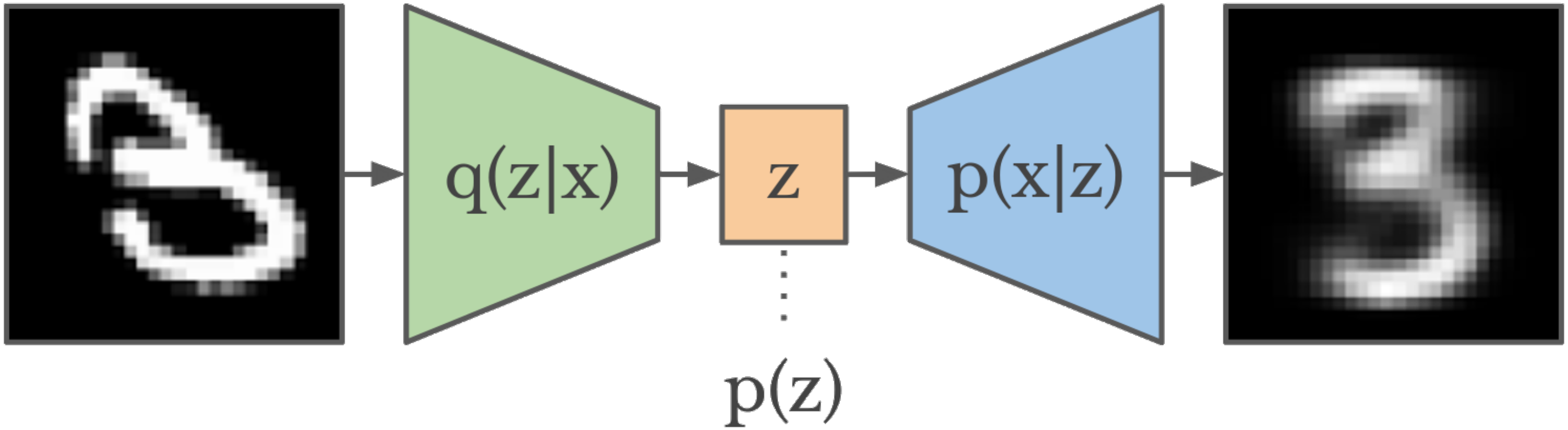
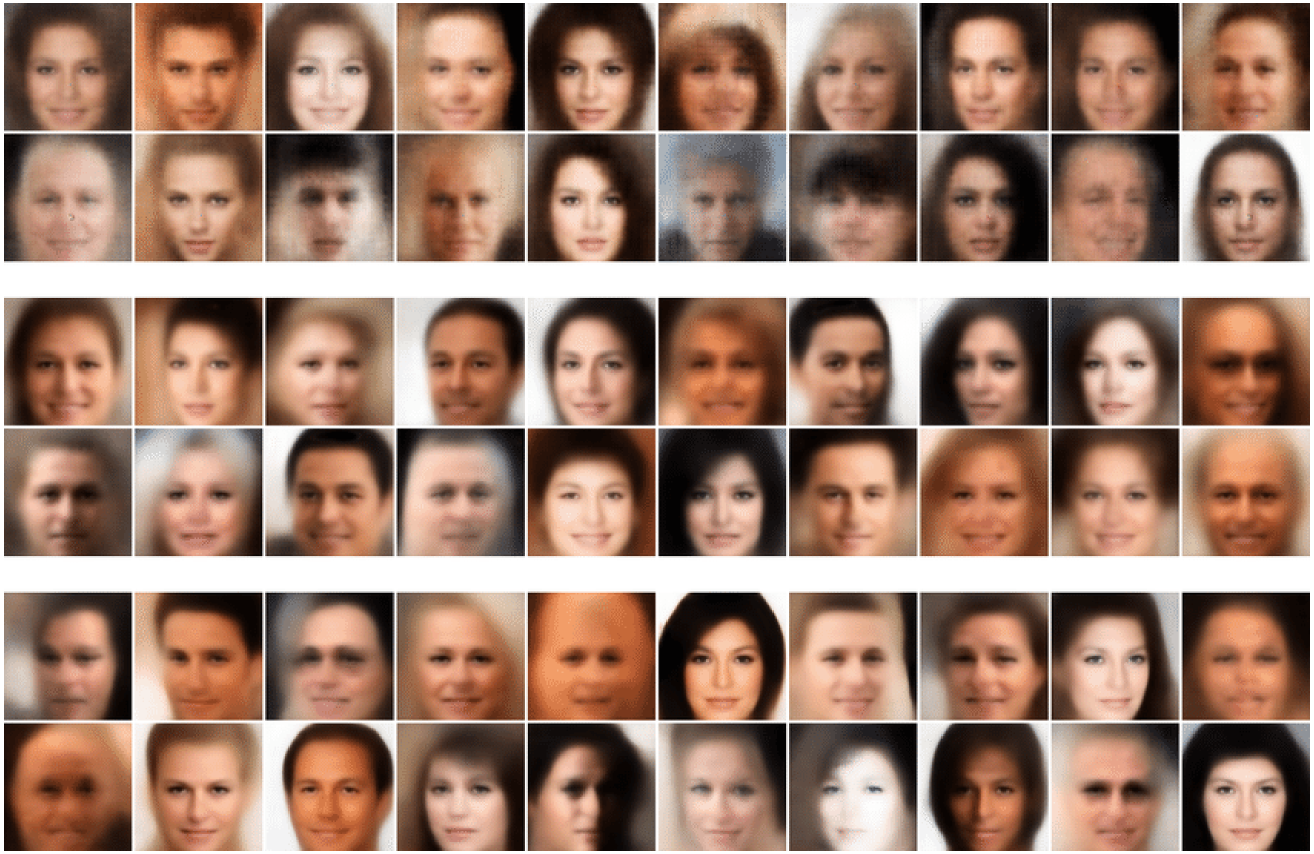
Variational auto-encoders (VAEs)

Variational autoencoder



Encoder Decoder

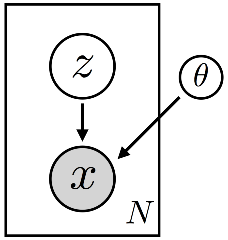
Computer generated faces



Interpolation between sampled points

EM algorithm in general

- Given a training set $\{x^1, \dots, x^{(N)}\}$ which we hypothesize to be generated from latent variables z



we wish to maximize the log-likelihood

$$
\begin{aligned}
l_\theta(\mathbf{x}) &= \sum_{i=1}^{N} \log p_\theta\left(x^{(i)}\right) \\
&= \sum_{i=1}^{N} \log \int p_\theta\left(x^{(i)}, z\right) \, \mathrm{d}z
\end{aligned}
$$

- The expectation-maximization (EM) algorithm in general is a technique for finding maximum likelihood solutions for probabilistic models with latent variables.
- In general, the *incomplete data likelihood function* $p_\theta(x)$ is hard to optimize, but the *complete data likelihood function* $p_\theta(x, z)$ is easier to work with.

Lower bound

Given any distribution $q(z)$, we have

$$
\begin{aligned}
\sum_{i=1}^{N} \log \int p_\theta\left(x^{(i)}, z\right) \mathrm{d}z &= \sum_{i=1}^{N} \log \int q(z) \frac{p_\theta\left(x^{(i)}, z\right)}{q(z)} \mathrm{d}z \\
&= \sum_{i=1}^{N} \log \mathbb{E}_{q(z)} \left[\frac{p_\theta\left(x^{(i)}, z\right)}{q(z)} \right] \\
&>= \sum_{i=1}^{N} \mathbb{E}_{q(z)} \left[\log \frac{p_\theta\left(x^{(i)}, z\right)}{q(z)} \right] = \sum_{i=1}^{N} \int q(z) \log \frac{p_\theta(x^{(i)}, z)}{q(z)} \mathrm{d}z,
\end{aligned}
$$

where the last line follows by Jensen's inequality.

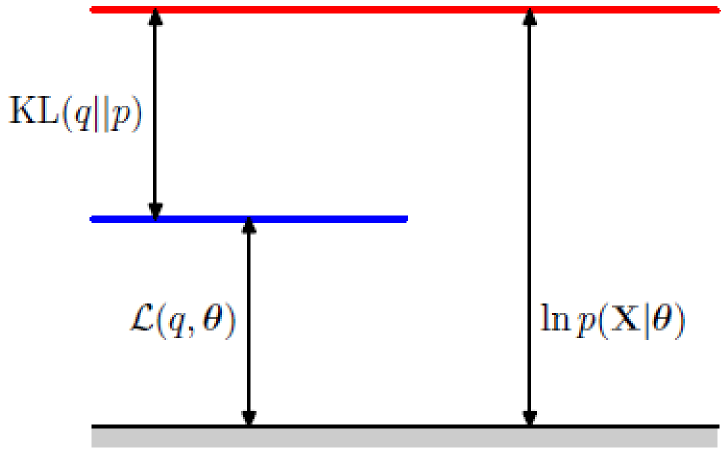
- The lower bound

$$\mathcal{L}(q,\theta) = \sum_{i=1}^{N} \int q(z) \log \frac{p_\theta\left(x^{(i)}, z\right)}{q(z)} \,\mathrm{d}z$$

holds for all distributions $q(z)$, but which one is the best?

- We have the following formula which gives the difference between the log-likelihood and the lower bound:

$$\log_\theta p\left(x^{(i)}\right) - \mathcal{L}(q,\theta) = D_{KL}\left[q(z) \,\middle|\, p_\theta\left(z \middle| x^{(i)}\right)\right].$$



- Recall that the KL-divergence is ≥ 0, and equals 0 when $q(z) = p_\theta\left(z \middle| x^{(i)}\right)$, in which case the lower bound is equal to the log-likelihood.

EM algorithm

(i) E-step: Optimize lower bound with respect to q

$$q_{t+1}(z) := \arg\max_{q} \mathcal{L}(q, \theta_t)$$

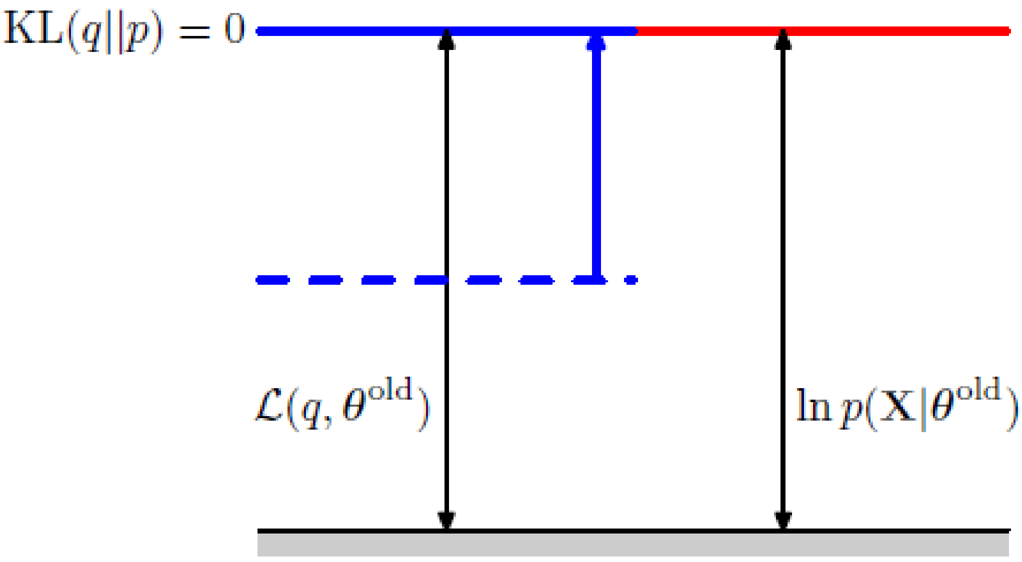
(ii) M-step: Optimize lower bound with respect to θ

$$\begin{aligned}\theta_{t+1} &:= \arg\max_{\theta} \mathcal{L}(q_{t+1}, \theta) \\ &= \arg\max_{\theta} \sum_{i=1}^{N} \int q_{t+1}(z) \log \frac{p_\theta\left(x^{(i)}, z\right)}{q_{t+1}(z)} \,\mathrm{d}z\end{aligned}$$

(iii) Go back to step (i) until the increase in $\ell_\theta(\mathbf{x})$ falls below some predetermined threshold.

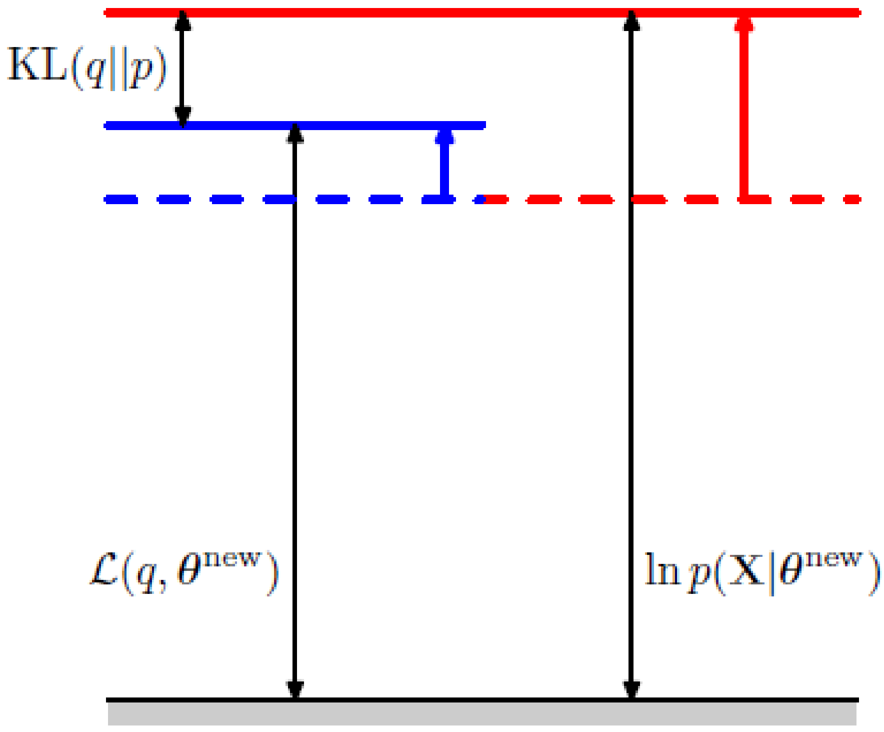
E-step

Illustration of the E step of the EM algorithm. The q distribution is set equal to the posterior distribution for the current parameter values θ^{old}, causing the lower bound to move up to the same value as the log likelihood function, with the KL divergence vanishing.



M-step

Illustration of the M step of the EM algorithm. The distribution $q(\mathbf{Z})$ is held fixed and the lower bound $\mathcal{L}(q, \boldsymbol{\theta})$ is maximized with respect to the parameter vector $\boldsymbol{\theta}$ to give a revised value $\boldsymbol{\theta}^{\text{new}}$. Because the KL divergence is nonnegative, this causes the log likelihood $\ln p(\mathbf{X}|\boldsymbol{\theta})$ to increase by at least as much as the lower bound does.



Monotone convergence theorem

Theorem

Let $\{a_n\}$ be an monotonically non-decreasing sequence; i.e. $a_{n+1} \geq a_n$ for all n. If $\{a_n\}$ is bounded above by some constant c, then the sequence converges.

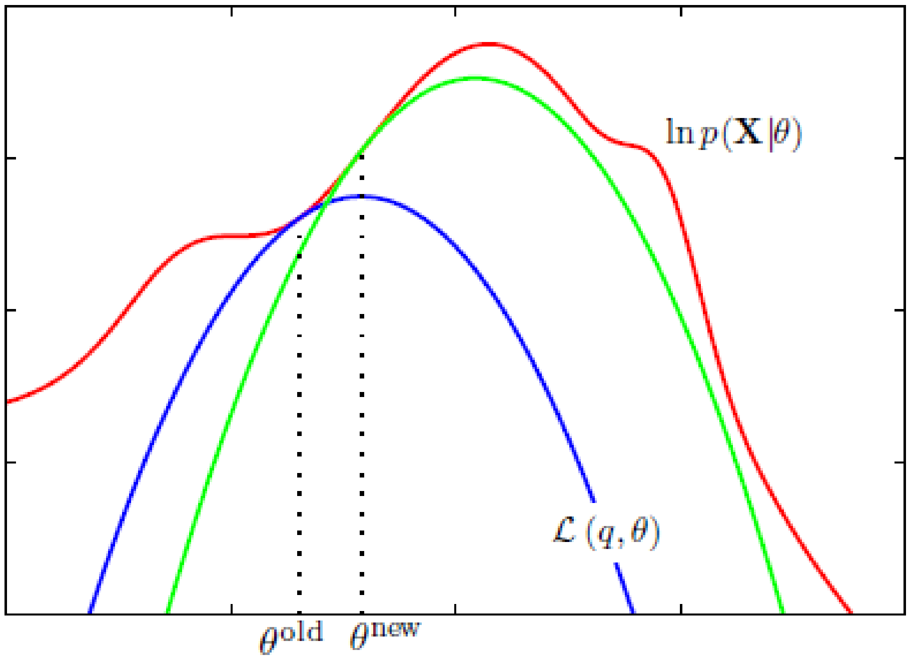
Convergence

- Note that

$$\begin{aligned}
\ell_{\theta_{t+1}}(\mathbf{x}) &\geq \sum_{i=1}^{N} \int q_{t+1}(z) \log \frac{p_{\theta_{t+1}}\left(x^{(i)}, z\right)}{q_{t+1}(z)} \, \mathrm{d}z \\
&\geq \sum_{i=1}^{N} \int q_{t+1}(z) \log \frac{p_{\theta_{t}}\left(x^{(i)}, z\right)}{q_{t+1}(z)} \, \mathrm{d}z \\
&= \ell_{\theta_t}(\mathbf{x}).
\end{aligned}$$

- The first inequality follows from the definition of the lower bound, the second follows from the M-step, and the third equality is a result of the E-step which sets $D_{KL}[q(z) \mid p_{\theta_t}(z|x_i)]$ to 0.
- Thus, we get convergence from Monotone convergence theorem since we have a monotonically non-decreasing sequence which is bounded above by 0.

Another view of EM



- Blue curve: Lower bound after E-step at previous iteration
- Green curve: Lower bound after E-step at current iteration

- In a complex model like a VAE, $p_\theta\left(z|x^{(i)}\right)$ is intractable, so we cannot directly set

$$q_{t+1}(z) := p_{\theta_t}\left(z \,\middle|\, x^{(i)}\right),$$

which also means the KL-divergence is never exactly 0.

- Instead, we approximate the conditional distribution by considering a restricted family of (parameterized) distributions for q. For VAEs, q is modeled using a neural network with parameters ϕ and the lower bound

$$\mathbb{E}_{q_\phi\left(z|x^{(i)}\right)}\left[\log \frac{p_\theta(x^{(i)}, z)}{q_\phi(z \mid x^{(i)})}\right]$$

is maximized with respect to θ and ϕ together.

- Furthermore, we have

$$\mathbb{E}_{q_\phi(z|x^{(i)})}\left[\log \frac{p_\theta(x^{(i)}, z)}{q_\phi(z \mid x^{(i)})}\right]$$
$$= \mathbb{E}_{q_\phi(z|x^{(i)})}\left[\log p_\theta(x^{(i)}|z)\right] - D_{KL}\left[q_\phi\left(z|x^{(i)}\right) \,\middle|\, p(z)\right].$$

- The second term in can be computed analytically, and the first term can be approximated by

$$\frac{1}{L}\sum_{l=1}^{L} \log p_\theta\left(x^{(i)} \mid z^{(i,l)}\right),$$

where $z^{(i,l)}$ is drawn (L times) from the distribution of Z.