Mean Field Approximation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Variational Bayesian (VB) methods are a family of techniques that are very popular in statistical Machine Learning. VB methods allows us to re-write statistical inference problems (i.e. infer the value of a random variable given the value of another random variable) as *optimization* problems (i.e. find the parameter values that minimize some objective funciton). The inference-optimization duality is powerful because it allows us to use the latest-and-greatest optimization algorithms to solve statistical ML problems and vice versa, minimize functions using statistical techniques.

1 Preliminaries and Notations

1. Uppercase X denotes a random variable.
2. Uppercase $P(X)$ denotes the probability distribution over that variable.
3. Lowercase $x \sim P(X)$ denotes a value x sampled from the prob distribution via some generative process.
4. Lowercase $p(X)$ is the density function of the distribution of X. It is a scalar function over the measure space of X.
5. $p(X = x)$ (shorthand $p(x)$) denotes the density function evaluated at x.

We model systems as a collection of random variables, where some variables (X) are *observable*, while other variables (Z) are *hidden*. We can draw this relationship via Fig 1.

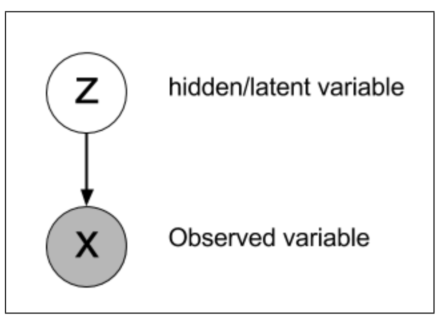

Figure 1: The edge drawn from Z to X relates the two variables together via the conditional distribution $P(X|Z)$.

Here is a more concrete example: X might represent the "raw pixel values of an image", while Z is a binary variable such that $Z = 1$ "if X is an image of a cat". Refer to Fig 2, 3, and 4

Submitted to 31st Conference on Neural Information Processing Systems (NIPS 2017). Do not distribute.

Figure 2: If X is this image, $P(Z = 1) = 1$ (definitely a cat).

Figure 3: If X is this image, $P(Z = 1) = 0$ (definitely not a cat).

Bayes' Theorem gives us a relationship between any pair of RVs.

$$p(Z|X) = \frac{p(X|Z)p(Z)}{p(X)} \tag{1}$$

$p(Z|X)$ is the **posterior probability**: the probability of Z after taking into account X, "given the image, what is the probability that this is of a cat". If we sample from $z \sim P(Z|X)$, we can use this to make a cat classifier.

$p(X|Z)$ is the **likelihood**: the initial degree of belief in X, given the proposition Z is true, "given a value of Z, compute how probable this image X is under the category cat/non-cat." If we sample from $x \sim P(X|Z)$, we generate images of cats and non-cats.

$p(Z)$ is the **prior probability**: the initial degree of belief in Z. This captures any prior information about Z, if we think that 1/3 of all images in existence are of cats, then $p(Z = 1) = 1/3$.

Hidden variables can be interpreted from a Bayesian Statistics framework as *prior beliefs* attached to the observed variables. For example, if we belief X is a multivariate gaussian, the hidden variable Z might represent the mean and variance of the Gaussian. The distribution over parameters $P(Z)$ is then a *prior* distribution to $P(X)$.

You are also free to choose which values X and Z represent. For example, Z could instead be "mean, cube root of variance, and $X + Y$ where $Y \sim N(0, 1)$". This is somewhat unnatural and weird, but the structure is still valid, as long as $P(X|Z)$ is modified accordingly.

You can even "add" variables to your system. The prior itself might be dependent on other random variables via $P(Z|\theta)$, which have prior distributions of their own $P(\theta)$, and those have priors still, and so on. *Any hyper-parameter can be thought of as a prior*, see Fig 5.

2 Problem Formulation

The key problem we are interested in is *posterior inference*, or computing functions on the hidden variable Z. Some canonical examples of posterior inference:

1. Given this surveillance footage X, did the suspect show up in it?
2. Given this twitter feed X, is the author depressed?
3. Given historical stock prices $X_{1:t-1}$, what will X_t be?

Figure 4: If X is this image, $P(Z = 1) = 0.1$ (sort of cat-like).

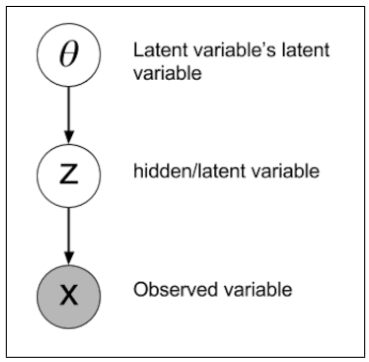

Figure 5: Latent variable's latent variable.

We usually assume that we know how to compute functions on likelihood function $P(X|Z)$ and priors $P(Z)$.

The problem is, for complicated tasks like above, we often don't know how to sample from $P(Z|X)$ or compute $p(X|Z)$. Alternatively, we might know the form of $p(Z|X)$, but the corresponding computation is so complicated that we cannot evaluate it in a reasonable amount of time. We could try to use sampling-based approaches like MCMC, but these are slow to converge.

3 Lower Bound for Mean-Field Approximation

The idea behind variational inference is to perform inference on an easy parametric distribution $Q_{Phi}(Z|X)$ (like a Gaussian) for which we know how to do posterior inference, but adjust the parameters Φ so that Q_Φ is as close to P as possible. See Fig 8 for better illustration. The blue curve is the true posterior distribution, and the green distribution is the variational approximation (Gaussian) that we fit to the blue density via optimization.

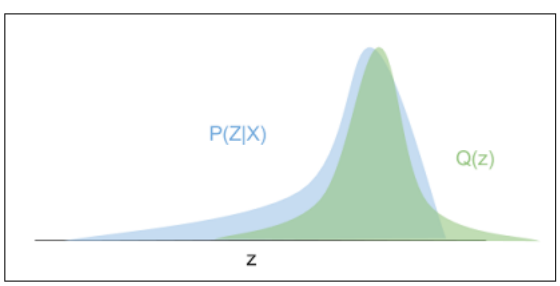

Figure 6: Variational approximation (Gaussian).

3.1 Mean Field Approximation

MF approximation assumes that:

1. Q(x) is our MF approximation.
2. Variables in the Q distribution are independent variables X_i.
3. In the standard MF approach, Q is completely factorized: $Q(x) = \prod_i Q_i(x_i)$

What does it mean for distributions to be "close"? Mean-field variational Bayes (the most common type) uses the Reverse KL Divergence to as the distance metric between two distributions.

$$KL(Q_\Phi(Z|X)||P(Z|X)) = \sum_{z \in Z} q_\Phi(z|x) \log \frac{q_\Phi(z|x)}{p(z|x)} \tag{2}$$

Reverse KL divergence measures the amount of information (in nats, or units of $\frac{1}{\log(2)}$ bits) required to "distort" $P(Z)$ into $Q_\Phi(Z)$. We wish to minimize this quantity w.r.t. Φ.

By definition of conditional distribution, $p(z|x) = \frac{p(x,z)}{p(x)}$. Let's substitute this expression into our original KL expression, and then distribute:

$$\begin{aligned}
KL(Q|P) &= \sum_{z \in Z} q_\Phi(z|x) \log \frac{q_\Phi(z|x) p(x)}{p(z,x)} \\
&= \sum_{z \in Z} q_\Phi(z|x) \big(\log \frac{q_\Phi(z|x)}{p(z,x)} + \log p(x) \big) \\
&= \Big(\sum_z q_\Phi(z|x) \log \frac{q_\Phi(z|x)}{p(z,x)} \Big) + \Big(\sum_z \log p(x) q_\Phi(z|x) \Big) \\
&= \Big(\sum_z q_\Phi(z|x) \log \frac{q_\Phi(z|x)}{p(z,x)} \Big) + \Big(\log p(x) \sum_z q_\Phi(z|x) \Big) \\
&= \Big(\sum_z q_\Phi(z|x) \log \frac{q_\Phi(z|x)}{p(z,x)} \Big) + \log p(x)
\end{aligned} \tag{3}$$

To minimize $KL(Q||P)$ w.r.t. variational parameters Φ, we just have to minimize $\sum_z q_\Phi(z|x) \log \frac{q_\Phi(z|x)}{p(z,x)}$, since $\log p(x)$ is fixed w.r.t. Φ. Let's rewrite this quantity as an expectation over the distribution $Q_\Phi(Z|X)$.

$$\begin{aligned}
\sum_z q_\Phi(z|x) \log \frac{q_\Phi(z|x)}{p(z,x)} &= \mathbb{E}_{z \sim Q_\Phi(Z|X)} \log \frac{q_\Phi(z|x)}{p(z,x)} \\
&= \mathbb{E}_Q \big[\log q_\Phi(z|x) - \log p(x,z) \big] \\
&= \mathbb{E}_Q \big[\log q_\Phi(z|x) - \log p(x|z) - \log p(z) \big]
\end{aligned} \tag{4}$$

Minimizing this is equavalent to maximizing the negation of this function:

$$\begin{aligned}
maximize \mathcal{L} &= - \sum_z q_\Phi(z|x) \log \frac{q_\Phi(z|x)}{p(z,x)} \\
&= \mathbb{E}_Q \big[- \log q_\Phi(z|x) + \log p(x|z) + \log p(z) \big] \\
&= \mathbb{E}_Q \big[\log p(x|z) + \log \frac{p(z)}{q_\Phi(z|x)} \big]
\end{aligned} \tag{5}$$

$\mathcal{L}$ is known as the *variational lower bound*, and is computationally tractable if we can evaluate $p(x|z)$, $p(z)$, $q(z|x)$. We can further re-arrange terms in a way that yields an intuitive formula:

$$\begin{aligned}\mathcal{L} &= \mathbb{E}_Q\big[\log p(x|z) + \log \frac{p(z)}{q_\Phi(z|x)}\big] \\ &= \mathbb{E}_Q\big[\log p(x|z)\big] + \sum_Q q(z|x) \log \frac{p(z)}{q_\Phi(z|x)} \\ &= \mathbb{E}_Q\big[\log p(x|z)\big] - KL\big(Q(Z|X)||P(z)\big)\end{aligned} \tag{6}$$

$\mathcal{L}$ itself contains a KL divergence term between the approximate posterior and the prior. If sampling $z \sim Q(Z|X)$ is an encoding process that converts an observation x into latent code z, then sampling $x \sim P(X|Z)$ is a decoding process that reconstructs the observation from z.

It follows that $\mathcal{L}$ is the sum of the expected *decoding likelihood* (how good our variational distribution can decode a sample of Z back to a sample of X), plus the KL divergence between the variational approximation and the prior on Z. If we assume $Q(Z|X)$ is conditional Gaussian, then prior Z is often to be a diagonal Gaussian distribution with mean 0 and standard deviation 1.

4 Forward KL and Reverse KL

Let's consider the forward KL first. We can write KL as the expectation of a "penalty" funciton $\log \frac{p(z)}{q(z)}$ over a weighting function $p(z)$.

$$KL(P||Q) = \sum_z p(z) log \frac{p(z)}{q(z)} = \mathbb{E}_{p(z)}\big[\log \frac{p(z)}{q(z)}\big] \tag{7}$$

The penalty function contributes loss to the total KL (that's why called penalty) wherever $p(Z) > 0$. For $p(Z) > 0$, $\lim_{q(Z)\to 0} \log \frac{p(z)}{q(z0} \to \infty$. This means that the forward-KL will be large wherever $Q(Z)$ fails to "cover up" $P(Z)$. Therefore, the forward-KL is minimized when we ensure that $q(z) > 0$ wherever $p(z) > 0$. The optimized variational distribution $Q(Z)$ is known as "zero-avoiding" (density avoids zero when $p(Z)$ is zero). See Fig

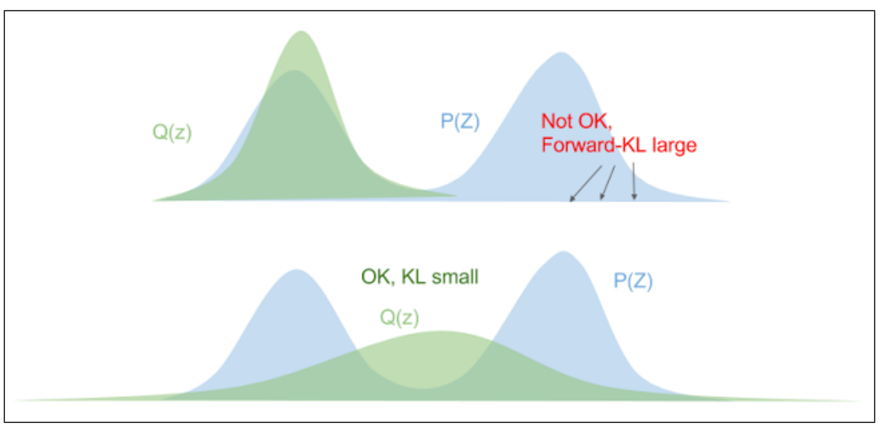

Figure 7: Forward KL.

KL divergence is always non-negative which can be proved by using Jensen's inequality. Minimizing the reverse-KL has the opposite behavior: If $p(Z) = 0$, we must ensure that the weighting function $q(Z) = 0$ wherever denominator $p(Z) = 0$, this is known as "zero-forcing". See Fig

So in summary, minimizing forward-KL "stretches" your variational distribution $Q(Z)$ to cover over the entire $P(Z)$ like a tarp, while minimizing reverse-KL "squeezes" the $Q(Z)$ under $P(Z)$.

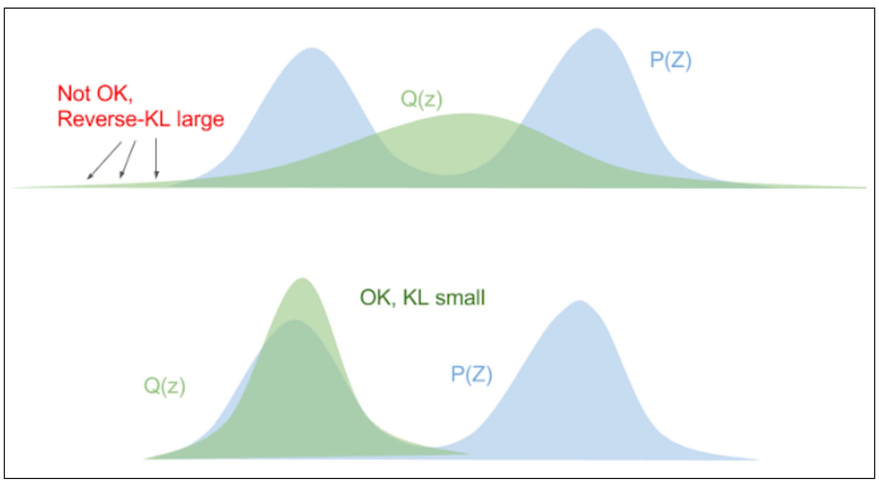

Figure 8: Reverse KL.