# Bayesian Methods II

Paolo Favaro

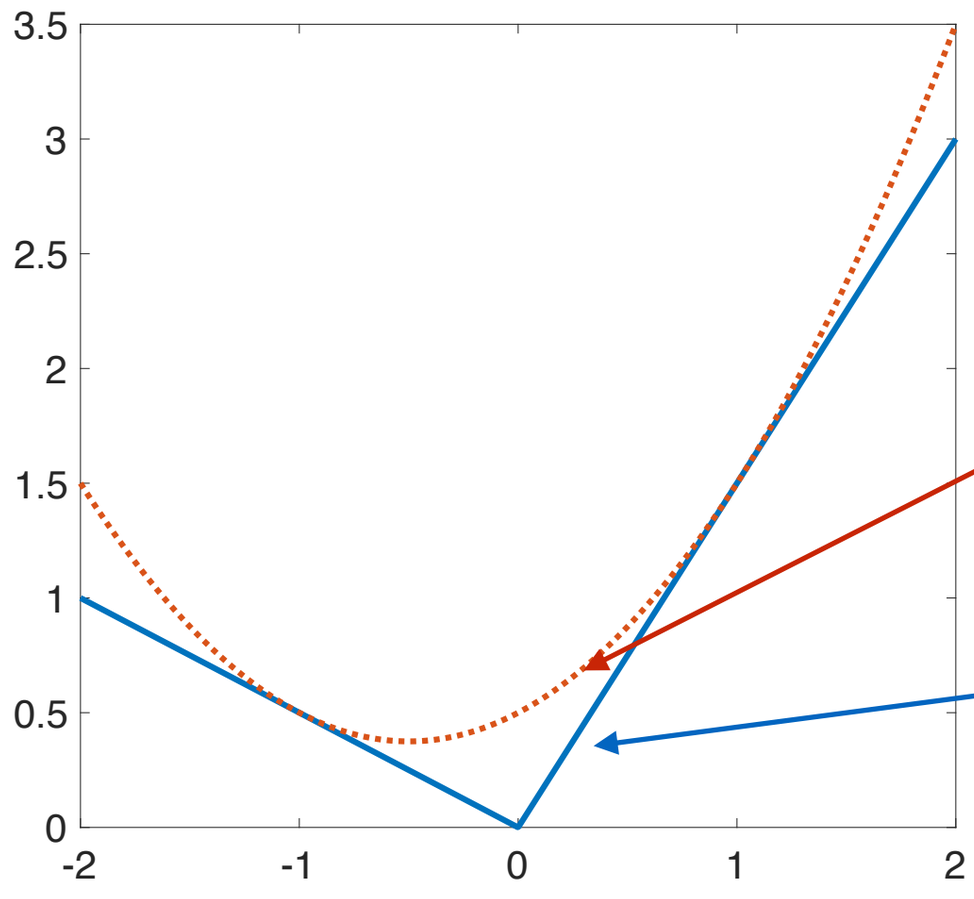
# Contents

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

# 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

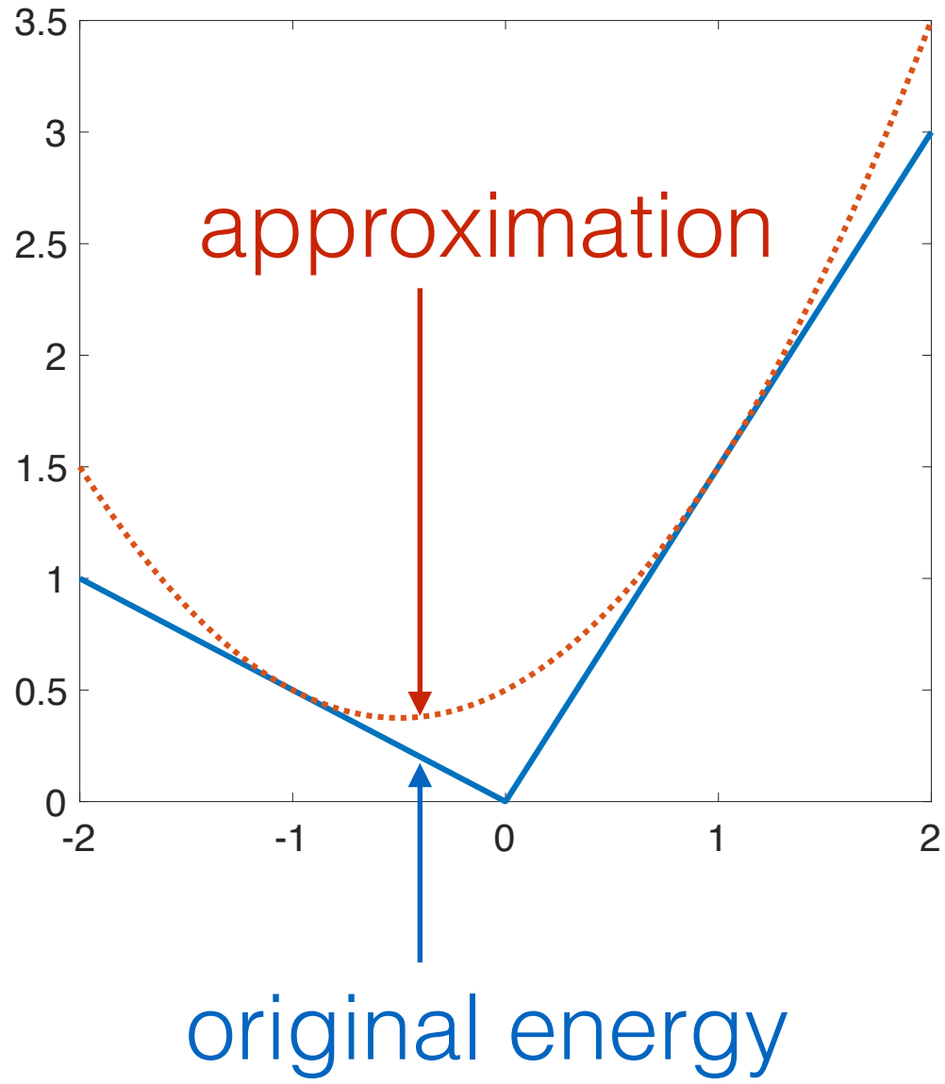
    4. Can deal with constraints (equalities/ inequalities)

# Majorization Minimization

- Probabilistic framework free
- Easy to apply
- MM includes EM (Expectation Maximization) as a special case

# 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 $\phi$, 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

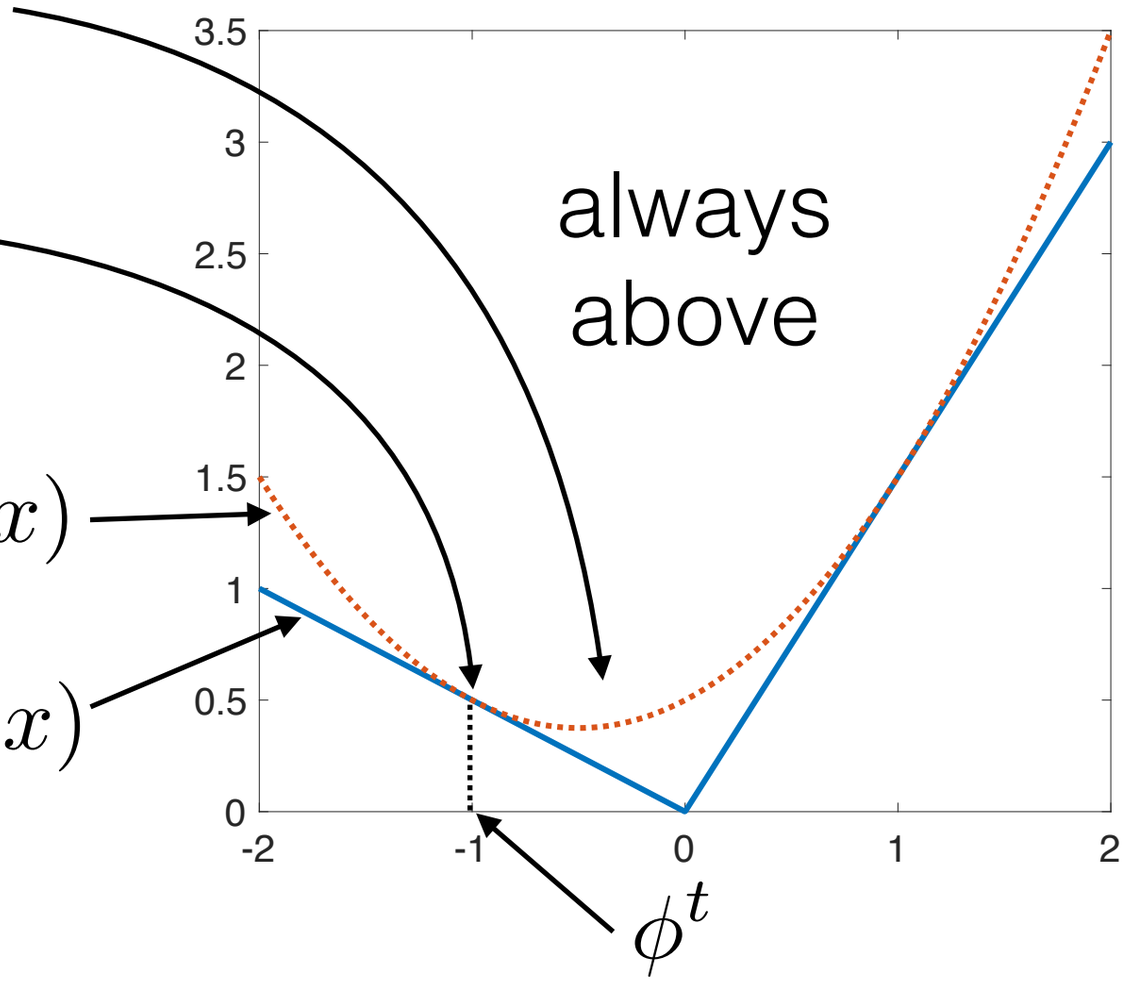
# Majorization Minimization

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

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

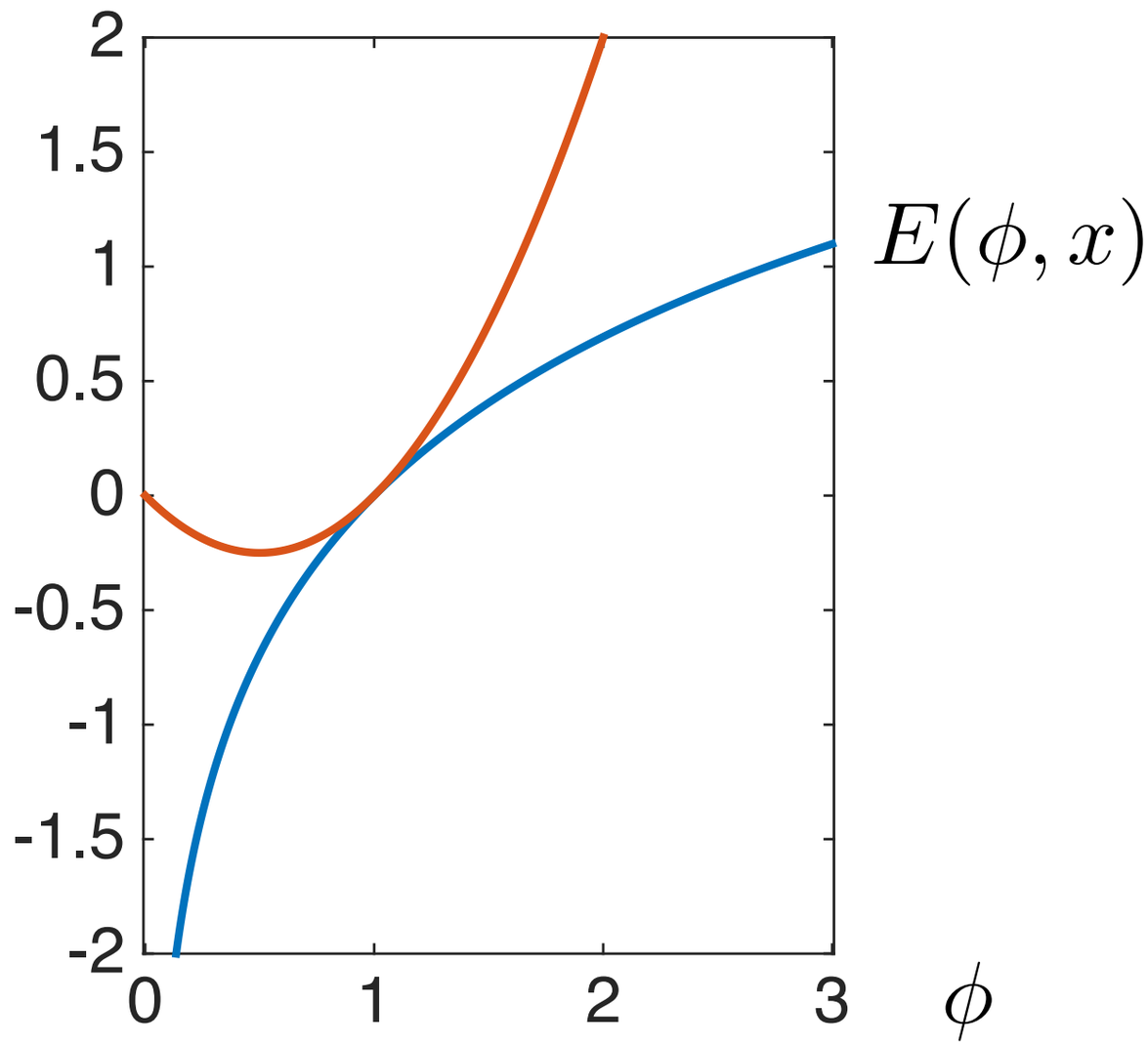
$$g(\phi^t|\phi^t, x) = E(\phi, x)$$

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



# Example

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

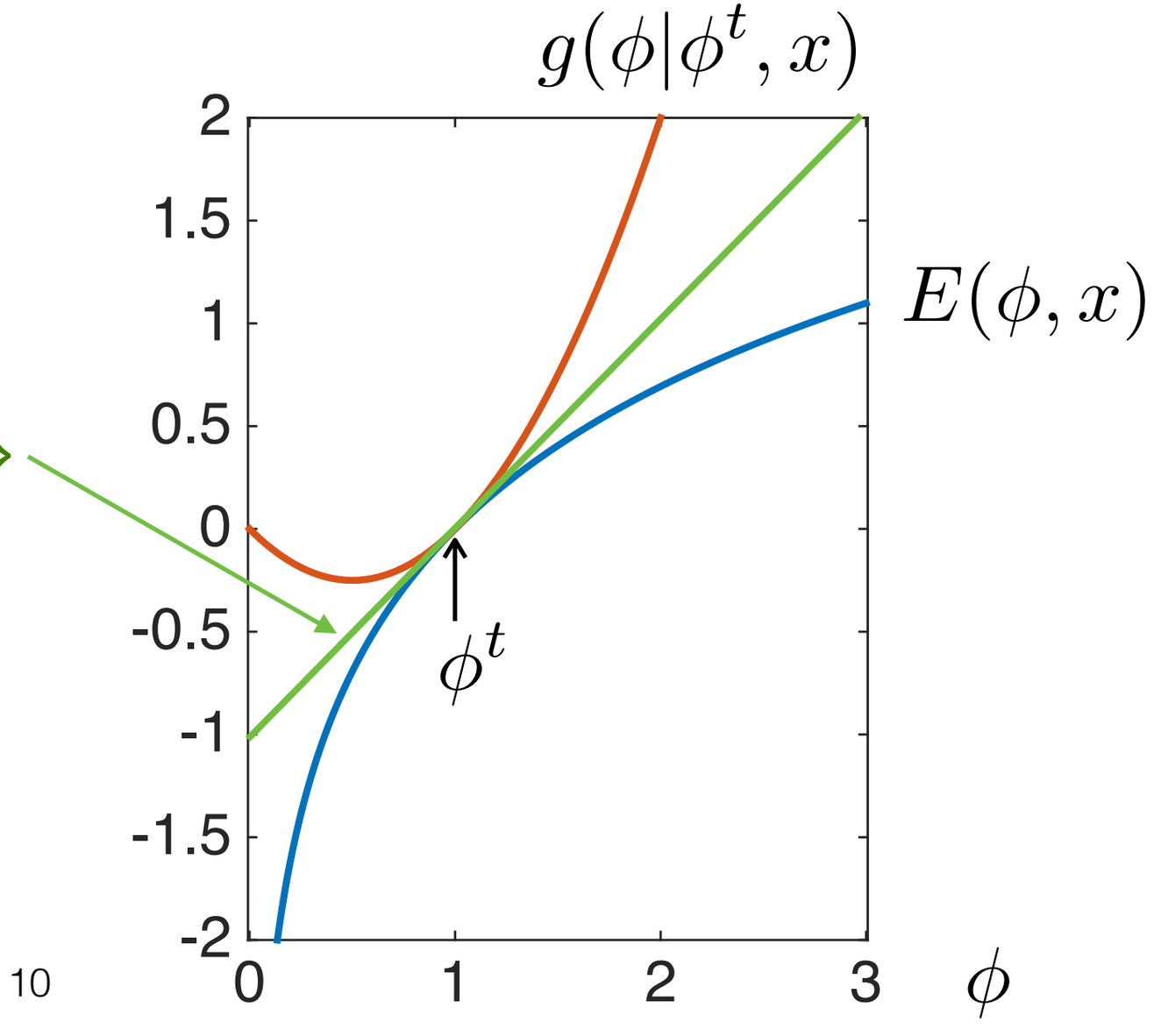


# Example

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

- Define

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



# 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

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

$$E(\phi^{t+1}, x) \leq g(\phi^{t+1}|\phi^t, x)$$

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

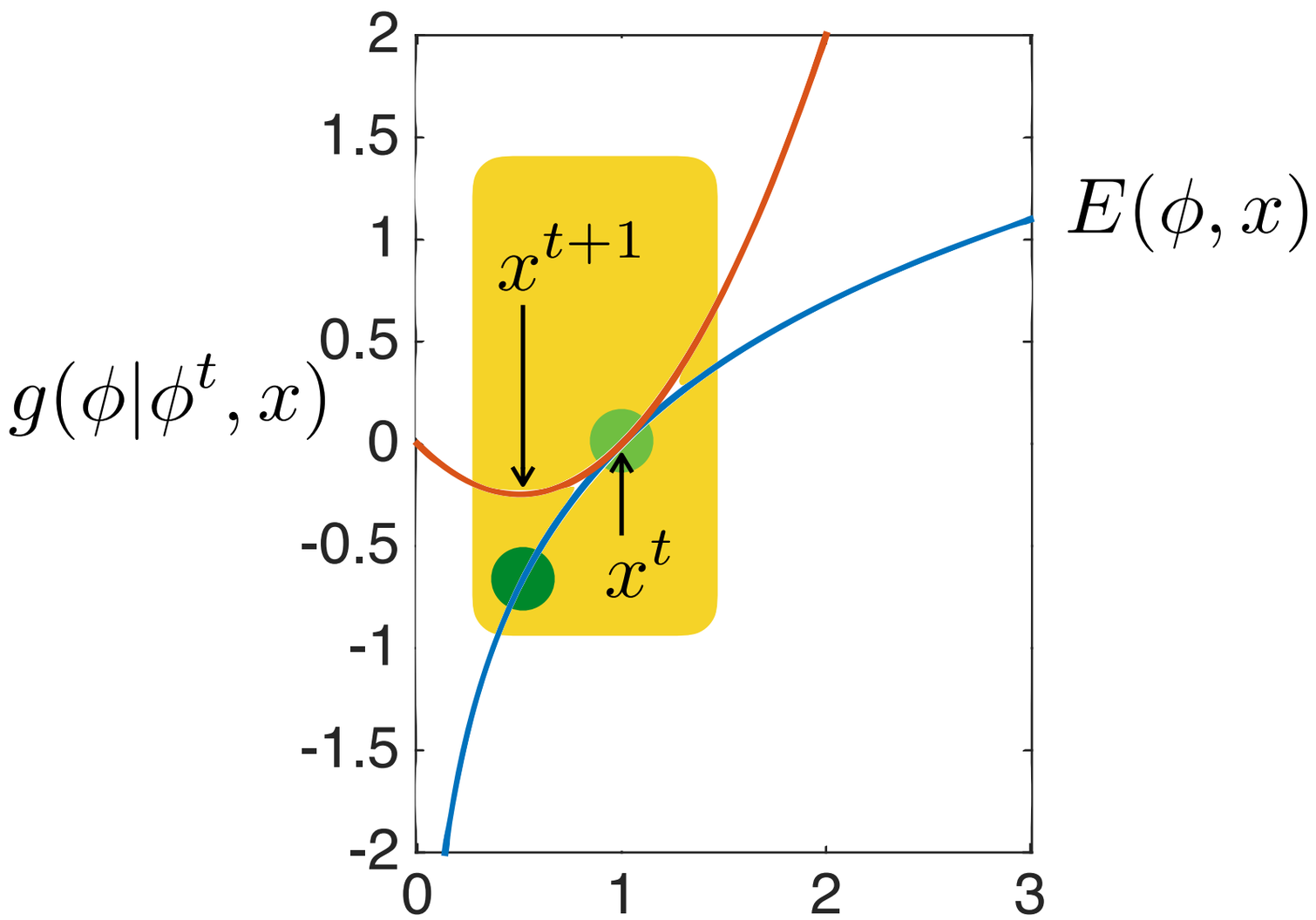
$$= E(\phi^t, x)$$

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

- The MM solution minimizes the original energy

# Example

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



# 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 $\varphi$ we have

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

random variable

expectation

- This is a useful upper bound

# Jensen's inequality

- For a convex function $\varphi$ we have

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

- If $\varphi$ 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 $\varphi(z) = -\log(z)$

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

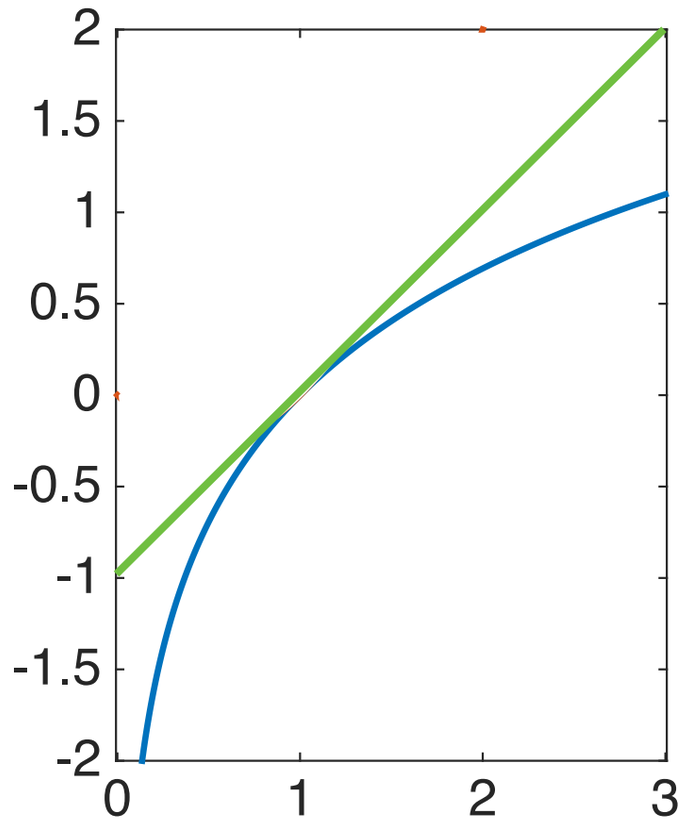
Kullback-Leibler divergence

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

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

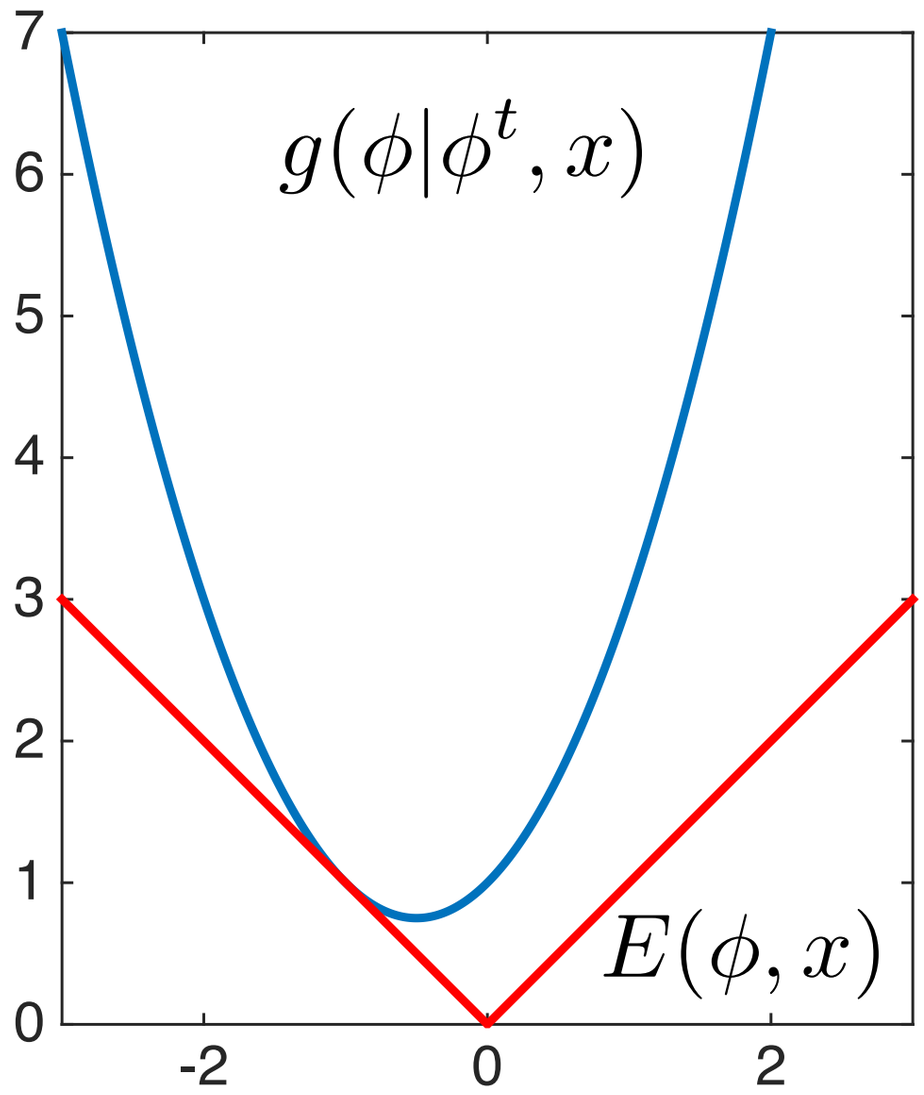
the MM algorithm $\phi^{t+1} = \arg\min_{\phi} g(\phi|\phi^t, x)$ becomes

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

# 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

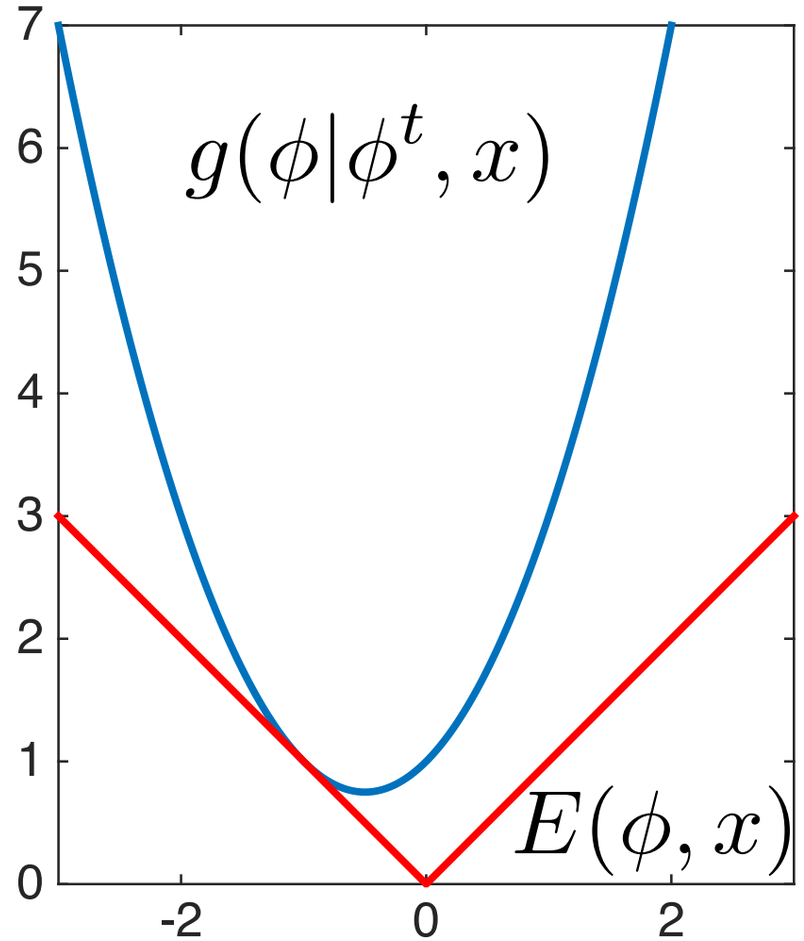
- Let $HE(\phi, x)$ be the Hessian and $\nabla E(\phi, x)$ the gradient of $E(\phi, x)$

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

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

then, we have

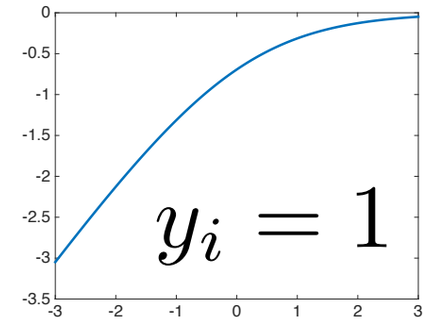
$$\begin{aligned} E(\phi, x) &\leq E(\phi^t, x) + \nabla E(\phi^t, x)^\top (\theta - \theta^t) \\ &\quad + (\theta - \theta^t)^\top M (\theta - \theta^t) \\ &= g(\phi | \phi^t, x) \end{aligned}$$

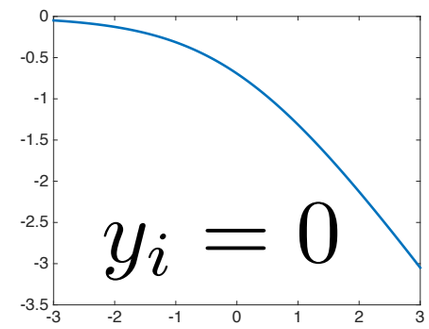


# Example

- In logistic regression we have

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




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

# Example

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

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

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

with $M = -\sum_i \frac{x_i x_i^\top}{8}$

# Example

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

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

# Arithmetic-Geometric mean inequality

- The function $\varphi$ 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

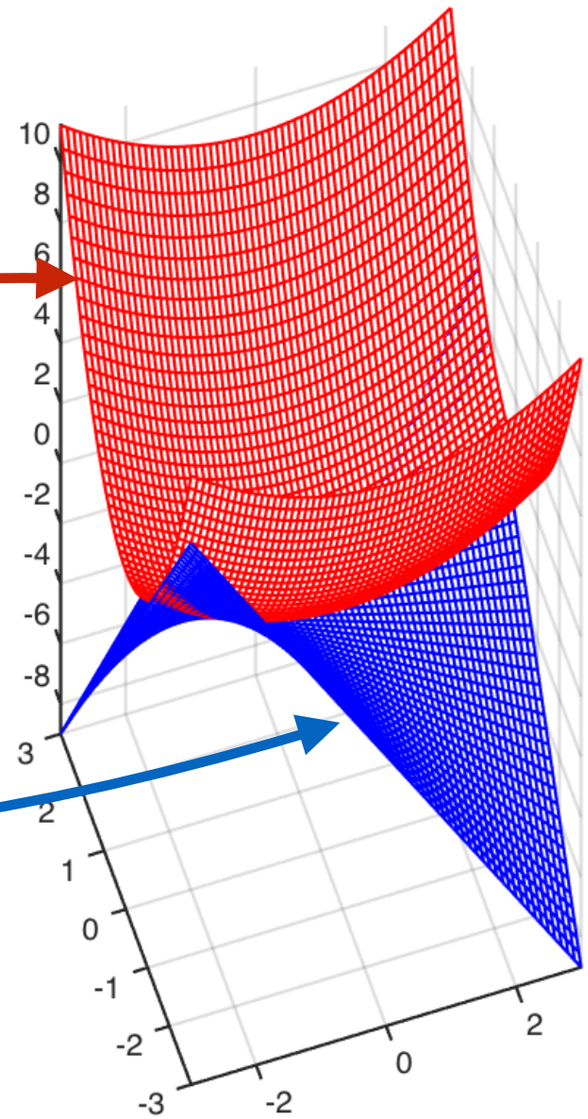
$$\sqrt[n]{\prod_i z_i} \leq \frac{1}{n}\sum_i z_i$$

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

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

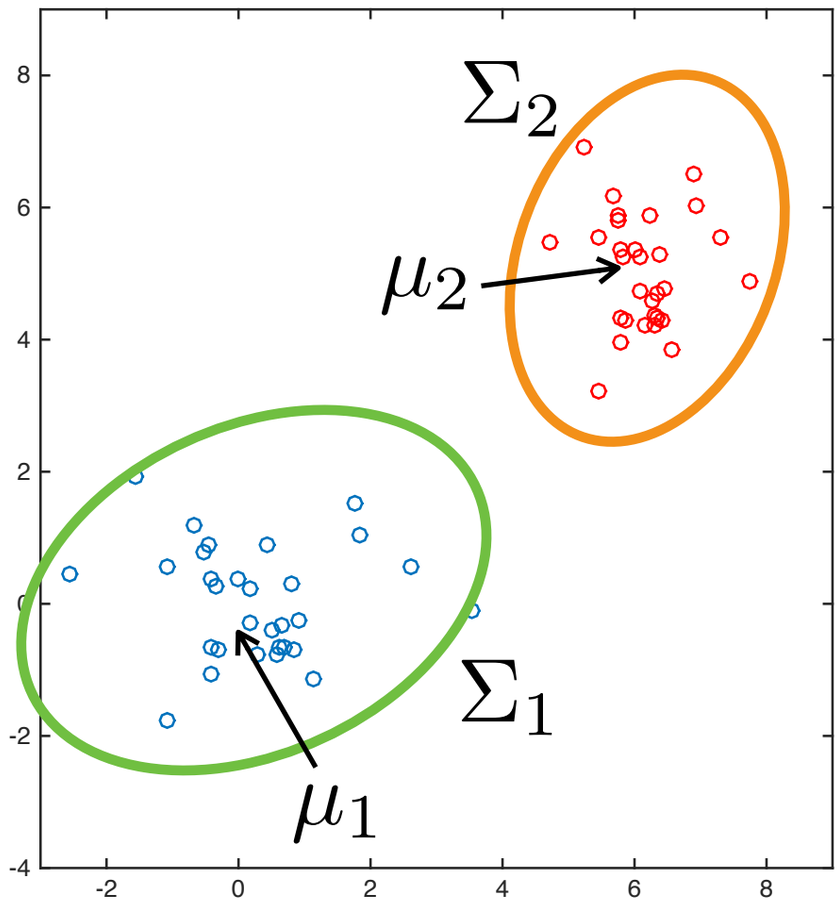
- The corresponding MM algorithm becomes

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

# 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

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

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

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