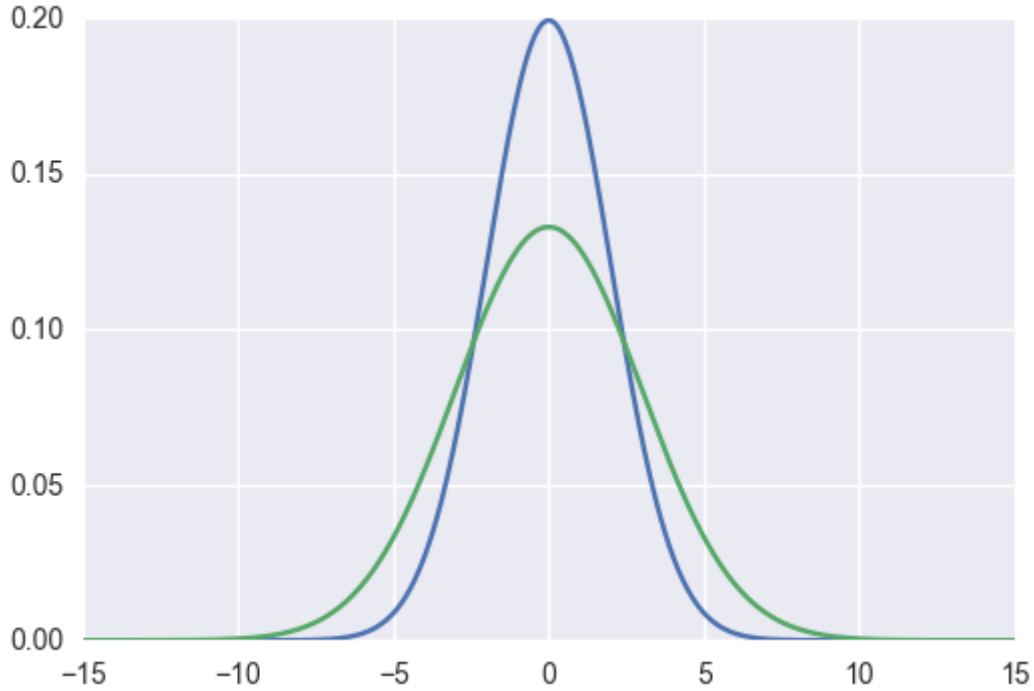
KL Divergence: Forward vs Reverse?

Kullback-Leibler Divergence, or KL Divergence is a measure on how "off" two probability distributions $P(X)$ and $Q(X)$ are. It measures the distance between two probability distributions.

For example, if we have two gaussians, $P(X) = N(0, 2)$ and $Q(X) = N(0, 3)$, how different are those two gaussians?



The KL Divergence could be computed as follows:

$$D_{KL}[P(X) \parallel Q(X)] = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

that is, for all random variable $x \in X$, KL Divergence calculates the weighted average on the difference between those distributions at x.

KL Divergence in optimization

In optimization setting, we assume that $P(X)$ as the true distribution we want to approximate and $Q(X)$ as the approximate distribution.

Just like any other distance functions (e.g. euclidean distance), we can use KL Divergence as a loss function in an optimization setting, especially in a probabilistic setting. For example, in Variational Bayes, we are trying to fit an approximate to the true posterior, and the process to make sure that $Q(X)$ fits $P(X)$ is to minimize the KL Divergence between them.

However, we have to note this important property about KL Divergence: it is not symmetric. Formally, $D_{KL}[P(X) \parallel Q(X)] \neq D_{KL}[Q(X) \parallel P(X)]$.

$D_{KL}[P(X) \parallel Q(X)]$ is called forward KL, whereas $D_{KL}[Q(X) \parallel P(X)]$ is called reverse KL.

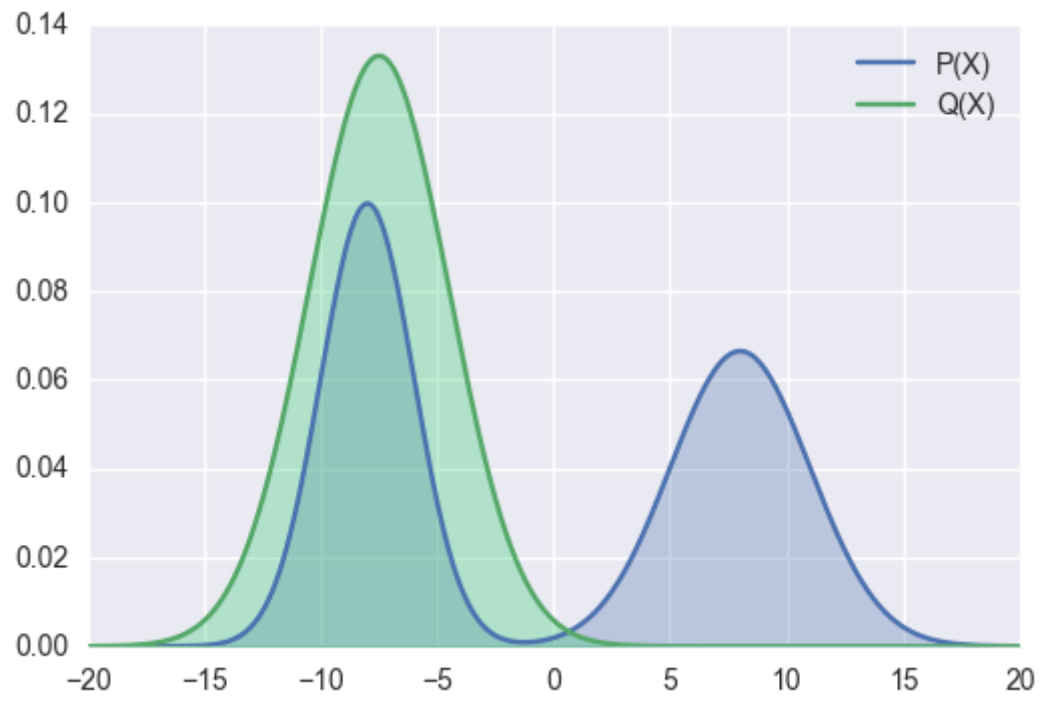
Forward KL

In forward KL, the difference between $P(x)$ and $Q(x)$ is weighted by $P(x)$. Now let's ponder on that statement for a while.

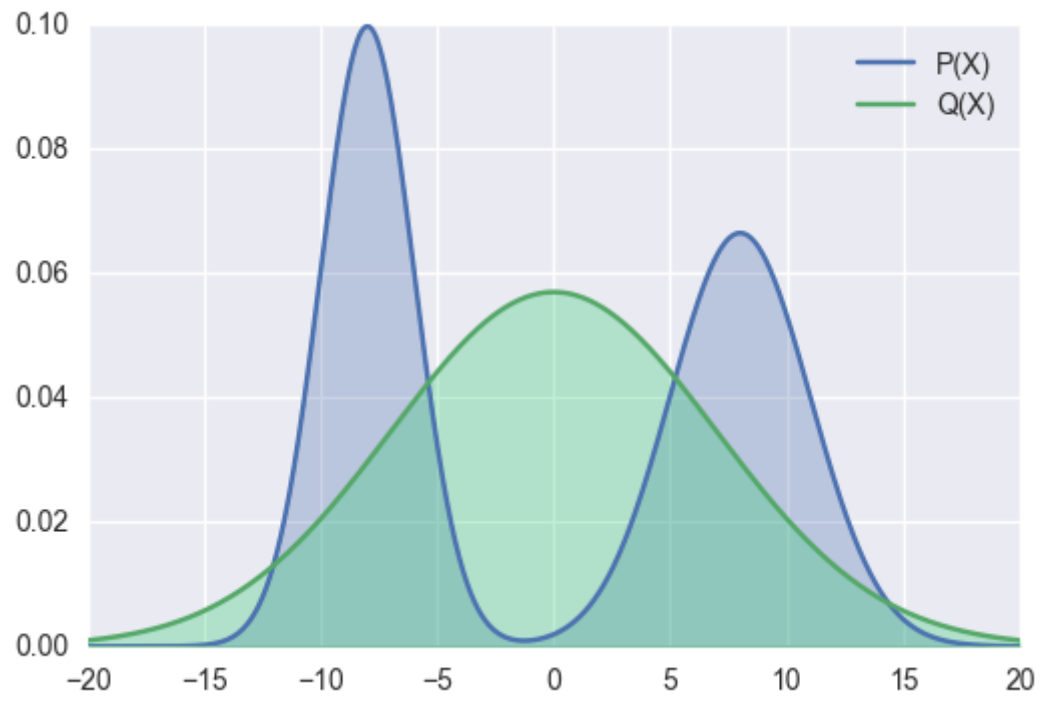
Consider $P(x) = 0$ for a particular x. What does that mean? As $P(x)$ is the weight, then it doesn't really matter what's the value of the other term. In other words, if $P(x) = 0$, there is no consequence at all to have very big difference between $P(x)$ and $Q(x)$. In this case, the total KL Divergence will not be affected when $P(x) = 0$, as the minimum value for KL Divergence is 0 (no distance at all, i.e. exact match). During the optimization process then, whenever $P(x) = 0$, $Q(x)$ would be ignored.

Reversely, if $P(x) > 0$, then the $\log\left(\frac{P(x)}{Q(x)}\right)$ term will contribute to the overall KL Divergence. This is not good if our objective is to minimize KL Divergence. Hence, during the optimization, the difference between $P(x)$ and $Q(x)$ will be minimized if $P(x) > 0$.

Let's see some visual examples.



In the example above, the right hand side mode is not covered by $Q(x)$, but it is obviously the case that $P(x) > 0$! The consequence for this scenario is that the KL Divergence would be big. The optimization algorithm then would force $Q(x)$ to take different form:



In the example above, $Q(x)$ is now more spread out, covering all $P(x) > 0$. Now, there is no $P(x) > 0$ that are not covered by $Q(x)$.

Although there are still some area that are wrongly covered by $Q(x)$, this is the desired optimization result as in this form of $Q(x)$, the KL Divergence is low.

Those are the reason why, Forward KL is known as *zero avoiding*, as it is avoiding $Q(x) = 0$ whenever $P(x) > 0$.

Reverse KL

In Reverse KL, as we switch the two distributions' position in the equation, now $Q(x)$ is the weight. Still keeping that $Q(x)$ is the approximate and $P(x)$ is the true distribution, let's ponder some scenarios.

First, what happen if $Q(x) = 0$ for some x, in term of the optimization process? In this case, there is no penalty when we ignore $P(x) > 0$.

Second, what happen if $Q(x) > 0$? Now the difference between $P(x)$ and $Q(x)$ must be as low as possible, as it now contribute to the overall divergence.

Therefore, the failure case example above for Forward KL, is the desireable outcome for Reverse KL. That is, for Reverse KL, it is better to fit just some portion of $P(X)$, as long as that approximate is good.

Consequently, Reverse KL will try avoid spreading the approximate. Now, there would be some portion of $P(X)$ that will not be approximated by $Q(X)$, i.e. $Q(x) = 0$.

As those properties suggest, this form of KL Divergence is know as *zero forcing*, as it forces $Q(X)$ to be 0 on some areas, even if $P(X) > 0$.

Conclusion

So, what's the best KL?

As always, the answer is "it depends". As we have seen above, both has its own characteristic. So, depending on what we want to do, we choose which KL Divergence mode that's suitable for our problem.

In Bayesian Inference, esp. in Variational Bayes, Reverse KL is widely used. As we could see at the derivation of Variational Autoencoder (/techblog/2016/12/10/variational-autoencoder/), VAE also uses Reverse KL (as the idea is rooted in Variational Bayes!).

References

1. Blei, David M. "Variational Inference." Lecture from Princeton, https://www. cs. princeton. edu/courses/archive/fall11/cos597C/lectures/variational-inference-i. pdf (2011).
2. Fox, Charles W., and Stephen J. Roberts. "A tutorial on variational Bayesian inference." Artificial intelligence review 38.2 (2012): 85-95.

← PREVIOUS POST (/TECHBLOG/2016/12/17/CONDITIONAL-VAE/)

NEXT POST → (/TECHBLOG/2016/12/24/CONDITIONAL-GAN-TENSORFLOW/)

 (/feed.xml)  (https://github.com/wiseodd)

Copyright © Agustinus Kristiadi's Blog 2018