Lecture: Variational Approximation for Bayes Inference

Qianheng Ma

March 7, 2017

Lecturer: John Novembre ***Note***: *These lecture notes are still rough, and have only have been mildly proofread.*

Background of Variational Approximation.

We are always insterested in the approximation of the posterior distribution when we are doing the Bayesian Inference. variational approximation is becoming a more and more popular in Bayesian Inference as an alternative approach for Markov Chain Monte Carlo Method that takes long time to converge and has the uncertainty of convergence.

To think of MCMC as a way of doing inference via sampling the posterior distribution, the variation approximation techniques allows for the inference as optimization. An optimization is an area with a lot of rich tools which have been developed broadly in computational sciences.

Pre-requisites of Variational Calculus.

- Functional: function of functions.
- Variational calculus is for optimazation of functionals.

Several useful results at hand:

- $\frac{\partial}{\partial q}\int q(x)dx = 1$
- $\frac{\partial}{\partial q}\int ln(q(x))dx = \frac{1}{q(x)}$
- $\frac{\partial}{\partial q}\int q(x)ln(q(x))dx = ln(q(x)) + 1$
- $\frac{\partial}{\partial q}\int q(x)f(x)dx = f(x)$

Kullback-Leibler Divergence

In particular, $q(\bullet)$ s an approximation to $p(\theta|x)$. Here we use the Kullback-Leibler Divergence(KL) as and measure of non-symmetric difference between two probability distributions.

$min_q KL(q|p) = E_g(\frac{log(q_\theta(x))}{log(p_\theta(x))}) = \int q(x)(log\frac{q(x)}{p(x)}dx)$

Note that, when $q(x) = p(x)$, $KL = 0$. KL divergence is non-symmetric since KL changes when we switch $q(x)$ and $p(x)$.The log ratio $log\frac{q(x)}{p(x)}$ can be served as an penalty function for KL, i.e., if $q(x)$ is large, and $p(x)$ is similarly large, then the ratio contributes little to the optimazition; if $q(x)$ is large, and $p(x)$ is small, then the ratio contributes a lot to the optimazition. Therefore, KL divergence captures more when $q(x)$ is large and $p(x)$ is small.

Now we take p to be fixed and $kL(q|p)$ is a function of q.

Example: Inference of Allele Frequency.

Now we are going to use this example to show how variational approximation works to infer allele frequency across individuals. Consider we have three genotypes: AA, Aa, aa, corresponding to $Z_i = 0, 1, 2$ for each individual i=1,2,... n. The proportion of type a is θ, Therefore the proportion of type A is $1 - \theta$.Therefore we assumed $\theta \sim Beta(\alpha, \beta)$ and $Z_i \sim Binomial(2, \theta)$. X_i denotes the counts of read of type a and $X_i \sim Binomial(n_i, \frac{Z_i}{2})$, n_i is the number of read at one location of the genome for individual i.

Remember we are interested in the posterior joint distribution $P(\theta, Z|X)$.

Start with the KL divergence.Denote Z is the vector storing the genotype and X is the vector storing the number of read, for each individual.

$KL(q|p) = \int q(\theta, Z) log \frac{q(\theta,Z)}{P(\theta,Z|X)} d\theta dZ$

Recall that, $P(\theta, Z|X) = \frac{P(\theta,Z,X)}{P(X)}$.

Therefore,

$KL(q|p) = \int q(\theta, Z) log(\frac{q(\theta,Z)}{P(\theta,Z,X)}) d\theta dZ + \int q(\theta, Z) log(p(X)) d\theta dZ$

Since for the second part of the equation above, $log(p(X))$ is free from Z and θ, it can be moved outside the intergral.

In addition,

$\int q(\theta, Z) d\theta dZ = 1$

Therefore,

$KL(q|p) = \int q(\theta, Z) log \frac{q(\theta,Z)}{P(\theta,Z,X)} d\theta dZ + log(p(X))$

Notice that the first part of the equation actually is the negative of the lower bound of $log(p(X))$

$log(p(X)) = log \int p(\theta, Z, X) d\theta dZ = log \int p(\theta, Z, X) \frac{q(\theta,Z)}{q(\theta,Z)} d\theta dZ$

Then by Jensen's Inequality,$log(E(X)) \geqslant E(log(X))$,

$log \int p(\theta, Z, X) \frac{q(\theta,Z)}{q(\theta,Z)} d\theta dZ = log E_q(\frac{p(\theta,Z,X)}{q(\theta,Z)}) \geqslant E_q(log(\frac{p(\theta,Z,X)}{q(\theta,Z)}))$, we call this evidence lower bound(ELBO), which is the negative first part of the equation above.

By the assumption of independence, $q(\theta, Z)$ can be decomposed into $q(\theta) \prod_i^n q(Z_i)$. Here we treat $\prod_i^n q(Z_i)$ a constant.

While $q(\theta) \sim Beta(\alpha\ , \beta\ )$, and $q(Z_i) \sim Multinomial(p0, p1, p2), 1)$, $p0, p1, p2$ are the probabilities of the three genotypes.

Now we focus on this ELBO, and treat it as a new functional of $q(\theta)$.

$\int q(\theta, Z) log \frac{q(\theta,Z)}{P(\theta,Z,X)} d\theta dZ$

$= \int q(\theta) \prod_{j=1}^n q(Z_j) log \frac{p(\theta) \prod (p(Z_i|\theta) p(Xi|Zi)))}{q(\theta) \prod q(Z_i)} d\theta dZ$

$= \int q(\theta) \prod_{j=1}^n q(Z_j) (log(p(\theta)) - log(q(\theta)) + \sum_{i=1}^n log(p(Zi|\theta) p(Xi|Zi)) - \sum_{i=1}^n q(Z_i)) d\theta dZ$

$= \int q(\theta) \prod_{j=1}^n q(Z_j) (log(p(\theta)) - log(q(\theta)) - \sum_{i=1}^n log(q(Z_i))) d\theta dZ + \int q(\theta) \prod_{j=1}^n q(Z_j) (\sum_{i=1}^n p(Z_i|\theta) p(X_i|Z_i)) d\theta dZ$

Focus on the second part of the equation above,

$\int q(\theta) \prod_{j=1}^n q(Z_j) (\sum_{i=1}^n p(Z_i|\theta) p(X_i|Z_i)) d\theta dZ$

$= \sum_{i=1}^n \int q(\theta) \prod_{j=1}^n q(Z_j) log p(Z_i|\theta) p(X_i|Z_i) dZ id\theta$

This inegration included a lot of terms regarding each j in $\prod_{j=1}^n q(Z_j)$. Other terms will vanish except for the term regarding $q(Z_j)$ when $j = i$

And we first integrate Z_i then θ

Therefore, $\int q(\theta)\prod_{j=1}^{n} q(Z_j)(\sum_{i=1}^{n} p(Z_i|\theta)p(X_i|Z_i))d\theta dZ$

$= \sum_{i=1}^{n} \int q(\theta)E_q(log(p(Z_i|\theta)p(X_i|Z_i)))d\theta$

On the other hand,

For $\int q(\theta)\prod_{j=1}^{n} q(Z_j)(log(p(\theta)) - log(q(\theta)) - \sum_{i=1}^{n} log(q(Z_i)))d\theta dZ$,

First for $\int_\theta q(\theta)(log(p(\theta))(\int_Z \prod_{j=1}^{n} q(Z_j)dZ)d\theta$, note that $\int_Z \prod_{j=1}^{n} q(Z_j)dZ = 1$

Thus,$\int_\theta q(\theta)(log(p(\theta))(\int_Z \prod_{j=1}^{n} q(Z_j)dZ)d\theta = \int_\theta q(\theta)(log(p(\theta))d\theta$

Similarly, $\int_\theta q(\theta)(log(q(\theta))(\int_Z \prod_{j=1}^{n} q(Z_j)dZ)d\theta = \int_\theta q(\theta)(log(q(\theta))d\theta$

To optimize the functional $L_\theta(q(\theta)) = \int_\theta q(\theta)(log(p(\theta))d\theta - \int_\theta q(\theta)(log(q(\theta))d\theta + \sum_{i=1}^{n} \int q(\theta)E_q(log(p(Z_i|\theta)p(X_i|Z_i)))d\theta$

and we set constraint by Lagrange multiplier $\lambda(\int q(\theta)d\theta - 1)$, since $\int q(\theta)d\theta) = 1$

Recall the variational calculus,

- $\frac{\partial}{\partial q}\lambda(\int q(\theta)d\theta - 1) = \lambda\frac{\partial}{\partial q}(\int q(\theta)d\theta - 1) = \lambda \times 1$
- $\frac{\partial}{\partial q}\int_\theta q(\theta)(log(p(\theta))d\theta = log(p(\theta))$
- $\frac{\partial}{\partial q}\int_\theta q(\theta)(log(q(\theta))d\theta = log(q(\theta)) + 1$

$\frac{\partial}{\partial q}\sum_{i=1}^{n} \int q(\theta)E_q(log(p(Z_i|\theta)p(X_i|Z_i)))d\theta = \frac{\partial}{\partial q}\int \sum_{i=1}^{n} q(\theta)E_q(log(p(Z_i|\theta)p(X_i|Z_i)))d\theta = \sum_{i=1}^{n} q(\theta)E_q(log(p(Z_i|\theta)p(X_i|Z_i)))$

Therefore,

$\frac{\partial}{\partial q}L_\theta(q(\theta)) = \lambda + log(p(\theta)) - (log(q(\theta)) + 1) + \sum_{i=1}^{n} E_q(log(p(Z_i|\theta)p(X_i|Z_i))) = 0$

Then we solve this equation,

$q(\theta) = p(\theta)exp(\lambda - 1)\prod_{i=1}^{n} exp(E(log(p(Z_j|\theta)p(X_i|Z_i)))$,

and,

$\int q(\theta)d\theta = \int p(\theta)exp(\lambda{-}1)\prod_{i=1}^{n} exp(E(log(p(Z_j|\theta)p(X_i|Z_i)))d\theta = exp(\lambda{-}1)\int p(\theta)\prod_{i=1}^{n} exp(E(log(p(Z_j|\theta)p(X_i|Z_i)))d\theta = 1$

therefore, $\lambda = 1 - log(\int p(\theta)\prod_{i=1}^{n} exp(E(log(p(Z_j|\theta)p(X_i|Z_i)))d\theta)$

Plug-in λ, we can get a very clean form,

$q(\theta) = \frac{p(\theta)\prod_{i=1}^{n} exp(E(log(p(Z_j|\theta)p(X_i|Z_i)))}{\int p(\theta)\prod_{i=1}^{n} exp(E(log(p(Z_j|\theta)p(X_i|Z_i)))d\theta}$.

Note that the above derivation is generic and does not assume any distribution.

Note we are going to use the assumption of distributions.

Focus on $E(log(p(Z_j|\theta)p(X_i|Z_i)))$, recall that $p(Z_i|\theta)$ is the binomial distribution with $BINOMIAL(2, \theta)$ and $p(X_i|Z_i)$ is the binomial distribution with $BINOMIAL(n_i, \frac{Z_i}{2})$.

And we finally figure out that $q(\theta)$ is $Dirichlet(\alpha + \sum(E_Z(Z_i = 0|\theta, X_i)), \beta + \sum(E_Z(Z_i = 1|\theta, X_i)))$

Similarly , we can get the same form of $q(Z_i)$, using the same optimizing procedure.

Expectation-Maximization algorithm can be treated as a problem to minimize the KL divergence.

Limitation

(1)The independence assumption may not be met.

(2)The KL divergence will not penalize when p is large and q is small.