**OPTIMIZATION OF THE LOSS FUNCTION**

As discussed, Ѳ\*, Ø\* = argmin Ѳ,Ø  L(Ѳ,Ø)

In variational Baysian method, this loss function is known as the variational lower bound or evidence lower bound (ELBO). This ‘lower bound ‘ part comes form the fact that KL divergence is always non-negative & thus L(Ѳ,Ø) is the lower bound of log PѲ(x).

L(Ѳ,Ø) = - E z ~ Q(ᵶ|x [ log (P(x|ᵶ)) + DkL (Q(ᵶ|x)||P (ᵶ)) ]

And we know DkL (Q(ᵶ|x)||P (ᵶ)) >=0

as a result **, L(Ѳ,Ø) <= log PѲ(x).**

therefore minimizing loss, we are maximixing the lower bound of the probability of generating the real data samples.

**REPARAMETRIZATION TRICK**

Recall :

L(Ѳ,Ø) = - E z ~ Q(ᵶ|x [ log (P(x|ᵶ)) + DkL (Q(ᵶ|x)||P (ᵶ)) ]

The optimization is carried out with respect to both Ѳ & Ø to learn QØ(ᵶ|x) PѲ(x|ᵶ) at the same time.

minѲ,Ø  L(Ѳ,Ø) = minѲ,Ø  { -E z ~ Q(ᵶ|x [ log (P(x|ᵶ)) + DkL (Q(ᵶ|x)||P (ᵶ)) ] }

we run the algorithm for a fixed number of iterations that is upto N. In that the first step is to find the derivative of L(Ѳ,Ø) with respect to Ѳ. The result will give the optimal value of Ѳi. The next step is to find optimal value of Øi. For that we will find the derivative of L(Ѳ,Ø) with respect to Ø. Here we replace the Ѳ with the optimal value find in last iteration that is Ѳi.

Ѳi = d/dѲ(L(Ѳ,Ø) )

Øi = d/dØ(L(Ѳ,Ø) )

The derivative d/dØ is harder to estimate because Ø appears in the distribution with respect to which expectation is taken.

If we can somehow rewrite the expectation in such a way that the Ø appears inside the expectation then we can push the gradient inside expectation i.e.,

EQ(ᵶ|x)[f(ᵶ)] = EP(£)[f(gØ(£,x))]

Such that Z = gØ(£,x) with £ ~ N(0,1)

In our case, gØ(£,x) = ψØ(x) + £. ∑Ø(x) = Z ~ N(ψØ(x), ∑Ø(x))

Here N(ψØ(x), ∑Ø(x)) is obtained from N(0,1) using the above linear transformation.

Instead of sampling ᵶ ~ QØ(ᵶ|x), we are sampling from N(0,1) i.e, £ ~ N(0,1) and then linear transformation using the above function to realize N(ψØ(x), ∑Ø(x)).

