Variational Inference

Chunpai Wang

June 01, 2016

Goal of Inference:

- Computing the likelihood of observed data (in models with latent variables)
- Computing the marginal distribution over a given subset of nodes in the model
- Computing the conditional distribution over a subsets of nodes given a disjoint subset of nodes
- Computing a mode of the density (for the above distributions)

Approximate Inferences:

- Variational Inference
  - Mean field approximation
  - Expectation propagation
  - Variational 2nd-order Taylor approximation
- Markov Chain Monte Carlo
  - Gibbs sampling
  - Stochastic gradient MCMC methods

Given a set of i.i.d. observed data $X = \{x_1, , , , ., x_N\}$, a set of latent variables and parameters $Y = \{y_1, ..., y_N\}$, and the joint distribution $P(X, Y)$, the goal of variational inference is to find an approximation for the posterior distribution $P(Y|X)$ as well as for the model evidence $P(X)$.

1 KL-Divergence

We would like to find a distribution $q(Y)$ to approximate the $p(Y|X)$ that minimize the KL-divergence

$$KL(q(Y)\|p(Y|X)) = \int q(Y) \log \frac{q(Y)}{p(Y|X)} dY \tag{1}$$

However, it is intractable to compute the $p(Y|X)$ for each possible value of variable Y, because it requires to compute the normalization term $p(X)$ so that we can get

$$p(Y|X) = \frac{p(X, Y)}{p(X)} \tag{2}$$

To avoid this, we change to minimize the KL-divergence between $q(Y)$ and $p(X, Y)$

$$\begin{aligned} KL(q(Y)\|p(X, Y)) &= KL(q(Y)\|p(Y|X)p(X)) && (3)\\ &= \int q(Y) \log \frac{q(Y)}{p(Y|X)p(X)} dY && (4)\\ &= \int q(Y) \left[\log \frac{q(Y)}{p(Y|X)} - \log p(X)\right] dY && (5)\\ &= \int q(Y) \log \frac{q(Y)}{p(Y|X)} dY - \log p(X) && (6)\\ &= KL(q(Y\|p(Y|X))) - \log p(X) && (7) \end{aligned}$$

We can see that to minimize the $KL(q(Y)\|p(X,Y))$ is same as to maximize the $-KL(q(Y)\|p(X,Y))$

$$
\begin{aligned}
-KL(q(Y)\|p(X,Y)) &= -\int q(Y)\log\frac{q(Y)}{p(X,Y)}dY && (8)\\
&= \int q(Y)\log\frac{p(X,Y)}{q(Y)}dY && (9)\\
&= E_{q(Y)}\left[\log p(X,Y)\right] - E_{q(Y)}\left[\log q(Y)\right] && (10)\\
& && (11)
\end{aligned}
$$

which is known as the evidence lower bound of log-likelihood of the observed variable, since we have

$$-KL(q(Y)\|p(X,Y)) = \log p(X) - KL(q(Y\|p(Y|X))) \leq \log p(X) \tag{12}$$

2 EM-Recap

Here, we will show the convergence of EM algorithm from functional perspective. As we know, the EM is to iteratively maximize the log likelihood of observed data, which is

$$L(\theta|X) = \log p(X|\theta) = \log \int_y p(X,Y|\theta)dy \tag{13}$$

Any distribution over the hidden variables, denoted by $q(Y)$, can be used to obtained a lower bound on the log-likelihood using Jensen's inequality, which can be expressed as

$$
\begin{aligned}
L(\theta|X) &= \log \int_y q(Y)\frac{p(X,Y|\theta)}{q(Y)}dy && (14)\\
&= \log E_{q(Y)}\left[\frac{p(X,Y|\theta)}{q(Y)}\right] && (15)\\
&\geq E_{q(Y)}\left[\log\frac{p(X,Y|\theta)}{q(Y)}\right] && (16)\\
&= \int_y q(Y)\log\frac{p(X,Y|\theta)}{q(Y)}dy && (17)\\
&= F(q,\theta) && (18)
\end{aligned}
$$

where $F(q,\theta)$ denotes the functional, which is always lower bound on the log-likelihood. In the EM algorithm, we alternatively optimize over q and θ.

First, we can find the lower bound of the log-likelihood using functional $F(q,\theta)$:

$$
\begin{aligned}
F(q,\theta) &= \int_y q(Y)\log\frac{p(X,Y|\theta)}{q(Y)}dy && (19)\\
&= \int_y q(Y)\log\frac{p(Y|X,\theta)\cdot p(X|\theta)}{q(Y)}dy && (20)\\
&= \int_y q(Y)\log\frac{p(Y|X,\theta)}{q(Y)}dy + \int_y q(Y)\log p(X|\theta)dy && (21)\\
&= -KL(q(Y)\|p(Y|X,\theta)) + L(\theta|X) && (22)\\
&= L(\theta|X) - KL(q(Y)\|p(Y|X,\theta)) && (23)\\
& && (24)
\end{aligned}
$$

since KL divergence is alway non-negative, we have

$$L(\theta|X) = F(q,\theta) + KL(q(Y)\|p(Y|X,\theta)) \geq F(q,\theta) \tag{25}$$

which means $F(q,\theta)$ is always lower bound of log-likelihood of observed data, when θ is fixed. Thus, at k-th iteration of E-step, we would like to **maximize $F(q,\theta)$ (the lower bound of $L(\theta^{(k-1)}|X)$) w.r.t the distribution over hidden variable given the parameters**. $KL[q\|p] = 0$ if and only if $q = p$, so we can simply use the maximal $q(Y) = p(Y|X,\theta^{(k-1)})$ at k-th iteration of E-step.

Secondly,

$$F(q,\theta) = \int_y q(Y)\log\frac{p(X,Y|\theta)}{q(Y)}dy \tag{26}$$

$$= \int_y q(Y)\log p(X,Y|\theta)dy - \int_y q(Y)\log q(Y)dy \tag{27}$$

$$= \int_y q(Y)\log p(X,Y|\theta)dy + H(q) \tag{28}$$

$$= \int_y p(Y|X,\theta)\log p(X,Y|\theta)dy + H(q) \tag{29}$$

$$= E_{Y|X}\left[\log p(X,Y|\theta)\right] + H(q) \tag{30}$$

After we fix the q, we can optimize $F(\theta, X)$ w.r.t θ.

In summary,
E-step:

$$q^{(k)}(Y) := \arg\max_{q(Y)} F(q(Y), \theta^{(k-1)}) \tag{31}$$

$$= \arg\max_{q(Y)} L(\theta^{(k-1)}|X) - KL[q(Y)\|p(Y|X,\theta^{(k-1)})] \tag{32}$$

M-step:

$$\theta^{(k)} = \arg\max_{\theta} F(q^{(k)}(Y), \theta^{(k-1)}) \tag{33}$$

$$= \arg\max_{\theta} E_{Y|X}(\log p(X,Y|\theta)) + H(q) \tag{34}$$

$$= \arg\max_{\theta} E_{Y|X}(\log p(X,Y|\theta)) \tag{35}$$

$$= \arg\max_{\theta} \int_y p(Y|X,\theta^{(k-1)}) \cdot \log p(X,Y|\theta)dy \tag{36}$$

$$= \arg\max_{\theta} \int_y q^{(k)}(Y)\log p(X,Y|\theta)dy \tag{37}$$

where the last equation is from E-step $q^{(k)} = p(Y|X,\theta^{(k-1)})$

$$L(\theta^{(k-1)}|X) \underbrace{=}_{\text{E-step}} F(q^{(k)},\theta^{(k-1)}) \underbrace{\leq}_{\text{M-step}} F(q^{(k)},\theta^{(k)}) \underbrace{\leq}_{\text{Jensen Inequality or Non-negative KL}} L(\theta^{(k)}|X) \tag{38}$$

3 Goal of Variational Inference

We can see that any distribution $q(Y)$ over the hidden variable (which can be the distribution parameters of observed data) or missing data Y in EM, can be used to obtained a lower bound $F(q,\theta)$ on the log-likelihood $L(\theta|X)$, where

$$L(\theta|X) = F(q(Y),\theta) + KL(q\|p(Y|X)) \tag{39}$$

The $F(q,\theta)$ is also known as the evidence lower bound (ELBO). Our ultimate target is to maximize the log-likelihood on observed data, and opt to optimize the RHS. Now, we turn the inference into an optimization problem.

Intuitively, we can minimize the KL-divergence and everything is done. However, the KL-divergence contains the posterior distribution, which is hard to minimize. We opt to maximize the lower bound, which is equivalent to minimize the KL-divergence, our goal is to find $q(Y) = p(Y|X,\theta)$. The problem here is the posterior distribution $p(Y|X,\theta)$ is usually intractable.

Note, when KL-divergence is small, we can say q is good approximation of distribution p. The variational inference is used to approximate the posterior distribution or even more general problem. Variational inference will iteratively reach closely to the posterior distribution q.

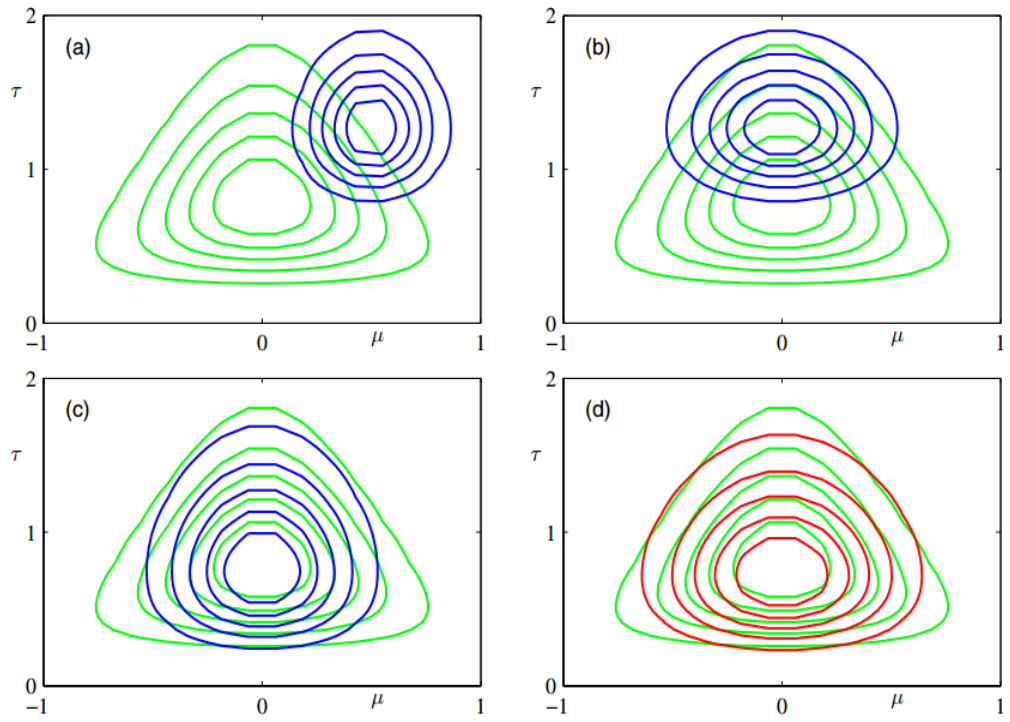

Figure 1: Visualization of variational inference. The green curve is the true distribution, which is hard to compute, and blue curve is our chosen simple distribution to approximate the true distribution.

4 Strategies

But how can we do that ? The distribution q should comprise some properties to ensure us to do this kind of approximation. Note, we can also use other metric rather than KL-divergence to measure the difference between p and q, but KL-divergence makes life easier.

There are several methods to approximate a complicated distribution:

1. restrict the family sufficiently that they comprise only tractable distribution
2. meanwhile, allow the family to be sufficiently rich and flexible that it can provide a good approximation to the true posterior distribution

Two ways to restrict the family of approximating distribution:

1. just use a parametric distribution $q(Y|w)$ governed by a set of parameters w. The lower bound $F(q, \theta)$ becomes a function of w, which can be optimized
2. Factorized distribution. Suppose we can partition latent elements of Y into disjoint groups that denote by Y_i where $i = 1, ..., M$, and we assume

$$q(Y) = \prod_{i=1}^{M} q_i(Y_i) \tag{40}$$

which is also known as **Mean Field Theory**, where this assumption can simplify our formulation later. Mean-field variational inference casts Bayesian computation as optimization.

5 Mean Field Variational Inference

We assume the variational distribution over the latent variables factorizes as

$$q(Y) = q(Y_1, ..., Y_M) = \prod_{i=1}^{M} q_i(Y_i) \tag{41}$$

This family does not contain the true posterior because it assumes the hidden variables are independent, since the dependencies are often what makes the posterior difficult to work with.

$$F(q, \theta) = \int_y q(Y) \log \left\{ \frac{p(X, Y)}{q(Y)} \right\} dy \tag{42}$$

$$= \int_y \prod_{i=1}^{M} q_i(Y_i) \left\{ \log p(X, Y) - \sum_{i=1}^{M} \log q_i(Y_i) \right\} dy \tag{43}$$

$$= \underbrace{\int_y \prod_{i=1}^{M} q_i(Y_i) \log p(X, Y) dy}_{part1} - \underbrace{\int_y \prod_{i=1}^{M} q_i(Y_i) \sum_{i=1}^{M} \log q_i(Y_i) dy}_{part2} \tag{44}$$

For part 1, we can expand the integrals for each variable and rearrange the expression by taking a particular $q_j(Y_j)$ out of the integral:

$$part1 = \int_{Y_1} \int_{Y_2} ... \int_{Y_M} \prod_{i=1}^{M} q_i(Y_i) \log p(X, Y) dY_1 dY_2 ... dY_M \tag{45}$$

$$= \int_{Y_j} q_j(Y_j) \Big(\underbrace{\int_{Y_{i \neq j}} \cdots \int \log p(X, Y) \cdot \underbrace{\prod_{i \neq j}^{M} [q_i(Y_i) dY_i]}_{\text{relative independence from mean-field theorem}}}_{\text{after integral, this term remains } \log p(X, Y_j)} \Big) dY_j \tag{46}$$

$$= \int_{Y_j} q_j(Y_j) \Big[E_{i \neq j}[\log(p(X, Y))] \Big] dY_j \tag{47}$$

Note that, we cannot always be able to rearrange the order of integrals.

For part 2, we also needs to integrate out all $Y = \{Y_1, ..., Y_M\}$. However, notice that each term in the sum, $\sum_{i=1}^{M} \log(q_{i(Y_i)})$ involves only a single i, therefore, we are able to simplify further

$$part2 = \int_{Y_1} \cdots \int_{Y_M} \prod_{i=1}^{M} q_i(Y_i) \sum_{i}^{M} \log q_i(Y_i) dY_i \cdots dY_M \tag{48}$$

$$= \int_{Y_1} \cdots \int_{Y_M} \log q_1(Y_1) \prod_{i=1}^{M} q_i(Y_i) dY_i \cdots dY_M + \cdots + \int_{Y_1} \cdots \int_{Y_M} \log q_M(Y_M) \prod_{i=1}^{M} q_i(Y_i) dY_i \cdots dY_M \tag{49}$$

$$= \int_{Y_1} \log q_1(Y_1) \int_{Y_2} \cdots \int_{Y_M} \prod_{i=1}^{M} q_i(Y_i) dY_i \cdots dY_M + \cdots + \int_{Y_M} \log q_M(Y_M) \int_{Y_1} \cdots \int_{Y_{M-1}} \prod_{i=1}^{M} q_i(Y_i) dY_i \cdots dY_M \tag{50}$$

$$= \int_{Y_1} [\log q_1(Y_1)] \cdot q_1(Y_1) dY_1 + \cdots + \int_{Y_M} [\log q_M(Y_M)] \cdot q_M(Y_M) dY_M \tag{51}$$

$$= \sum_{j=1}^{M} \Big(\int_{Y_j} [\log q_j(Y_j)] \cdot q_j(Y_j) dY_j \Big) \tag{52}$$

For a particular $p_j(Y_j)$, the rest of the sum can be treated like a constant, part 2 can be written as:

$$part2 = \int_{Y_j} [\log q_j(Y_j)] \cdot q_j(Y_j) dY_j + const. \tag{53}$$

Combine part 1 and part 2, and we let

$$\log \tilde{p}(X, Y_j) = E_{i \neq j}[\log(p(X, Y))] \tag{54}$$

$$\begin{aligned}
F(q_j, \theta) &= \int_{Y_j} q_j(Y_j) \Big[E_{i \neq j}[\log(p(X, Y))] \Big] dY_j - \sum_{j=1}^{M} \Big(\int_{Y_j} [\log q_j(Y_j)] \cdot q_j(Y_j) dY_j \Big) && (55) \\
&= \int_{Y_j} q_j(Y_j) \Big[E_{i \neq j}[\log(p(X, Y))] \Big] dY_j - \int_{Y_j} [\log q_j(Y_j)] \cdot q_j(Y_j) dY_j + const. && (56) \\
&= \int_{Y_j} q_j(Y_j) \Big[\log \tilde{p}(X, Y_j) \Big] dY_j - \int_{Y_j} [\log q_j(Y_j)] \cdot q_j(Y_j) dY_j + const. && (57) \\
&= \int_{Y_j} q_j(Y_j) \log \frac{\tilde{p}(X, Y_j)}{q_j(Y_j)} dY_j + const. && (58) \\
&= -KL(q_j(Y_j) \| \tilde{p}(X, Y_j)) + const. && (59)
\end{aligned}$$

Therefore, our goal is to maximize the $F(q_j, \theta)$, which is same as minimize $KL(q_j(Y_j \| \tilde{p}(X, Y_j)))$ and the minimum occurs when

$$\begin{aligned}
q_j^*(Y_j) &= \tilde{p}(X, Y_j) && (60) \\
\log q_j^*(Y_j) &= \log \tilde{p}(X, Y_j) = E_{i \neq j}[\log(p(X, Y))] && (61)
\end{aligned}$$

5.1 Coordinate Ascent Inference

We will use coordinate ascent inference, iteratively optimizing each variational distribution holding the others fixed.

1. initialize every $q_j(Y_j)$
2. calculate $q_j^*(Y_j)$ with all other fixed
3. the iterative procedure will converge because the $F(q_j, \theta)$ is convex for each factor $q_j(Y_j)$.

However, this is not the only possible optimization algorithm.

5.2 Posterior Conditional

Now, we will show another way to get the coordinate ascent update. Recall that

$$\begin{aligned}
F(q, \theta) &= \int_y q(Y) \log \left\{ \frac{p(X, Y)}{q(Y)} \right\} dy && (62) \\
&= \int_y q(Y) \log \{p(X, Y)\} \, dy - \int_y q(Y) \log \{q(Y)\} \, dy && (63) \\
&= E_{q(Y)}[\log(p(X_{1:N}, Y_{1:M}))] - E_{q(Y)}[\log(q(Y_{1:M}))] && (64) \\
& && (65)
\end{aligned}$$

First, recall that the probability chain rule gives:

$$p(Y_{1:M}, X_{1:N}) = p(X_{1:N}) \prod_{j=1}^{M} p(Y_j | Y_{1:j-1}, X_{1:N}) \tag{66}$$

Note that, the latent variables in this product can occur in any order ! This will be important later.

Second, we can decompose the entropy term of the ELBO (using the mean field variational approximation) as

$$E_{q(Y)}[\log(q(Y_{1:M}))] = \sum_{j}^{M} E_{q_j(Y_j)}[\log(q_j(Y_j))] \tag{67}$$

Then, we can decompose the ELBO for the mean field variational approximation into a nice form:

$$F(q,\theta) = E_{q(Y)}[\log(p(X_{1:N}, Y_{1:M}))] - E_{q(Y)}[\log(q(Y_{1:M}))] \tag{68}$$

$$= E_{q(Y)}\big[\log(p(X_{1:N}) \prod_{j=1}^{M} p(Y_j|Y_{1:j-1}, X_{1:N})\big] - E_{q(Y)}[\log(q(Y_{1:M}))] \tag{69}$$

$$= E_{q(Y)}\big[\log(p(X_{1:N}) + \sum_{j=1}^{M} \log p(Y_j|Y_{1:j-1}, X_{1:N})\big] - E_{q(Y)}[\log(q(Y_{1:M}))] \tag{70}$$

$$= \log(p(X_{1:N}) + E_{q(Y)}\big[\sum_{j=1}^{M} \log p(Y_j|Y_{1:j-1}, X_{1:N})\big] - E_{q(Y)}[\log(q(Y_{1:M}))] \tag{71}$$

Since we assume the latent variable Y_j are independent, we have

$$p(Y_j|Y_{1:j-1}, X_{1:N})) = p(Y_j|Y_{-j}, X_{1:N}) \tag{72}$$

where the notation $-j$ denotes all indices other than the j^{th}, which is also called the "posterior conditional" of Y_j, given all other latent variables and observations. **This posterior conditional is very important in mean field variational bayes, and will be important in Gibbs sampling.**

Next, we want to derive the coordinate ascent update for a latent variable, keeping all other latent variables fixed. Removing the parts that do not depend on $q_j(Y_j)$, we can write,

$$\arg\max_{q_j} F(q,\theta) \tag{73}$$

$$= \arg\max_{q_j}(E_q[\log p(Y_j|Y_{-j}, X_{1:N})] - E_{q_j(Y_j)}[\log(q_j(Y_j))]) \tag{74}$$

$$= \arg\max_{q_j}\Big(\int q_j(Y_j) E_{q_{-j}}[\log p(Y_j|Y_{-j}, X_{1:N})]dY_j - \int q_j(Y_j)\log q_j(Y_j)dY_j\Big) \tag{75}$$

$$= \arg\max F(q_j,\theta) \tag{76}$$

- To find the argmax, we take the derivative of $F(q_j,\theta)$ with respect to $q_j(Y_j)$, use Lagrange multipliers, and set the derivative to zero:

$$\frac{dF(q_j,\theta)}{dq_j(Y_j)} = E_{q_{-j}}[\log p(Y_j|Y_{-j}, X)] - \log q_j(Y_j) - 1 = 0 \tag{77}$$

- From this, we arrive at the coordinate ascent update:

$$q_j^*(Y_j) \propto \exp\{E_{q_{-j}}[\log p(Y_j|Y_{-j}, X)]\} \tag{78}$$

- Based on the Bayes' rule, since the denominator of the conditional does not depend on Y_j, we can equivalently write

$$q_j^*(Y_j) \propto \exp\{E_{q_{-j}}[\log p(Y_j, Y_{-j}, X)]\} \tag{79}$$

5.3 Exponential Family Conditionals

Even though we have the coordinate ascent update:

$$q_j^*(Y_j) \propto \exp\{E_{q_{-j}}[\log p(Y_j|Y_{-j}, X)]\} \tag{80}$$

which is not a closed form, it is also difficult to compute. The question here is that is there a general form for models in which the coordinate updates in mean field variational inference are easy to compute and lead to closed-form updates? **Yes, the answer is exponential family conditionals**, i.e. models with conditional densities that are in an exponential family, i.e. of the form:

$$p(Y_j|Y_{-j}, X)] = h(Y_j)\exp\{\eta(Y_{-j}, X)^\top \phi(Y_j) - A(\eta(Y_{-j}, X))\} \tag{81}$$

where η, h, A and ϕ are functions that parameterize the exponential family. And different choices of these parameters lead to many popular densities (normal, gamma, exponential, Bernouilli, Dirichlet, categorical, beta, Poisson, geometric, etc.). We call these **exponential-family-conditional models**, a.k.a **conditionally conjugate models**.

We can derive a general formula for the coordinate ascent update for all exponential-family conditional models.

1. we will choose the form of our **local variational approximation** $q(Y_j; \Lambda_j)$ to be the same as the conditional distribution (i.e. in an exponential family), where Λ_j are parameters for this local distribution.
2. When we perform our coordinate ascent update, we will see that the update yields an optimal $q(Y_j; \Lambda_j)$ in the same family, which only change the value of Λ_j, which significantly simplify the computation.

We need to go through one example about the exponential family. Please check the note Exponential Family for details.

6 Stochastic Gradients of the ELBO

The classical variational inference is inefficient when data is massive:

- Do some local computation for each data point
- aggregate these computation to re-estimate global structure
- repeat

Stochastic variational inference scales it to massive data.

7 Black-Box Variational Inference

References

[1] Variation Inference Tutorial https://www.zhihu.com/question/41765860 https://www.bilibili.com/video/av24062247/ https://www.bilibili.com/video/av24075851/ https://www.bilibili.com/video/av24093797/

[2] David Blei's Talk https://www.youtube.com/watch?v=Dv86zdWjJKQ

[3] Eric Xing's Lecture https://www.cs.cmu.edu/~epxing/Class/10708-15/slides/lecture13-VI.pdf

[4] Variational Methods by Zoubin Ghahramani http://www.cs.cmu.edu/~tom/10-702/Zoubin-702.pdf

[5] NIPS Tutorial on Variational Inference https://media.nips.cc/Conferences/2016/Slides/6199-Slides.pdf

[6] Black Box Variational Inference http://www.cs.columbia.edu/~blei/papers/RanganathGerrishBlei2014.pdf

[7] LDA Tutorial http://www.cnblogs.com/pinard/p/6831308.html

[8] CMU Lectures Spring 2017 https://www.cs.cmu.edu/~epxing/Class/10708-17/lecture.html