Generative Models

Niels Nygaard

September 27, 2023

1 Introduction

The object of a *Generative Model* is to be able to generate samples from an unknown distribution

Consider a dataset $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N\}$ of vectors in $\mathbb{R}^d$.

It could be images formed out of pixels, each with a color defined by a 3-dimensional vector of RGB values. In this case each data vector has dimension $d = 3 \times H \times W$ where H and W are the height and width resp. measured in pixels. Or it could be vectors of data of a financial instrument or data from a scientific experiment.

Common for this situation is the assumption that the data are independent samples of some random variable X with values in $\mathbb{R}^d$, say with density function

$$p_X : \mathbb{R}^d \to \mathbb{R}$$

1.1 Reminder of some concepts of Random Variables

Random Variables

A vector valued *Random Variable* (with values in $\mathbb{R}^d$) is a function

$$\mathbf{X} : \Omega \to \mathbb{R}^d$$

where $(\Omega, \mathbb{P})$ is some fixed probability space (it doesn't really matter what it is).

A vector valued random variable has coordinate functions

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_d)$$

where each $\mathbf{X}_i : \Omega \to \mathbb{R}$ is a usual $\mathbb{R}$ valued random variable

Distribution and Density

We can express properties of this abstract function in terms of usual real functions:

- The *Distribution Function* $d_{\mathbf{X}} : \mathbb{R}^d \to [0,1]$

  $d_{\mathbf{X}}(a_1, a_2, \ldots, a_d) = \mathbb{P}(\{\omega \in \Omega \,|\, \mathbf{X}(\omega) \in \{(t_1, t_2, \ldots, t_d) \in \mathbb{R}^d \,|\, t_1 < a_1, t_2 < a_2, \ldots, t_d < a_d\}\}$

  In other words

  $d_{\mathbf{X}}(a_1, a_2, \ldots, a_d) = \mathbb{P}\left(\mathbf{X}^{-1}(\{(t_1, t_2, \ldots, t_d) \in \mathbb{R}^d \,|\, t_1 < a_1, t_2 < a_2, \ldots, t_d < a_d\}\right)$

- The *Density Function* $p_{\mathbf{X}} : \mathbb{R}^d \to \mathbb{R}_+$

$$p_{\mathbf{X}} = \frac{\partial}{\partial t_1}\frac{\partial}{\partial t_2}\cdots\frac{\partial}{\partial t_d} d_{\mathbf{X}}$$

If $A \subset \mathbb{R}^d$ is a reasonable (= Borel measurable) subset then we can express the probability that a value of $\mathbf{X}$ lands in A, as a usual d-dimensional integral

$$\mathbb{P}(\mathbf{X}^{-1}(A)) = \int_A p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}$$

Hence

$$\int_{\mathbb{R}^d} p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} = 1$$

Marginalization

We can recover the usual density functions of each of the coordinate functions by *marginalization* i.e. integrating out all the variables except one

$$p_{\mathbf{X}_i}(\mathbf{x}_i) = \int_{\mathbb{R}^{d-1}} p_{\mathbf{X}}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i, \ldots, \mathbf{x_d}) d\mathbf{x}_1, \ldots, \widehat{d\mathbf{x}_i} \ldots \mathbf{x}_d$$

Mean and Covariance Matrix

The *mean* $\mu_{\mathbf{X}}$ (which is a vector in $\mathbb{R}^d$) is

$$\mu_{\mathbf{X}} = (\mu_{\mathbf{X}_1}, \mu_{\mathbf{X}_2}, \ldots, \mu_{\mathbf{X}_d}) = \left(\int_{\mathbb{R}^d} x_1 p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}, \int_{\mathbb{R}^d} x_2 p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}, \ldots, \int_{\mathbb{R}^d} x_d p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}\right)$$

where $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_d)$ are the coordinate functions

Remark that

$$\int_{\mathbb{R}^d} \mathbf{x}_i p_{\mathbf{X}}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d) d\mathbf{x}_1 d\mathbf{x}_2 \ldots d\mathbf{x}_d = \int_{\mathbb{R}} \mathbf{x}_i \left(\int_{\mathbb{R}^{d-1}} p_{\mathbf{X}}(\mathbf{x}_1, \ldots, \mathbf{x}_d) d\mathbf{x}_1 \ldots \widehat{d\mathbf{x}_i} \ldots d\mathbf{x}_d\right) d\mathbf{x}_i$$

and by marginalization

$$\int_{\mathbb{R}^{d-1}} p_{\mathbf{X}}(\mathbf{x}_1, \ldots, \mathbf{x}_d) d\mathbf{x}_1 \ldots \widehat{d\mathbf{x}_i} \ldots d\mathbf{x}_d = p_{\mathbf{X}_i}$$

is the density function $p_{\mathbf{X}_i}$ of the coordinate function $\mathbf{X}_i$ so

$$\mu_{\mathbf{X}_i} = \int_{\mathbb{R}} \mathbf{x}_i p_{\mathbf{X}_i}(\mathbf{x_i}) d\mathbf{x}_i$$

the usual mean of the 1-dimensional random variable $\mathbf{X}_i$

The covariance matrix Σ is defined by

$$\begin{aligned}\Sigma_{ij} &= \int_{\mathbb{R}^d} (\mathbf{x}_i - \mu_{\mathbf{X}_i})(\mathbf{x}_j - \mu_{\mathbf{X}_j}) p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathbb{R}^2} (\mathbf{x}_i - \mu_{\mathbf{X}_i})(\mathbf{x}_j - \mu_{\mathbf{X}_j}) p_{\mathbf{X}_i,\mathbf{X}_j}(\mathbf{x}_i, \mathbf{x}_j) d\mathbf{x}_i d\mathbf{x}_j\end{aligned}$$

An example of a multi-dimensional density function, is the d-variate normal distribution with mean $\mu_{\mathbf{X}}$ and covariance matrix Σ

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_{\mathbf{X}})\, \Sigma^{-1}\, (\mathbf{x} - \mu_{\mathbf{X}})^T\right)$$

Conditional Density

If $\mathbf{X}$ and $\mathbf{Y}$ are vector valued random variables of dimension d_1 and d_2, the *conditional density function* is defined as

$$p_{\mathbf{X}|\mathbf{Y}}(x_1, x_2, \ldots, x_{d_1} | y_1, y_2, \ldots, y_{d_2}) = \frac{p_{\mathbf{X},\mathbf{Y}}(x_1, x_2, \ldots, x_{d_1}, y_1, y_2, \ldots, y_{d_2})}{p_{\mathbf{Y}}(y_1, y_2, \ldots, y_{d_2})}$$

This is related to the usual definition of *conditional probability*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Indeed if $\mathbf{x} \in \mathbb{R}^{d_1}$ and $\mathbf{y} \in \mathbb{R}_{d_2}$. We take small neighborhoods $\mathcal{N}_1$ and $\mathcal{N}_2$ of volumes $\Delta\mathbf{x}$ and $\Delta\mathbf{y}$ around these points.

Then

$$\mathbb{P}(\mathbf{X} \in \mathcal{N}_1 | \mathbf{Y} \in \mathcal{N}_2) = \frac{\mathbb{P}((\mathbf{X}, \mathbf{Y}) \in \mathcal{N}_1 \times \mathcal{N}_2)}{\mathbb{P}(\mathbf{Y} \in \mathcal{N}_2)}$$

We have

$$\mathbb{P}((\mathbf{X}, \mathbf{Y}) \in \mathcal{N}_1 \times \mathcal{N}_2) \approx p_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) \Delta\mathbf{x} \Delta\mathbf{y}$$

and

$$\mathbb{P}(\mathbf{Y} \in \mathcal{N}_2) \approx p_{\mathbf{Y}}(\mathbf{y}) \Delta\mathbf{y}$$

so

$$\mathbb{P}(\mathbf{X} \in \mathcal{N}_1 | \mathbf{Y} \in \mathcal{N}_2) \approx \frac{p_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) \Delta\mathbf{x} \Delta\mathbf{y}}{p_{\mathbf{Y}}(\mathbf{y}) \Delta\mathbf{y}} = \frac{p_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) \Delta\mathbf{x}}{p_{\mathbf{Y}}(\mathbf{y})}$$

so if we let $\Delta\mathbf{y} \to 0$ we get

$$\mathbb{P}(\mathbf{X} \in \mathcal{N}_1 | \mathbf{Y} = \mathbf{y}) \approx p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})\Delta\mathbf{x}$$

We have Bayes' formula which is a trivial consequence of the definition

$$p_{\mathbf{X}|\mathbf{Y}} = \frac{p_{\mathbf{Y}|\mathbf{X}}}{p_{\mathbf{Y}}} p_{\mathbf{X}}$$

Maximum Likelihood Estimation (MLE)

Assume we have a set of data vectors $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ in $\mathbb{R}^d$ which we assume are samples of some random variable. We want to estimate the density of this random variable.

This question is too broad to be solvable so we restrict ourselves to consider a parametrized family of distributions

$$\{p_\theta\}_{\theta \in \Theta}$$

To estimate the fit of a distribution we look at the log likelihood

$$LL(\theta) = \log \prod_{\mathbf{x} \in \mathcal{D}} p_\theta(\mathbf{x}) = \sum_{\mathbf{x} \in \mathcal{D}} \log p_\theta(\mathbf{x})$$

The problem then becomes to find a parameter $\theta_0 \in \Theta$ such that

$$LL = \sum_{\mathbf{x} \in \mathcal{D}} \log p_{\theta_0}(\mathbf{x})$$

is maximal

This is *Maximum Likelihood Estimation*

To justify Maximum Likelihood we can give a heuristic argument.

Assume our data are samples of some density p_{data}. If $\mathbf{X}_\theta$ is a random variable with density p_θ.

We want to find θ that maximizes the probability that the data set $\mathcal{D}$ are samples from $\mathbf{X}_\theta$

For each $\mathbf{x} \in \mathcal{D}$ we choose a small neighborhood $\mathcal{N}_\mathbf{x}$ around $\mathbf{x}$

The probability that $\mathbf{X}_\theta(\omega) \in \mathcal{N}_\mathbf{x}$ is

$$\mathbb{P}(X_\theta(\omega) \in \mathcal{N}_\mathbf{x})) = \int_{\mathcal{N}_\mathbf{x}} p_\theta(\mathbf{t}) d\mathbf{t} \approx p_\theta(\mathbf{x}) vol(\mathcal{N}_\mathbf{x})$$

Given a sequence of elements $\omega_1, \omega_2, \ldots, \omega_N$ the probability that

$$\mathbf{X}_\theta(\omega_i) \approx \mathbf{x}_i \text{ for } i = 1, 2, \ldots, N$$

is

$$\approx \prod_i p_\theta(\mathbf{x}_i) vol(\mathcal{N}_{\mathbf{x}_i}) = \prod_{\mathbf{x}\in\mathcal{D}} p_\theta(\mathbf{x}) \prod_{\mathbf{x}\in\mathcal{D}} vol(\mathcal{N}_{\mathbf{x}_i})$$

We we want this to be maximal for any choice of the neighborhoods $\{\mathcal{N}_{\mathbf{x}}\}$ so we want to maximize $\prod_{\mathbf{x}\in\mathcal{D}} p_\theta(\mathbf{x})$

A case where we can solve MLE explicitly the case where the family of distributions are Gaussians

$$p_\theta(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_{\mathbf{X}})\,\Sigma^{-1}\,(\mathbf{x}-\mu_{\mathbf{X}})^T\right)$$

Here the parameter $\theta = (\mu, \Sigma)$

Differentiating with respect to μ we get

$$0 = \Sigma^{-1} \sum_{\mathbf{x}\in\mathcal{D}} (\mathbf{x} - \mu)$$

Solving for μ we get

$$\mu = \frac{1}{N} \sum_{\mathbf{x}\in\mathcal{D}} \mathbf{x}$$

Differentiating with respect to Σ we get

$$\begin{aligned} 0 &= \nabla_\Sigma(-\log (2\pi)^{d/2}\sqrt{\det \Sigma} - \frac{1}{2}(\mathbf{x}-\mu)\Sigma^{-1}(\mathbf{x}-\mu)^T) \\ &= \nabla_\Sigma(-\frac{1}{2}\log\det\Sigma - \frac{1}{2}(\mathbf{x}-\mu)\Sigma^{-1}(\mathbf{x}-\mu)^T \end{aligned}$$

Using formulas for differentiating matrices (these formulas hold for symmetric positive definite matrices)

- $\nabla_\Sigma \log\det\Sigma = \Sigma^{-1}$
- $\nabla_\Sigma(a\Sigma^{-1}a^T) = -\Sigma^{-1}a^Ta\Sigma^{-1}$

we get

$$0 = \sum_{\mathbf{x}\in\mathcal{D}} (\Sigma^{-1} - \Sigma^{-1}(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T\Sigma^{-1}) = N\Sigma^{-1} - \Sigma^{-1}\sum_{\mathbf{x}\in\mathcal{D}} (\mathbf{x}-\mu)^T(\mathbf{x}-\mu)\,\Sigma^{-1}$$

so

$$\Sigma = \frac{1}{N}\sum_{\mathbf{x}\in\mathcal{D}} (\mathbf{x}-\mu)^T(\mathbf{x}-\mu)$$

Remark that $(\mathbf{x}-\mu)$ is $1\times d$ dimensional so $\sum_{\mathbf{x}\in\mathcal{D}} (\mathbf{x}-\mu)^T(\mathbf{x}-\mu)$ is $d\times d$

1.2 Kullback-Liebler Divergence and Variational Inference

Consider a model where we are given a dataset $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots . \mathbf{x}_N\}$ and we assume the data are generated by some parametrized distribution with unknown parameters θ,

$$p_\theta(\mathbf{x})$$

In *Bayesian Inference* we consider the parameters themselves as random variables and so we shall write

$$p_\theta(\mathbf{x}) \text{ as } p(\mathbf{x}|\theta)$$

and we can try to find the θ that maximizes the *posterior probability* $p_{post}(\theta|\mathcal{D})$, the distribution of the parameters given the dataset $\mathcal{D}$ (this is known as the MAP= the Maximal Aposteriori Probability)

We choose a prior distribution $p_{prior}(\theta)$ of the parameters (in practice this choice does not matter too much)

For a given sample θ from the prior we have the parametrized distribution of the data $p(\mathbf{x}|\theta)$ and the *likelihood* of $\mathcal{D}$ (conditional on θ)

$$p(\mathcal{D}|\theta) = \prod_{\mathbf{x}\in\mathcal{D}} p(\mathbf{x}|\theta)$$

Using Bayes' formula

$$p_{post}(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)}{p(\mathcal{D})} p_{prior}(\theta)$$

The problem here is again the intractability of an integral

$$p(\mathcal{D}) = \int p(\mathcal{D}|\theta)p(\theta)d\theta$$

This is simply marginalizing out θ since $p(\mathcal{D}|\theta)p(\theta)$ is the joint distribution $p(\mathcal{D}, \theta)$. Even though we know $p(\mathcal{D}|\theta)$ and $p(\theta)$ the integral is generally intractable.

The idea of *Variational Inference* is to replace the posterior $p_{post}(\theta|\mathcal{D})$ with some simpler distribution $q(\theta)$. We choose q from a family of parametrized distributions so $q(\theta) = q_\phi(\theta)$ and we want to find the parameters ϕ such that q_ϕ is 'close' to the posterior. The meaning of 'close' of course has to be defined.

KL-Divergence

Definition Let p and q be distributions with the same support i.e. $p(\mathbf{x}) = 0$ iff $q(\mathbf{x}) = 0$. We define the *Kullback-Liebler Divergence* by

$$D_{KL}(q||p) = -\int \log \frac{p(x)}{q(x)} q(x)dx = -\mathbb{E}_q(\log \frac{p}{q})$$

If p and q are distributions over finite sets the integral reduces to a sum

$$D_{KL}(q||p) = -\sum_x q(x) \log \frac{p(x)}{q(x)}$$

The KL-divergence is like a measure of distance between distributions, it is, however not a metric, it is not symmetric and it does not satisfy a triangle inequality which are requirements for a being a metric.

It does satisfy some weaker conditions

- $D_{KL}(q||p) \geq 0$. This follows from *Jensen's Inequality*: If f is a concave function then $\mathbb{E}(f(\mathbf{X})) \leq f(\mathbb{E}(\mathbf{X}))$ Since log is concave we get

$$\begin{aligned} D_{KL}(q||p) = -\mathbb{E}_q(\log \frac{p}{q}) &\geq -\log \mathbb{E}_q(\frac{p}{q}) = -\log \int \frac{p(x)}{q(x)} q(x)dx \\ &= -\log \int p(x)dx = -\log(1) = 0 \end{aligned}$$

- $D_{KL}(q||p) = 0$ if and only if $p = q$ a.e. . Indeed if $p = q$ a.e. then $\log \frac{p(x)}{q(x)} = \log 1 = 0$ a.e. The other direction follows from the inequality (Pinsker's inequality)

  $$TV(q, p) \leq D_{KL}(q||p)$$

  where $TV(q, p)$ is the total variation distance

  $$TV(q, p) = \sup_A (\int_A |p - q|)$$

  so if $D_{KL}(q||p) = 0$, $TV(q, p) = 0$ and $\int_A p - q = 0$ for all A implies $p = q$ a.e.

In spite of the fact that the KL divergence is not a real metric it is very useful for approximating distributions.

Example Consider again the Maximum Likelihood estimation with a data set $\mathcal{D}$.

Let $\mathcal{D}$ be samples from some distribution p_{data} and let p_θ be our parametrized distribution. We want to find θ that maximizes

$$LL(\theta) = \sum_{\mathbf{x}\in\mathcal{D}} \log p_\theta(\mathbf{x})$$

or equivalently the average

$$\frac{1}{|\mathcal{D}|}\sum_{\mathbf{x}\in\mathcal{D}} \log p_\theta(\mathbf{x})$$

We are averaging over a set of samples from p_{data} so

$$\frac{1}{|\mathcal{D}|}\sum_{\mathbf{x}\in\mathcal{D}} \log p_\theta(\mathbf{x}) \approx \mathbb{E}_{p_{data}}(\log p_\theta)$$

Now

$$-\mathbb{E}_{pdata}(\log p_\theta) = -\mathbb{E}_{p_{data}}\left(\log \frac{p_\theta}{p_{data}}\right) - \mathbb{E}_{p_{data}}(\log p_{data})$$

The last term does not depend on θ (it is called the *Shannon Entropy*) so minimizing $-\mathbb{E}_{p_{data}}(\log p_\theta)$ (approximately equivalent to maximizing $\frac{1}{|\mathcal{D}|}\sum_{\mathbf{x}\in\mathcal{D}} \log p_\theta(\mathbf{x})$) is equivalent to minimizing

$$-\mathbb{E}_{p_{data}}\left(\log \frac{p_\theta}{p_{data}}\right) = D_{KL}(p_{data}||p_\theta)$$

Thus the θ that maximizes $LL(\theta)$ also minimizes the KL-divergence between p_{data} and p_θ

Variational Inference

In the general Bayesian situation we have a random variable $\mathbf{X}$, depending on a *latent variable* (meaning hidden) $\mathbf{Z}$ with density $p_{\mathbf{Z}}(\mathbf{z})$ where we know $p_{\mathbf{X}|\mathbf{Z}}$ and $p_{\mathbf{Z}}$. The joint density $p_{\mathbf{X},\mathbf{Z})}(\mathbf{x},\mathbf{z}) = p_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}|\mathbf{z})p_{\mathbf{Z}}(\mathbf{z})$. The posterior distribution

$$p_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}) = \frac{p_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}|\mathbf{z})}{p_{\mathbf{X}}(\mathbf{x})}p_{\mathbf{Z}}(\mathbf{z})$$

is intractable because of the intractibility of

$$p_{\mathbf{X}}(\mathbf{x}) = \int p_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}|\mathbf{z})p_{\mathbf{Z}}(\mathbf{z})d\mathbf{z}$$

The idea of Variational Inference is to approximate the posterior distribution with a parametrized distribution $q_\phi(\mathbf{z})$ from a family of distributions e.g. Gaussians, which are parametrized by a parameter vector ϕ (q_ϕ can also

be conditional on $\mathbf{X}$ so instead of $q_\phi(\mathbf{z})$ we are looking for a conditional density $q_\phi(\mathbf{z}|\mathbf{x})$ We want q_ϕ to be as close to $p(\mathbf{z}|\mathbf{x})$ as possible in the sense that $D_{KL}(q_\phi(\mathbf{z})||p(\mathbf{z}||\mathbf{x}))$ (or $D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x}))$) is minimized.

The problem of finding q now becomes a problem of finding ϕ_0

$$\phi_0 = \underset{\phi}{\operatorname{argmin}}\ D_{KL}(q_\phi(\mathbf{z})||p(\mathbf{z}|\mathbf{x}))$$

Of course this seems to be equally hard because to compute the KL-divergence it seems that we need to know the distribution $p(\mathbf{z}|\mathbf{x})$ which we are trying to approximate.

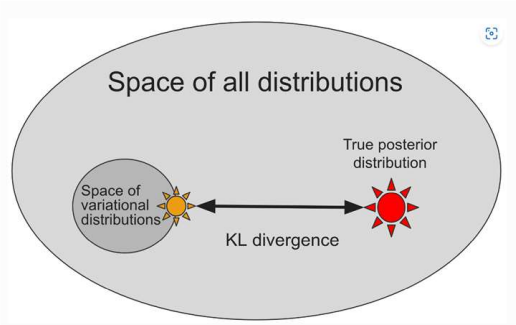

Figure 1:

We can rewrite the KL-divergence

$$\begin{aligned}
D_{KL}(q_\phi(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) &= -\mathbb{E}_{q_\phi}\left(\log \frac{p(\mathbf{z}|\mathbf{x}))}{q_\phi(\mathbf{z})}\right) \\
&= \mathbb{E}_{q_\theta}(\log q_\theta(\mathbf{z})) - \mathbb{E}_{q_\phi}(\log p(\mathbf{z}|\mathbf{x})) \\
&= \mathbb{E}_{q_\phi}(\log q_\phi(\mathbf{z})) - \mathbb{E}_{q_\phi}\left(\log \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})}\right) \\
&= \mathbb{E}_{q_\phi}(\log q_\phi(\mathbf{z})) - \mathbb{E}_{q_\phi}(\log p(\mathbf{x}|\mathbf{z})) - \mathbb{E}_{q_\phi}(\log p(\mathbf{z})) + \mathbb{E}_{q_\phi}(\log p(\mathbf{x}))
\end{aligned}$$

Now

$$\mathbb{E}_{q_\phi}(\log p(\mathbf{x})) = \int q_\phi(\mathbf{z}) \log p(\mathbf{x}) d\mathbf{z} = \log p(\mathbf{x}) \int q_\phi(\mathbf{z}) d\mathbf{z} = \log p(\mathbf{x})$$

since $\log p(\mathbf{x})$ does not depend on $\mathbf{z}$ and since $\int q_\phi(\mathbf{z})d\mathbf{z} = 1$ (q_ϕ is a density)

The term $\mathbb{E}_{q_\phi}(\log q_\phi) - \mathbb{E}_{q_\phi}(\log p(\mathbf{z})) = D_{KL}(q_\phi(\mathbf{z})||p(\mathbf{z}))$ so we get

$$D_{KL}(q_\phi(\mathbf{z})||p(\mathbf{z}|\mathbf{x}) = D_{KL}(q_\phi(\mathbf{z})||p(\mathbf{z})) - \mathbb{E}_{q_\phi}(\log p(\mathbf{x}|\mathbf{z})) + \log p(\mathbf{x})$$

Since $\log p(\mathbf{x})$ is constant i.e. does not depend on q_ϕ we are reduced to minimizing the expression

$$D_{KL}(q_\phi(\mathbf{z})||p(\mathbf{z})) - \mathbb{E}_{q_\phi}(\log p(\mathbf{x}|\mathbf{z}))$$

Rearranging the terms we get

$$\log p(\mathbf{x}) = \mathbb{E}_{q_\phi}(\log p(\mathbf{x}|\mathbf{z})) - D_{KL}(q_\phi(\mathbf{z})||p(\mathbf{z})) + D_{KL}(q_\phi(\mathbf{x})||p(\mathbf{z}|\mathbf{x}))$$

Since KL-divergence is always ≥ 0 we get

$$\log p(\mathbf{x}) \geq \mathbb{E}_{q_\phi}(\log p(\mathbf{x}|\mathbf{z})) - D_{KL}(q_\phi(\mathbf{z})||p(\mathbf{z}))$$

The log probability, $\log p(\mathbf{x})$, is called the *evidence* and so

$$\mathbb{E}_{q_\phi}(\log p(\mathbf{x}|\mathbf{z})) - D_{KL}(q_\phi(\mathbf{z})||p(\mathbf{z}))$$

is a lower bound for the evidence. This expression is called the *ELBO = Evidence Lower BOund* and our problem is then reduced to *maximizing* the ELBO.

In order to solve this maximization i.e. to find

$$\underset{\phi}{\operatorname{argmax}}\ \mathbb{E}_{q_\phi}(\log p(\mathbf{x}|\mathbf{z})) - D_{KL}(q_\phi(\mathbf{z})||p(\mathbf{z}))$$

we can use gradient ascent so we need to compute

$$\nabla_\phi ELBO(\phi) = \nabla_\phi \mathbb{E}_{q_\phi}\left(\log p(\mathbf{x}|\mathbf{z}) + \log \frac{p(\mathbf{z})}{q_\phi(\mathbf{z}))}\right)$$

In the real world we have observations given as a dataset $\mathcal{D}$, the latent variable is of the form $\mathbf{z}, \theta$ where θ are parameters for the distribution $p = p_\theta$. We could also write this as

$$p_\theta(\mathbf{x}|\mathbf{z}) = p(\mathbf{x}|\mathbf{z}, \theta)$$

i.e. the parameters are also latent variables.

The Maximum Likelihood principle would be to compute the argmax of the log-likelihood

$$\theta_{max} = \underset{\theta}{\operatorname{argmax}}\ \log p_\theta(\mathcal{D})$$