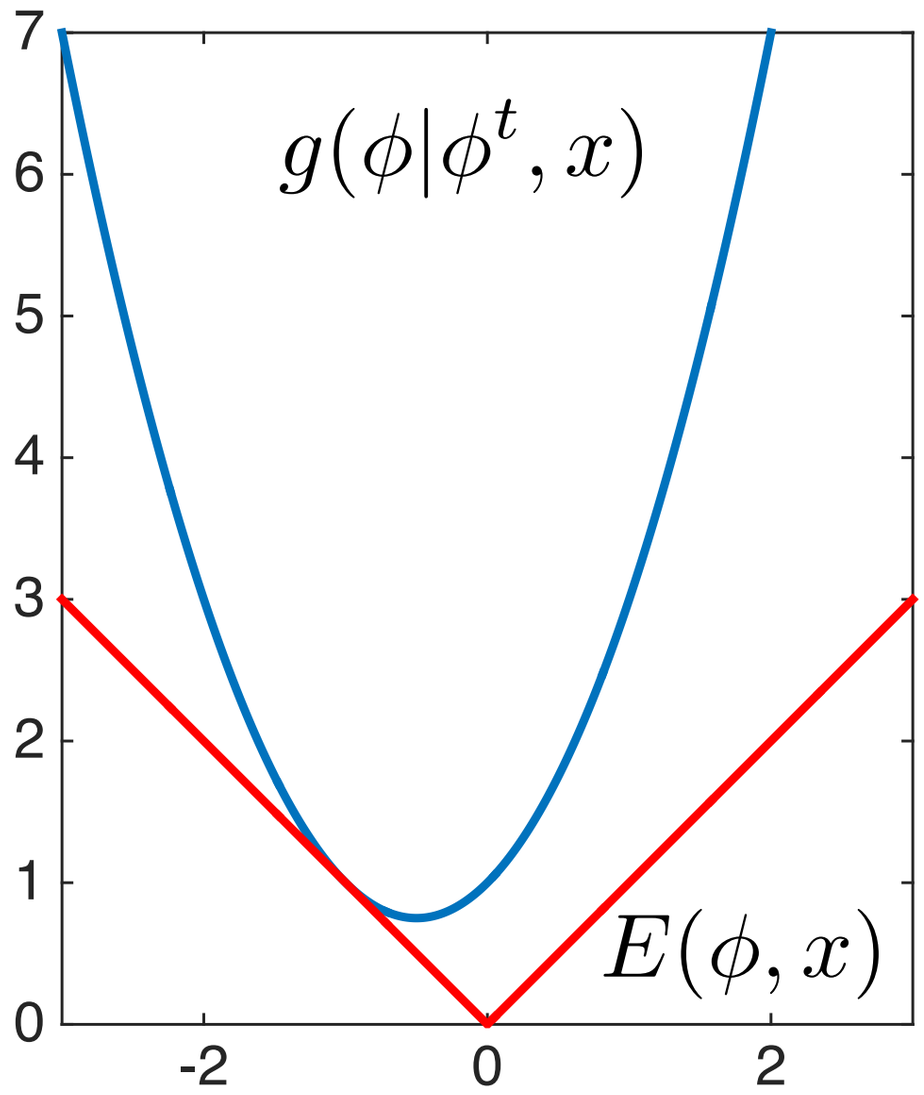
$$\phi^{t+1} = 1 - \frac{1}{2(|\phi^t| + 1)}$$

*No trigonometric functions, no exponentials or logarithms, no series expansions, but only the basic 4 operations

Quadratic upper-bound

- If the objective function $E(\phi, x)$ is convex

- Use the Taylor expansion and the mean value theorem on the Hessian to find an upper bound $g(\phi|\phi^t, x)$



Quadratic upper-bound

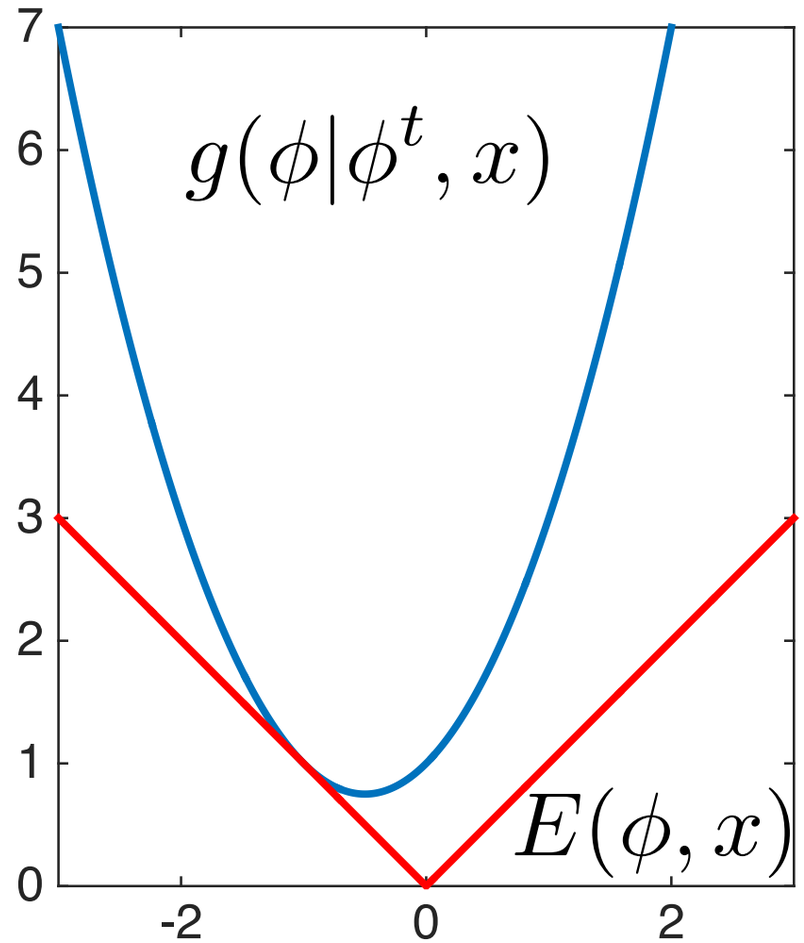
- Let $HE(\phi, x)$ be the Hessian and $\nabla E(\phi, x)$ the gradient of $E(\phi, x)$

- Let us pick $M \succ 0$ such that

$$M - HE(\phi^t, x) \succeq 0$$

then, we have

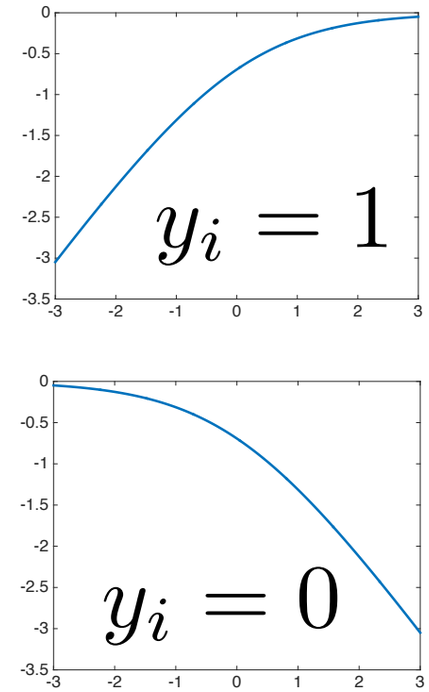
$$\begin{aligned} E(\phi, x) &\leq E\left(\phi^t, x\right) + \nabla E\left(\phi^t, x\right)^\top \left(\theta - \theta^t\right) \\ &\quad + \left(\theta - \theta^t\right)^\top \frac{M}{2} \left(\theta - \theta^t\right) \\ &= g\left(\phi \mid \phi^t, x\right) \end{aligned}$$



Example

- In logistic regression we have

$$E(\phi, x, y) = \sum_i \log\left(\frac{1}{1+e^{-x_i^\top \phi}}\mathbf{1}(y_i = 1)\right.$$
$$\left. + \frac{e^{-x_i^\top \phi}}{1+e^{-x_i^\top \phi}}\mathbf{1}(y_i = 0)\right)$$



- This objective function is concave, thus we can apply the quadratic **lower bound**

Example

- Through the quadratic lower-bound we obtain the surrogate function

$$g(\phi|\phi^t, x, y) = E(\phi^t, x, y) + \sum_i \left(\frac{e^{-x_i^\top \phi^t} \mathbf{1}(y_i = 1)}{1 + e^{-x_i^\top \phi^t}} - \frac{\mathbf{1}(y_i = 0)}{1 + e^{-x_i^\top \phi^t}} t \right).$$

$$x_i^\top (\phi - \phi^t) - (\phi - \phi^t)^\top \sum_i \frac{x_i x_i^\top}{8} (\phi - \phi^t)$$

with $M = -\sum_i \frac{x_i x_i^\top}{4}$, which is the lowest Hessian value

Example

- The algorithm $\phi^{t+1} = \arg\min_{\phi} g(\phi|\phi^t, x, y)$ becomes

$$\phi^{t+1} = \phi^t + 4\left(\sum_i x_i x_i^\top\right)^{-1} \sum_j x_j \left(\mathbf{1}(y_i = 1) - \frac{1}{1 + e^{-x_i^\top \phi^t}}\right)$$

Arithmetic-Geometric mean inequality

- The function φ is convex if

$$\varphi\left(\sum_i \alpha_i z_i\right) \leq \sum_i \alpha_i \varphi(z_i)$$

with $\alpha_i \geq 0$ and $\sum_i \alpha_i = 1$

Arithmetic-Geometric mean inequality

- Let $\varphi(x) = e^x$ and $\alpha_i = \frac{1}{n}$ then

$$\exp\left(\frac{1}{n}\sum_i x_i\right) \leq \frac{1}{n}\sum_i e^{x_i}$$

and with $z_i = e^{x_i}$ we finally obtain

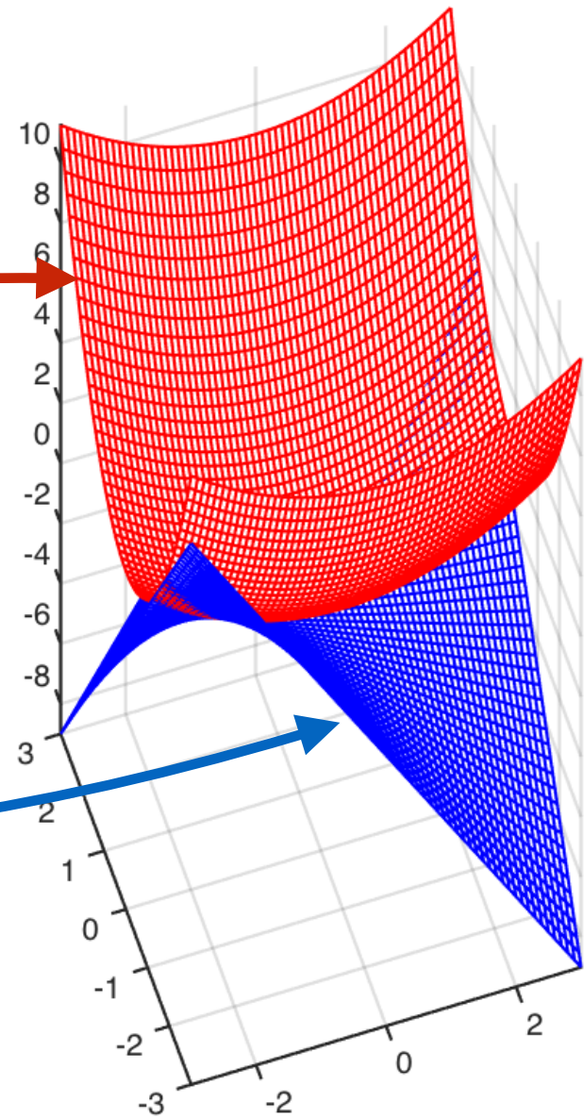
$$\sqrt[n]{\prod_i z_i} \leq \frac{1}{n}\sum_i z_i$$

Arithmetic-Geometric mean inequality

- The Arithmetic-Geometric mean inequality holds with equality when all terms are identical

- With two terms we obtain

$$x_1 x_2 \leq x_1^2 \frac{x_2^t}{2x_1^t} + x_2^2 \frac{x_1^t}{2x_2^t}$$



Arithmetic-Geometric mean inequality

- We obtain $x_1 x_2 \leq x_1^2 \frac{x_2^t}{2x_1^t} + x_2^2 \frac{x_1^t}{2x_2^t}$

by chosing $z_1 = x_1^2 \frac{x_2^t}{x_1^t}$ and $z_2 = x_1^2 \frac{x_1^t}{x_2^t}$

in

$$\sqrt[n]{\prod_i z_i} \leq \frac{1}{n} \sum_i z_i$$

Example

- Reweighed least squares uses this inequality

$$|x| \leq \frac{|x|^2 + |x^t|^2}{2|x^t|}$$

obtained by using $z_2 = |x^t|^2$ and $z_1 = |x|^2$ in

$$\sqrt[n]{\prod_i z_i} \leq \frac{1}{n} \sum_i z_i$$

Example

- Reweighed least squares uses this inequality

$$|x| \leq \frac{|x|^2 + |x^t|^2}{2|x^t|}$$

- In the optimization the constant term is irrelevant

$$\arg\min_x \frac{|x|^2 + |x^t|^2}{2|x^t|} = \arg\min_x \frac{|x|^2}{2|x^t|}$$

Cauchy-Schwarz inequality

- The Euclidean norm is convex, thus the supporting hyperplanes inequality yields

$$|x| \geq |x^t| + \frac{(x^t)^T}{|x^t|}(x - x^t)$$

- By rearranging we obtain the Cauchy-Schwarz inequality

$$|x^t||x| \geq x^T x^t$$

- Or, alternatively

$$|x| \geq \frac{x^T x^t}{|x^t|}$$

Expectation Maximization

- Solves problems with latent variables (z)

$$\tilde{\phi} = \arg\max_{\phi} \log p(\phi, x)$$

$$= \arg\max_{\phi} \log \sum_{z} \underbrace{p(\phi, z, x)}$$

model parameters

data

optimization of this term should be easy

Expectation Maximization

- Can apply **EM** to solve **Maximum a Posteriori** problems with latent variables

$$\begin{aligned}\phi &= \arg\max_{\phi} \log p(\phi, x) \\ &= \arg\max_{\phi} \log \sum_{z} p(\phi, z, x)\end{aligned}$$

by introducing a **lower bound**

Expectation Maximization

- The decomposition

$$\log p(\phi, x) = \underbrace{\int q(z) \log \frac{p(\phi, z, x)}{q(z)} dz}_{-\mathrm{KL}(q \| p(\phi, z, x))} - \underbrace{\int q(z) \log \frac{p(z|\phi, x)}{q(z)} dz}_{\mathrm{KL}(q \| p(z|\phi, x))}$$

gives the following **lower bound**

$$\log p(\phi, x) \geq -\mathrm{KL}(q \| p(\phi, z, x))$$

because $\mathrm{KL}(q \| p(z|\phi, x)) \geq 0$

EM as a special case

- By splitting the algorithm in two steps we obtain

$$q^{t+1}(z) = p(z|\phi^t, x)$$ **E-step**

and

$$\begin{aligned}\phi^{t+1} &= \arg\max_{\phi} -\mathrm{KL}(q^{t+1}||p(\phi, z, x)) \\ &= \arg\min_{\phi} \mathrm{KL}(q^{t+1}||p(\phi, z, x))\end{aligned}$$ **M-step**

which is the **EM** algorithm

Jensen's inequality

- By applying Jensen's inequality we choose

$$g(\phi|\phi^t, x) = \mathrm{KL}(p(z|\phi^t, x)\|p(\phi, z, x))$$

as surrogate function

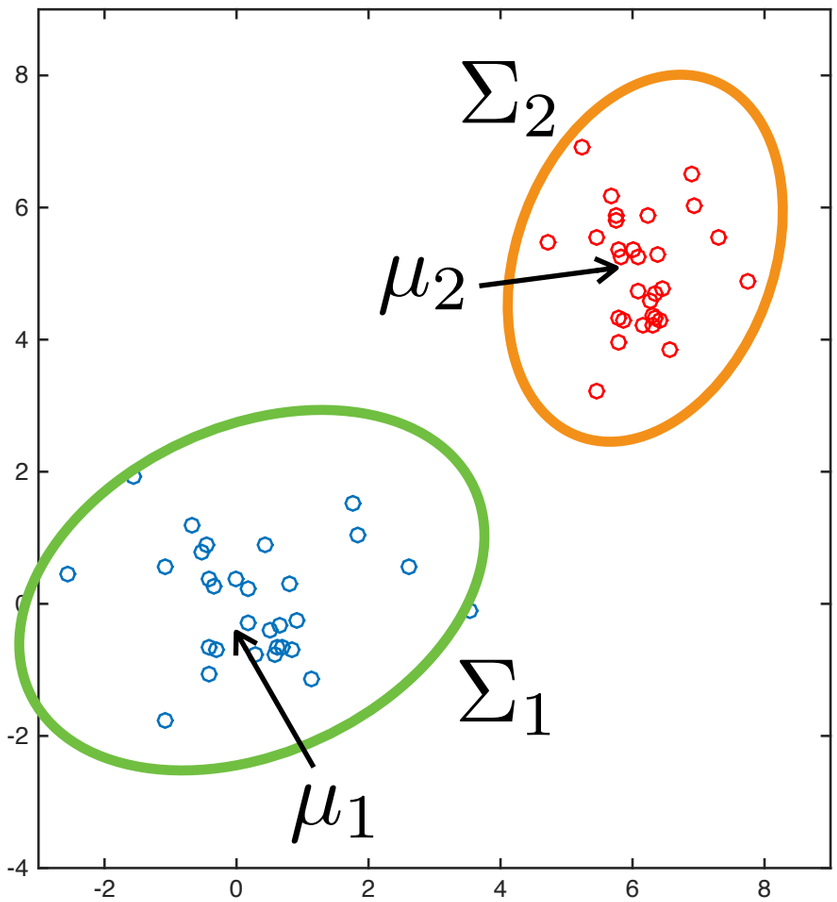
- The corresponding MM algorithm becomes

$$\phi^{t+1} = \arg\min_{\phi} \mathrm{KL}(p(z|\phi^t, x)\|p(\phi, z, x))$$

Example

- Collect m samples $x_1, \ldots, x_m$ with corresponding latent variables $z_1, \ldots, z_m$

- $p(z_i = j|\phi)$ is the probability that x_i belongs to the j-th Gaussian



- We model the samples as a **mixture of n Gaussians**

$$x_i|z_i = j, \phi \sim \mathcal{N}(\mu_j, \Sigma_j)$$

$$\varphi_j = p(z_i = j|\phi)$$

$$\phi = \{\varphi_1, \ldots, \varphi_n, \mu_1, \ldots, \mu_n, \Sigma_1, \ldots, \Sigma_n\}$$

Example

- **E-step**

Gaussian Multinomial

$$p(x_i, z_i|\phi) = p(x_i|z_i, \phi)p(z_i|\phi)$$

$$q_i^{t+1}(z_i) = p(z_i|x_i, \phi^t) = \frac{p(x_i, z_i|\phi^t)}{\sum_j p(x_i, z_i = j|\phi^t)}$$

- **M-step**

$$\phi^{t+1} = \arg\max_{\phi} \sum_i \sum_j q_i^{t+1}(z_i = j) \log p(\phi, z_i = j, x_i)$$

Example

- **M-step**

Gaussian Multinomial

$$p(\phi, z_i = j, x_i) = p(x_i|\phi, z_i = j)p(z_i = j|\phi)p(\phi)$$

$$\phi^{t+1} = \arg\max_{\phi} \sum_i \sum_j q_i^{t+1}(z_i = j) \log p(\phi, z_i = j, x_i)$$

$$= \arg\max_{\phi} \sum_i \sum_j q_i^{t+1}(j) \log p(x_i|\phi, j)p(j|\phi)p(\phi)$$

prior
(a constant if uniform)

Example

- Suppose that the prior is uniform, then we obtain

$$L(\phi) \doteq \sum_i \sum_j q_i^{t+1}(j) \log p(x_i|\phi, j) p(j|\phi) p(\phi)$$

$$= \sum_i \sum_j q_i^{t+1}(j) \left(\log p(x_i|\phi, j) + \log p(j|\phi) + \log p(\phi) \right)$$

Gaussian

Multinomial

constant
(we can ignore it)

$$\log \varphi_j$$

$$-\frac{n}{2} \log 2\pi - \frac{1}{2} \log |\Sigma_j| - \frac{1}{2} (x_i - \mu_j)^\top \Sigma_j^{-1} (x_i - \mu_j)$$

Example

- To find the maxima, compute the derivatives with respect to the parameters and set them to zero

- For example $\nabla_{\mu_j} L(\phi) = 0$

$$-\sum_i q_i^{t+1}(j)\Sigma_j^{-1}(x_i - \mu_j) = 0$$

$$\sum_i q_i^{t+1}(j)(x_i - \mu_j) = 0$$

$$\sum_i q_i^{t+1}(j)x_i = \sum_i q_i^{t+1}(j)\mu_j$$

$$\mu_j = \frac{\sum_i q_i^{t+1}(j)x_i}{\sum_i q_i^{t+1}(j)}$$

Example

- Suppose that the prior is uniform, then we obtain

$$\mu_j = \frac{\sum_i q_i^{t+1}(j) x_i}{\sum_i q_i^{t+1}(j)}$$

$$\varphi_j = \frac{1}{m} \sum_i q_i^{t+1}(j)$$

$$\Sigma_j = \frac{\sum_i q_i^{t+1}(j)(x_i - \mu_j)(x_i - \mu_j)^\top}{\sum_i q_i^{t+1}(j)}$$