towardsdatascience.com

Intuitive Guide to Variational Bayes Inference | Towards Data Science

Anwesh Marwade

8-10 minutes

Variational Bayes: the intuition behind Variational Auto-Encoders (VAEs)

The high computational cost of inferring from a complicated and often intractable, **'true posterior distribution'** has always been a stumbling block in the Bayesian framework. However (and thankfully), there are certain inference techniques that are able to achieve a reasonable approximation of this intractable posterior with something… *tractable.* See what I did there?

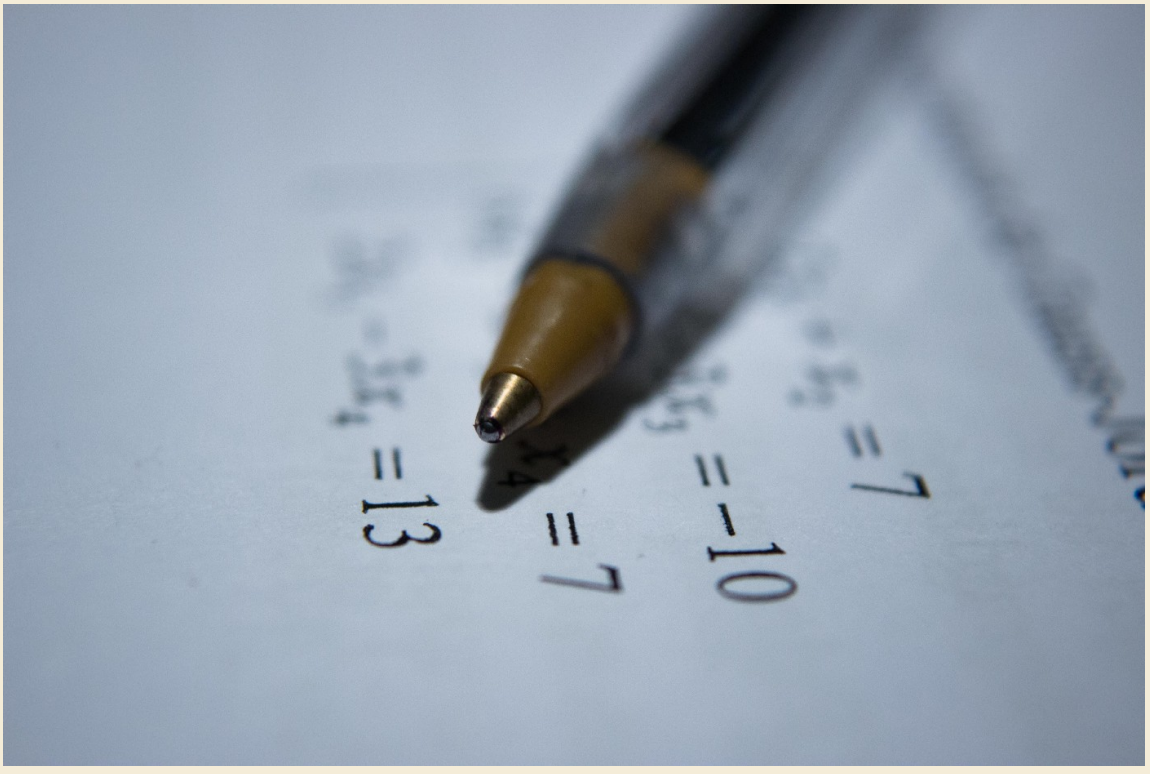


Photo by Antoine Dautry on Unsplash

One such approximate inference technique that has gained popularity in recent times is the **Variational Bayes (VB).** Having a relatively **low computational cost** and a **good empirical approximation** has propelled it to drive the intuition behind successful models like the **Variational Auto-encoders** and more. In this article, I look to build an intuition behind Variational Bayes as a latent variable model looking to approximate closely the *'true posterior distribution'* by optimising a statistical measure called the **Kullback-Leibler divergence.**

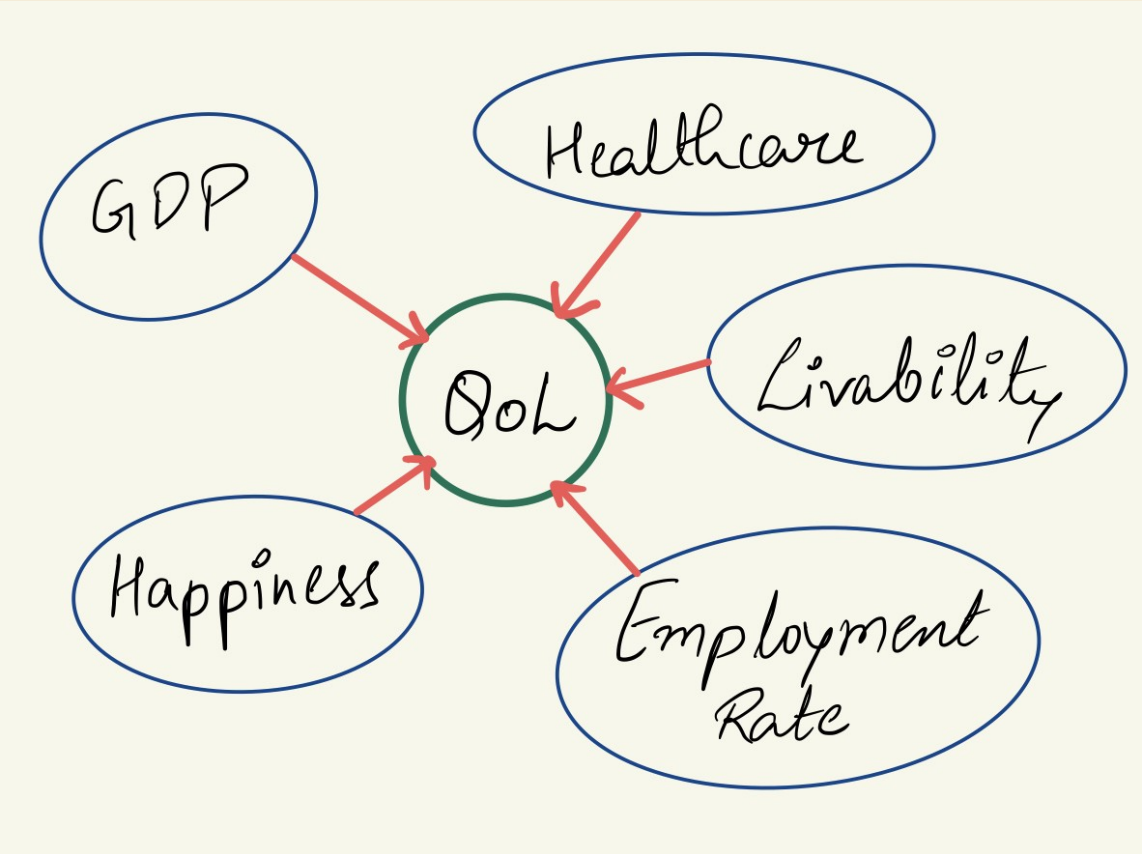
> **Fun fact:** Although we are using VB to construct an approximation of a posterior distribution, it is not its primary motivation. The posterior approximation idea appears when attempting to maximize the log marginal likelihood. You'll see!
> — A first course in Machine Learning by Simon Rogers and Mark Girolami.

Before we proceed any further, let us take a simple (toy) example for laying some ground-work on the idea of latent variables (and possibly understanding the distinction between latent variables and model parameters).

Toy Example:

Quality of Life (QoL) is a common measure used by WHO to describe the state of living conditions in a given geographic region. It is often defined as the degree to which an individual is healthy, comfortable, and able to participate in or enjoy life events [1]. Even though quantifying such abstract quantities are usually difficult, based on general understanding, we can come up with some (simple) indicators that would affect this QoL measure. For example, the per Capita GDP of the region, the

employment rate, quality of education, healthcare, livability, happiness index are all fairly tractable features that might influence the QoL. (see figure)



A simple (totally made-up) latent variable model. *Image by Author.*

Talking about our made-up model, we can see that there are multiple factors, both directly observable or otherwise, that (might) have an influence on our quantity of interest i.e. the QoL measure. Since we cannot measure this abstract value by itself, we have used multiple surrogate (observable) variables to *somewhat* quantify it. Such a quantity is termed as a **latent variable**. The value of a latent variable can be inferred from measurements of the observable variables that might influence it. Such is the case in many real-world scenarios and this is where latent variable models are useful.

> Even when we (wrongly) assume to know of a relationship between a quantity we can observe and what we wish to

observe, there might be some unknown or hidden variables at play which are left unaccounted.

These unaccounted or hidden quantities, however, are taken care of when using latent variable models. Otherwise, down the road, these quantities might come up as errors in our model. For some intuition about intentional modelling of error, check out this article:

Variation Inference: The Bayesian way

Having digested our short example on the idea of latent variables, let us contextualize our main idea of Variational Inference by considering a general modelling scenario with some observable data ***Y,*** and model parameters and /or latent variables defined by ***θ****.* In the Bayesian domain, usually, we define both the model parameters and latent variables by treating them as random variables and lumping them together in the ***θ*** term. In other words, we consider all that we do not know about in the given scenario as a part of this ***θ*** variable*.* To the ones who were paying attention, onto maximizing the log marginal likelihood now! The marginal looks like this,

$$p(Y) = \int p(Y, \theta) d\theta = \int p(Y \mid \theta) p(\theta) d\theta$$

the joint density ***p(Y, θ)*** can be broken up into ***p(Y|θ)p(θ).*** *Image by Author.*

where we sum up the likelihood, ***p(Y|θ)*** over all possible values of the parameter ***θ,*** weighted by the prior ***p(θ)*** and we are looking to maximize this quantity. However, owing to the potentially high dimensional integral over the entire parameter space, we won't be able to achieve an exact evaluation of this expression; which is why we will make use of **Jensons's**

inequality to place a lower bound over the ***log*** (which is a convex function) of our ***marginal term***.

$$log\ E_{p(x)}\left\{f(x)\right\} \geq E_{p(x)}\left\{log\ f(x)\right\}$$

Jenson's inequality allows us to place a lower bound. Image by Author.

Taking a log over the marginal likelihood, we have the following:

$$log\ p(Y) = log \int p(Y, \theta) d\theta$$

Image by Author.

Before applying Jenson's inequality we introduce ***Q(θ),*** which is an arbitrary distribution over ***θ,*** by multiplying and dividing it on the right-hand side. Looking at the inequality expression we mentioned above, our right-hand term can be thought of as an expectation of the ***p(Y, θ)/Q(θ)*** term w.r.t ***Q(θ).*** Now applying Jenson's inequality, we can get a lower bound on our log marginal.

$$\begin{aligned} log\ p(Y) &= log \int Q(\theta) \frac{p(Y, \theta)}{Q(\theta)} d\theta \\ &\geq \int Q(\theta)\ log\ \frac{p(Y, \theta)}{Q(\theta)} d\theta \\ &= L(Q)\ (lower\ bound) \end{aligned}$$

We obtain a lower bound in terms of ***Q(θ)*** *and* ***the posterior.*** Image by Author.

> Given that we have chosen this distribution ***Q(θ)*** arbitrarily, it is a variational knob that we can fine-tune and this fact is going to be quite an important revelation as you shall see.

Now, let us compute the difference between our log marginal likelihood term and the lower bound ***L(Q)*** that we just found.

$$
\begin{aligned}
log\ p(Y) - L(Q) &= log\ p(Y) - \int Q(\theta)\ log \frac{p(Y,\theta)}{Q(\theta)} d\theta \\
&= log\ p(Y) - \int Q(\theta)\ log \frac{p(\theta \mid Y)p(Y)}{Q(\theta)} d\theta \\
&= log\ p(Y) - \int Q(\theta)\ log \frac{p(\theta \mid Y)p(Y)}{Q(\theta)} d\theta - \int Q(\theta)\ log\ p(Y) d\theta \\
&= log\ p(Y) - \int Q(\theta)\ log \frac{p(\theta \mid Y)p(Y)}{Q(\theta)} d\theta - log\ p(Y) \int Q(\theta)\ d\theta \\
&\text{The integral of } Q(\theta) \text{ over all } \theta \text{ is equal to 1 as it is a probability density} \\
&= log\ p(Y) - \int Q(\theta)\ log \frac{p(\theta \mid Y)p(Y)}{Q(\theta)} d\theta - log\ p(Y)
\end{aligned}
$$

Working out the difference: ***log p(Y) — L(Q).*** *Image by Author*

Does the integral term seem familiar yet? Let us clean it up a bit, like so:

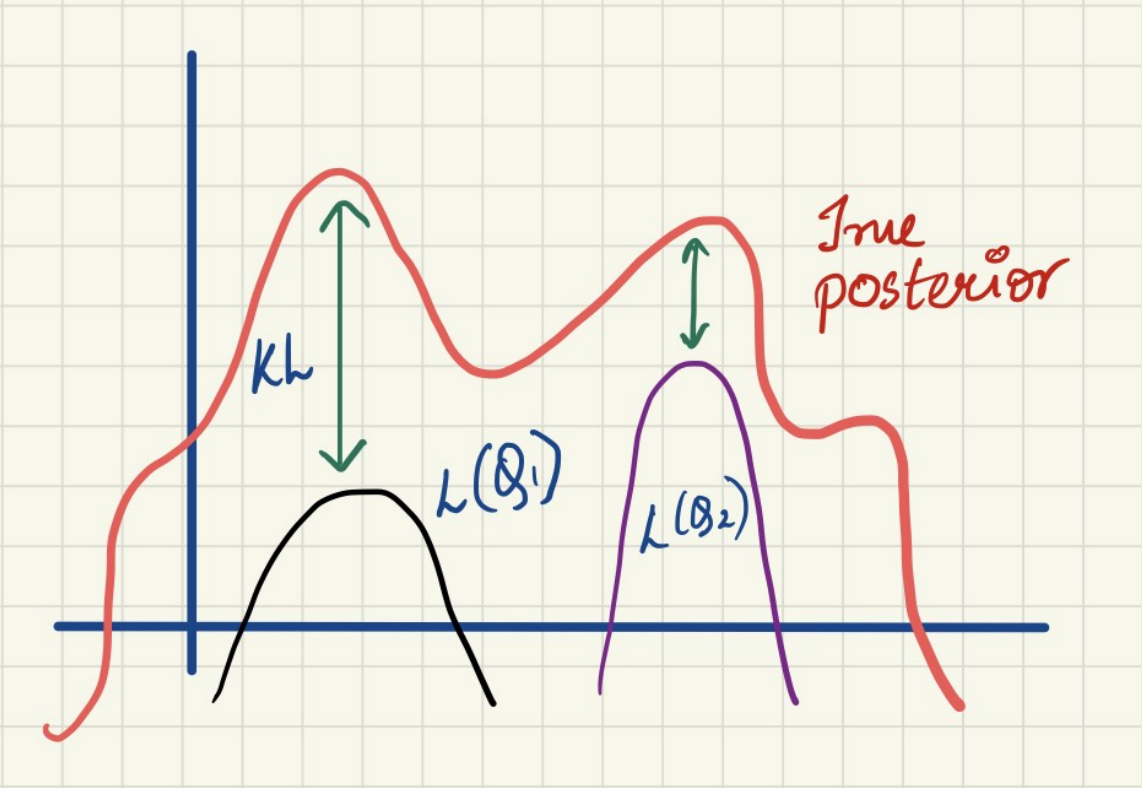
$$
\begin{aligned}
log\ p(Y) - L(Q) &= -\int Q(\theta)\ log \frac{p(\theta \mid Y)p(Y)}{Q(\theta)} d\theta \\
&= -KL\left[Q(\theta) \parallel p(\theta \mid Y)\right].
\end{aligned}
$$

The difference between the marginal and the lower bound (which we found using Jenson's inequality), works out to be equal to the negative KL divergence! *Image by Author.*

The expression under the integral works out (quite nicely) to be the **Kullback-Leibler (KL) divergence** between our true posterior ***p(θ|Y)*** and the arbitrarily chosen ***Q(θ).*** Interestingly, KL divergence is a measure that is often used to quantify the difference between two probability distributions and in our case, the posterior (that we wish to approximate) and arbitrarily chosen ***Q(θ)*** are those two distributions. The appearance of this KL divergence term in the derivation of our Variational Bayes

exercise isn't just a coincidence; we are indeed trying to find an approximate distribution that is most similar to the form of our true posterior and voila! ***Q(θ)*** is the variational knob that will allow us to tweak the ***L(Q)*** term and correspondingly maximising ***L(Q)*** will allow ***Q(θ)*** to become more similar to the true posterior ***p(θ|Y)*** hence reducing the negative of the KL divergence!

> **Note**: **KL divergence** as a measure is generally **less than or equal to zero**, with its maximum value i.e. zero occurring **when the two distributions are equal.** So, in case, when we find the true posterior, the **KL divergence** term will be zero and the bound **L*(Q)*** will be equal to the **log marginal likelihood**!



Visualising the variations of ***Q(θ)*** *in an attempt to approximate the true posterior. Image by Author.*

Basically, we have approximated the true distribution of the posterior using a lower bound term instead of a whole distribution which makes it easier to optimize. As can be seen in the figure above, the idea is to attempt to find the true posterior using an arbitrary distribution ***Q(θ).*** This optimisation of ***Q(θ)*** as

the best approximation of the posterior, ***p(θ|Y)*** is usually achieved through an iterative optimisation procedure like the **Expectation Maximisation** (EM) algorithm. ’s article on the EM algorithm is a good read (quite mathematically involved):

This concludes the journey of inference towards (potentially) understanding the powerful idea that drives state-of-the-art models like *Variational Auto-Encoders* and *Generative Adversarial Networks.* For an exhaustive conceptualization of the VAEs, this article by is a great read.

Thank you for your time! If you liked this article, please like or share the content as a way of showing your support, letting me know about your interests. Follow me on Medium or connect via LinkedIn.

Until next time!

References:

1. *Wikipedia article on Quality of Life (QOL): https://en.wikipedia.org/wiki/Quality_of_life*
2. *Simon Rogers and Mark Girolami. 2016. A First Course in Machine Learning, Second Edition (2nd. ed.). Chapman & Hall/CRC.*