Probability and Random Processes

Zhenrong
April 4, 2021

1 Probability Distributions

1.1 Gaussian Distribution

The Gaussian PDF is given by:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]. \tag{1}$$

The CDF of a Gaussian RV is given by,

$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-(z-\mu)^2/2\sigma^2\right] dz. \tag{2}$$

1.2 Multivariate Gaussian Distribution

The multivariate Gaussian distribution is a generalization of the one-dimensional Gaussian distribution to higher dimensions, which is given by,

$$p(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}. \tag{3}$$

- $\mu = \mathbb{E}(X)$.
- $\Sigma = \mathrm{Cov}(X) = \mathbb{E}\left[(X-\mu)(X-\mu)^T\right]$ (symmetric, positive semi-definite matrix).

1.2.1 Conditional and Marginal Distribution

Given multivariate Gaussian $N(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{pmatrix} \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix} \tag{4}$$

The marginal Gaussian can be given as,

$$\mu_2^m = \mu_2, \quad \Sigma_2^m = \Sigma_{22} \tag{5}$$

The conditional Gaussian can be given as,

$$\begin{aligned} \mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{aligned} \tag{6}$$

1.3 Correlated to Uncorrelated Gaussian

If the variables are correlated, then their covariance matrix Σ will not be a diagonal matrix. We can find the eigenvectors and associated eigenvalues of Σ by solving

$$\Phi^T \Sigma \Phi = \Lambda.$$

By performing the following transformation

$$\mathbf{y} = \mathbf{\Phi}^T \mathbf{x},$$

the elements of data y are uncorrlelated: its covariance, $E\left[\mathrm{yy}^T\right]$ is now a diagonal matrix, Λ.

1.4 Functions of Random Variables

1.4.1 Functional Transformations and Jacobians (for N=2)

Consider one-to-one differentiable functions $v = g(x, y)$, $w = h(x, y)$ with a unique inverse $x = \phi(v, w)$, $y = \varphi(v, w)$. Then the quantity $\tilde{J}$ is called the Jacobian of the transformation $x = \phi(v, w), y = \varphi(v, w)$.

$$\tilde{J} = \begin{vmatrix} \frac{\partial \phi}{\partial v} & \frac{\partial \phi}{\partial w} \\ \frac{\partial \varphi}{\partial v} & \frac{\partial \varphi}{\partial w} \end{vmatrix} = \frac{\partial \phi}{\partial v}\frac{\partial \varphi}{\partial w} - \frac{\partial \phi}{\partial w}\frac{\partial \varphi}{\partial v}. \tag{7}$$

The Jacobian is necessary to preserve probability measure since

$$\iint_{\varphi} f_{XY}(x, y)dxdy = \iint_{\varphi} f_{XY}(\phi(v, w), \varphi(v, w))|\tilde{J}|dvdw. \tag{8}$$

Sometimes it may be easier to deal with the original functions $v = g(x, y), w = h(x, y)$ than the inverse functions $x = \phi(v, w), y = \varphi(v, w)$. To get the desired result, we can compute,

$$J = \begin{vmatrix} \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \\ \frac{\partial h}{\partial x} & \frac{\partial h}{\partial y} \end{vmatrix} \tag{9}$$

where

$$\tilde{J}| = 1/|J \mid \text{ or } |\tilde{J}J| = 1. \tag{10}$$

1.5 Moments of Random Variables

Generally, a random variable will have many nonzero higherorder moments and it is possible to completely describe the behavior of the random variable, reconstruct its pdf from knowledge of all the moments.

The rth moment of X is defined as

$$m_r \triangleq E\left[X^r\right] = \int_{-\infty}^{\infty} x^r f_X(x)dx, \quad \text{where } r = 0, 1, 2, 3, \ldots \tag{11}$$

1.6 Moment-generating Function

The moment-generating function (MGF), if it exists, is defined by

$$\begin{aligned} M(t) &\triangleq E\left[e^{tX}\right] \\ &= \int_{-\infty}^{\infty} e^{tx} f_X(x) dx \end{aligned} \tag{12}$$

where t is a complex variable.

From Equation (12) we see that except for a sign reversal in the exponent, the MGF is the two-sided Laplace transform of the pdf for which there is a known inversion formula.

If we expand e^{tX} and take expectations, then

$$\begin{aligned} E\left[e^{tX}\right] &= E\left[1 + tX + \frac{(tX)^2}{2!} + \ldots + \frac{(tX)^n}{n!} + \ldots\right] \\ &= 1 + tm_1 + \frac{t^2}{2!}m_2 + \ldots + \frac{t^n}{n!}m_n + \ldots . \end{aligned}$$

If $M(t)$ does exist, computing any moment is easily obtained by differentiation.

$$m_k = M^{(k)}(0) = \frac{d^k}{dt^k}(M(t))\bigg|_{t=0} \qquad k = 0, 1, \ldots$$

The moment-generating function can be used to solve problems involving the computation of the sums of independent random variables since Laplace transform of a convolution product is the product of the individual transforms. If $Z = X_1 + \ldots + X_N$, where $X_i, i = 1, \ldots, N$, are independent random variables, the pdf of Z is furnished by

$$f_Z(z) = f_{X_1}(z) * f_{X_2}(z) * \ldots * f_{X_N}(z)$$

which is the repeated convolution product.

1.7 The law of large numbers and the central limit theorem

Suppose $X_1, X_2, \ldots, X_n$ are independent random variables with the same distribution or i.i.d. Let $\bar{X}_n$ be the average of $X_1, \ldots, X_n$

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- Law of large number: As n grows, the probability that $\bar{X}_n$ is close to μ goes to 1.
- Central limit theorem: As n grows, the distribution of $\bar{X}_n$ converges to the normal distribution $N\left(\mu, \sigma^2/n\right)$.

1.8 Bayes' Theorem

Let $A_i, i = 1, \ldots, n$, be a set of disjoint and exhaustive events defined on a probability space P. Then, $\bigcup_{i=1}^{n} A_i = \Omega, A_i A_j = \phi$ for all $i \neq j$. With B any event defined on P with $P[B] > 0$ and $P\left[A_i\right] \neq 0$ for all i

$$P\left[A_j \mid B\right] = \frac{P\left[B \mid A_j\right] P\left[A_j\right]}{\sum_{i=1}^{n} P\left[B \mid A_i\right] P\left[A_i\right]}. \tag{13}$$

2 Random Process

2.1 Introduction

Random process can be thought of as a family of jointly distributed random variables indexed by time t or n. For example, for random process $X(t) = A\sin(\omega t + \phi)$, the amplitude A, frequency ω and phase ϕ are all random variables. The value $X(t_1)$ at some specific time t_1 is also a random variable. In another point of view, a random process is a set of signals, for which the outcome of the probabilistic experiment could be any of the waveforms. Each waveform is deterministic, but the process is probabilistic or random because it is not known a priori which waveform will be generated by the probabilistic experiment. The first and second moments of the process is summarized in the following functions

$$\text{Mean:} \quad \mu_X(t_i) = E[X(t_i)], \tag{14}$$

$$\text{Auto-correlation: } R_{XX}(t_i, t_j) = E[X(t_i)X(t_j)] \text{ , and} \tag{15}$$

$$\begin{aligned}\text{Auto-covariance: } C_{XX}(t_i, t_j) &= E[(X(t_i) - \mu_X(t_i))(X(t_j) - \mu_X(t_j))] \\ &= R_{XX}(t_i, t_j) - \mu_X(t_i)\mu_X(t_j).\end{aligned} \tag{16}$$

Cross-moment functions:

$$\text{Cross-correlation: } R_{XY}(t_i, t_j) = E[X(t_i)Y(t_j)], \text{ and} \tag{17}$$

$$\begin{aligned}\text{Cross-covariance: } C_{XY}(t_i, t_j) &= E[(X(t_i) - \mu_X(t_i))(Y(t_j) - \mu_Y(t_j))] \\ &= R_{XY}(t_i, t_j) - \mu_X(t_i)\mu_Y(t_j).\end{aligned} \tag{18}$$

If $C_{XY}(t_1, t_2) = 0$ for all t_1, t_2, we say that the processes $X(t)$ and $Y(t)$ are uncorrelated.

2.2 Strict-sense Stationary

In general, we would expect that the joint PDFs associated with the random variables obtained by sampling a random process at an arbitrary number k of arbitrary times will be time-dependent, i.e., the joint PDF

$$f_{X(t_1),\cdots,X(t_k)}(x_1, \cdots, x_k) \tag{19}$$

will depend on the specific values of $t_1, \cdots, t_k$. If all the joint PDFs stay the same under arbitrary time shifts, i.e., if

$$f_{X(t_1),\cdots,X(t_k)}(x_1, \cdots, x_k) = f_{X(t_1+\tau),\cdots,X(t_k+\tau)}(x_1, \cdots, x_k) \tag{20}$$

for arbitrary τ, then the random process is said to be strict-sense stationary (SSS).

2.3 Wide-sense Stationary

For simplicity, we seek for a less restricted type of stationary. Specifically, if the mean value $\mu_X(t_i)$ is independent of time and the autocorrelation $R_{XX}(t_i, t_j)$ or equivalently the autocovariance $C_{XX}(t_i, t_j)$ is dependent only on the time difference $(t_i - t_j)$, then the process is said to be wide-sense stationary (WSS). For a WSS random process $X(t)$, we have

$$\begin{aligned} \mu_X(t) &= \mu_X, \\ R_{XX}(t_1, t_2) &= R_{XX}(t_1 + \alpha, t_2 + \alpha) \text{ for every } \alpha \\ &= R_{XX}(t_1 - t_2, 0). \end{aligned} \tag{21}$$

(Note that for a Gaussian process (i.e., a process whose samples are always jointly Gaussian) WSS implies SSS, because jointly Gaussian variables are entirely determined by the their joint first and second moments.)

Two random processes $X(t)$ and $Y(t)$ are jointly WSS if their first and second moments (including the cross-covariance) are stationary. In this case we use the notation $R_{XY}(\tau)$ to denote $E[X(t+\tau)Y(t)]$.

2.4 Some Properties of WSS Correlation and Covariance Functions

For real-valued WSS processes $x(t)$ and $y(t)$ the correlation and covariance functions have the following symmetry properties:

$$\begin{aligned} R_{xx}(\tau) &= R_{xx}(-\tau), \quad C_{xx}(\tau) = C_{xx}(-\tau) \\ R_{xy}(\tau) &= R_{yx}(-\tau), \quad C_{xy}(\tau) = C_{yx}(-\tau) \end{aligned}$$

2.5 The Effect of LTI Systems on WSS Processes

For an LTI system whose impulse response is $h(t)$, the output $y(t)$ is given by the convolution

$$y(t) = \int_{-\infty}^{+\infty} h(v)x(t-v)dv = \int_{-\infty}^{+\infty} x(v)h(t-v)dv \tag{22}$$

for any specific input $x(t)$ for which the convolution is well-defined.

Taking the expected value of both sides of (22) , we find

$$\begin{aligned} E[y(t)] &= E\left\{\int_{-\infty}^{+\infty} h(v)x(t-v)dv\right\} \\ &= \int_{-\infty}^{+\infty} h(v)E[x(t-v)]dv \\ &= \int_{-\infty}^{+\infty} h(v)\mu_x dv \\ &= \mu_x \int_{-\infty}^{+\infty} h(v)dv \\ &= H(j0)\mu_x = \mu_y. \end{aligned} \tag{23}$$

Next consider the cross-correlation between output and input:

$$\begin{aligned} E\{y(t+\tau)x(t)\} &= E\left\{\left[\int_{-\infty}^{+\infty} h(v)x(t+\tau-v)dv\right]x(t)\right\} \\ &= \int_{-\infty}^{+\infty} h(v)E\{x(t+\tau-v)x(t)\}dv. \end{aligned} \tag{24}$$

Since $x(t)$ is WSS, $E\{x(t+\tau-v)x(t)\} = R_{xx}(\tau-v)$, so

$$\begin{aligned} E\{y(t+\tau)x(t)\} &= \int_{-\infty}^{+\infty} h(v)R_{xx}(\tau-v)dv \\ &= h(\tau) * R_{xx}(\tau) \\ &= R_{yx}(\tau). \end{aligned} \tag{25}$$

The cross-correlation depends only on the lag τ between the sampling instants of the output and input processes. Also, this cross-correlation between the output and input is deterministically related to the autocorrelation of the input. We can also conclude that

$$R_{xy}(\tau) = R_{yx}(-\tau) = R_{xx}(-\tau) * h(-\tau) = R_{xx}(\tau) * h(-\tau). \tag{26}$$

Next we consider the autocorrelation of the output $y(t)$:

$$\begin{aligned} E\{y(t+\tau)y(t)\} &= E\left\{\left[\int_{-\infty}^{+\infty} h(v)x(t+\tau-v)dv\right]y(t)\right\} \\ &= \int_{-\infty}^{+\infty} h(v)\underbrace{E\{x(t+\tau-v)y(t)\}}_{R_{xy}(\tau-v)}dv \\ &= \int_{-\infty}^{+\infty} h(v)R_{xy}(\tau-v)dv \\ &= h(\tau) * R_{xy}(\tau) \\ &= R_{yy}(\tau). \end{aligned} \tag{27}$$

Putting this together with the earlier results, we conclude that $x(t)$ and $y(t)$ are jointly WSS.

The corresponding result for covariances is

$$C_{yy}(\tau) = h(\tau) * C_{xy}(\tau). \tag{28}$$

Combining (27) with (26) ,

$$R_{yy}(\tau) = R_{xx}(\tau) * \underbrace{h(\tau) * h(-\tau)}_{h(\tau)*h(-\tau)\triangleq\bar{R}_{hh}(\tau)} = R_{xx}(\tau) * \bar{R}_{hh}(\tau). \tag{29}$$

The function $\bar{R}_{hh}(\tau)$ is typically referred to as the deterministic autocorrelation function of $h(t)$, and is given by

$$\bar{R}_{hh}(\tau) = h(\tau) * h(-\tau) = \int_{-\infty}^{+\infty} h(t+\tau)h(t)dt. \tag{30}$$

For the covariance function version of (29) , we have

$$C_{yy}(\tau) = C_{xx}(\tau) * \underbrace{h(\tau) * h(-\tau)}_{h(\tau)*h(-\tau)\triangleq \bar{R}_{hh}(\tau)} = C_{xx}(\tau) * \bar{R}_{hh}(\tau). \tag{31}$$

3 Application

3.1 Linear Pridection of Random Process

Consider a discrete-time process $x[n]$ and we limit the prediction strategy to a linear one, i.e., with $\hat{x}\left[n_0+m\right]$ denoting the predicted value, we restrict $\hat{x}\left[n_0+m\right]$ to be of the form

$$\hat{x}\left[n_0+m\right] = ax\left[n_0\right] + b, \tag{32}$$

and choose a and b to minimize

$$\epsilon = E\left\{\left(x\left[n_0+m\right] - \hat{x}\left[n_0+m\right]\right)^2\right\}. \tag{33}$$

To minimize ϵ we set to zero its partial derivative with respect to each of the two parameters and solve for the parameter values. The resulting equations are

$$\begin{aligned} E\left\{\left(x\left[n_0+m\right] - ax\left[n_0\right] - b\right)x\left[n_0\right]\right\} = E\left\{\left(x\left[n_0+m\right] - \widehat{x}\left[n_0+m\right]\right)x\left[n_0\right]\right\} = 0 \\ E\left\{x\left[n_0+m\right] - ax\left[n_0\right] - b\right\} = E\left\{x\left[n_0+m\right] - \widehat{x}\left[n_0+m\right]\right\} = 0 \end{aligned} \tag{34}$$

Carrying out the multiplications and expectations in the preceding equations results in the following equations.

$$\begin{aligned} R_{xx}\left[n_0+m, n_0\right] - aR_{xx}\left[n_0, n_0\right] - b\mu_x\left[n_0\right] = 0, \\ \mu_x\left[n_0+m\right] - a\mu_x\left[n_0\right] - b = 0. \end{aligned} \tag{35}$$

If we assume that the process is WSS so that $R_{xx}\left[n_0+m, n_0\right] = R_{xx}[m], R_{xx}\left[n_0, n_0\right] = R_{xx}[0]$, and also assume that it is zero mean, $(\mu_x = 0)$, then equations (35) reduce to

$$\begin{aligned} a &= R_{xx}[m]/R_{xx}[0], \\ b &= 0. \end{aligned} \tag{36}$$

so that

$$\widehat{x}\left[n_0+m\right] = \frac{R_{xx}[m]}{R_{xx}[0]}x\left[n_0\right]. \tag{37}$$

If the process is not zero mean, then

$$\widehat{x}\left[n_0+m\right] = \mu_x + \frac{C_{xx}[m]}{C_{xx}[0]}\left(x\left[n_0\right] - \mu_x\right). \tag{38}$$

3.2 Linear FIR Filter

Consider a discrete-time signal $s[n]$ that has been corrupted by additive noise $d[n]$. The received signal $r[n]$ is then

$$r[n] = s[n] + d[n]. \tag{39}$$

Assume that both $s[n]$ and $d[n]$ are zero-mean random processes and are uncorrelated. If $h[n]$ is a FIR filter of length L, then

$$\widehat{s}[n] = \sum_{k=0}^{L-1} h[k]r[n-k]. \tag{40}$$

We would determine the filter coefficients $h[k]$ to minimize the mean square error between $\widehat{s}[n]$ and $s[n]$, i.e., minimize ϵ given by

$$\begin{aligned} \epsilon &= E(s[n] - \widehat{s}[n])^2 \\ &= E\left(s[n] - \sum_{k=0}^{L-1} h[k]r[n-k]\right)^2. \end{aligned} \tag{41}$$

To determine h, we set $\frac{\partial \epsilon}{\partial h[m]} = 0$ for each of the L values of m. Taking this derivative, we get

$$\begin{aligned} \frac{\partial \epsilon}{\partial h[m]} &= -E\left\{2\left(s[n] - \sum_k h[k]r[n-k]\right)r[n-m]\right\} \\ &= -E\{2(s[n] - \widehat{s}[n])r[n-m]\} \\ &= 0. \quad m = 0, 1, \cdots, L-1 \end{aligned} \tag{42}$$

Carrying out the multiplications in the above equations and taking expectations results in

$$\sum_{k=0}^{L-1} h[k]R_{rr}[m-k] = R_{sr}[m], \quad m = 0, 1, \cdots, L-1 \tag{43}$$

Equation (43) constitute L equations that can be solved for the L parameters $h[k]$. With $r[n] = s[n] + d[n]$, it is straightforward to show that $R_{sr}[m] = R_{ss}[m] + R_{sd}[m]$ and since we assumed that $s[n]$ and $d[n]$ are uncorrelated, then $R_{sd}[m] = 0$. Similarly, $R_{rr}[m] = R_{ss}[m] + R_{dd}[m]$.

3.3 Expectation-Maximization for Gaussian Mixture Model

Expectation-maximization (EM) is a method of numerically determining the MLE. Take Gaussian mixture model for example, the parameter estimation problem is given by

$$\begin{aligned} \Theta_{MLE} &= \arg\max \mathcal{L}(\Theta \mid X) \\ &= \arg\max_{\Theta} \left(\sum_{i=1}^{\theta} \log \sum_{l=1}^{k} \alpha_l \mathcal{N}\left(X \mid \mu_l, \Sigma_l\right)\right), \end{aligned} \tag{44}$$

with constraint $\sum \alpha_l = 1$. To maximize the objective function we need to set its partial derivative to 0. However, in this case the analytical solution doesn't exist. By utilizing EM algorithm in GMM, the analytical solution is easily derived in each iteration.

For each iteration of the EM algorithm, we perform

$$\Theta^{(g+1)} = \arg\max_{\theta} \left(\int_z \log(p(X, Z \mid \theta)) p\left(Z \mid X, \Theta^{(g)}\right) \right) dz, \tag{45}$$

with $p(X, Z \mid \Theta)$

$$p(X, Z \mid \Theta) = \prod_{i=1}^{n} p\left(x_i, z_i \mid \Theta\right) = \prod_{i=1}^{n} \underbrace{p\left(x_i \mid z_i, \Theta\right)}_{\mathcal{N}\left(\mu_{z_i}, \Sigma_{z_i}\right)} \underbrace{p\left(z_i \mid \Theta\right)}_{\alpha_{z_i}} = \prod_{i=1}^{n} \alpha_{z_i} \mathcal{N}\left(\mu_{z_1}, \Sigma_{z_i}\right), \tag{46}$$

and $p(Z \mid X, \Theta)$

$$p(Z \mid X, \Theta) = \prod_{i=1}^{n} p\left(z_i \mid x_i, \Theta\right) = \prod_{i=1}^{n} \frac{\alpha_{z_1} \mathcal{N}\left(\mu_{z_i}, \Sigma_{z_1}\right)}{\sum_{l=1}^{k} \alpha_l \mathcal{N}\left(\mu_l, \Sigma_l\right)}. \tag{47}$$

Now we compute $Q\left(\Theta, \Theta^{(g)}\right)$,

$$\begin{aligned} Q\left(\Theta, \Theta^{(g)}\right) &= \sum_{l=1}^{k} \sum_{i=1}^{n} \ln\left[\alpha_l \mathcal{N}\left(x_i \mid \mu_l, \Sigma_l\right)\right] p\left(l \mid x_i, \Theta^{(g)}\right) \\ &= \sum_{l=1}^{k} \sum_{i=1}^{n} \ln\left(\alpha_l\right) p\left(l \mid x_i, \Theta^{(g)}\right) + \sum_{l=1}^{k} \sum_{i=1}^{n} \ln\left[\mathcal{N}\left(x_i \mid \mu_l, \Sigma_l\right)\right] p\left(l \mid x_i, \Theta^{(g)}\right) \end{aligned} \tag{48}$$

Taking derivative with respect to the parameters, the solution can be derived as

$$\alpha_l^{(g+1)} = \frac{1}{N} \sum_{i=1}^{N} p\left(l \mid x_l, \Theta^{(g)}\right) \tag{49}$$

$$\mu_l^{(g+1)} = \frac{\sum_{i=1}^{N} x_l p\left(l \mid x_i, \Theta^{(g)}\right)}{\sum_{i=1}^{N} p\left(l \mid x_i, \Theta(g)\right)} \tag{50}$$

$$\Sigma_l^{(g+1)} = \frac{\sum_{i=1}^{N} \left[x_i - \mu_I^{(i+1)}\right] \left[x_i - \mu_l^{(i+1)}\right]^T p\left(l \mid x_i, \Theta^{(g)}\right)}{\sum_{i=1}^{N} p\left(l \mid x_i, \Theta^{(g)}\right)} \tag{51}$$

3.4 Varitional Inference

The aiming of varitional inference is using a simple distribution i.e., $q(Z)$ to approximation a target distribution i.e., $p(Z|X)$. Among considerations, we are interested in how to

evaluate the similarity between two distributions. In this case we use KL divergence as our measurement of similarity,

$$KL(q|p) = \int q(Z) \ln \left\{ \frac{q(Z)}{p(Z \mid X)} \right\} dZ. \tag{52}$$

By introducing KL divergence, our objective function is derived as

$$q^*(Z) = \arg\min_{q(Z)} KL(q|p). \tag{53}$$

The idea of varitional inference is to find a equivalent form of equation (53) to simply the calculation. Starting with the factorization of $P(X)$, we have the following results,

$$p(X) = \frac{p(X, Z)}{p(Z \mid X)}, \tag{54}$$

$$\begin{aligned} \ln p(X) &= \ln p(X, Z) - \ln p(Z \mid X) \\ &= \ln \tfrac{p(X,Z)}{q(Z)} - \ln \tfrac{p(Z|X)}{q(Z)}, \end{aligned} \tag{55}$$

$$\begin{aligned} \ln p(X) &= \int q(Z) \ln \left\{ \tfrac{p(X,Z)}{q(Z)} \right\} dZ - \int q(Z) \ln \left\{ \tfrac{p(Z|X)}{q(Z)} \right\} dZ \\ &= \int q(Z) \ln \left\{ \tfrac{p(X,Z)}{q(Z)} \right\} dZ + \int q(Z) \ln \left\{ \tfrac{p(q(Z)}{Z|X)} \right\} dZ \\ &= \mathcal{L}(q) + KL(q|p). \end{aligned} \tag{56}$$

where $\mathcal{L}(q)$ is called evidence lower bound. The relationship between evidence lower bound, KL divergence can be shown in Figure 1.

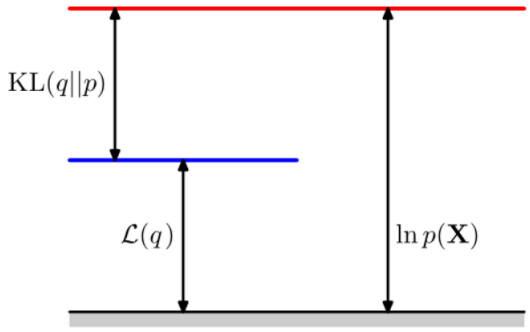

Figure 1

Rather then minimize KL divergence, we can maximize $\mathcal{L}(q)$ and an equivalent form of equation (53) is

$$q^*(Z) = \arg\max_{q(Z)} \int q(Z) \ln \left\{ \frac{p(X, Z)}{q(Z)} \right\} dZ. \tag{57}$$