

# Variational Inference with Coin Toss example

AntiCodeOn  
email: anticodeon@gmail.com

September 2017

## 1 Introduction

Variational inference (VI) is a machine learning method for approximation of the difficult to compute probability densities that appear in Bayesian methods. In general, the goal is to find a surrogate distribution over hidden variables and parameters of the original Bayesian model that is close to the original distribution but where the evaluation is computationally efficient. The closeness is defined in terms of a divergence measure between the approximate distribution and the original distribution.

## 2 Motivation

Consider classical Bayesian setup

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

where  $X = x_{1:n}$  denotes the observed data while  $Z = z_{1:k}$  denotes all the hidden variables and parameters of the model. Numerator of the fraction is called joint distribution

$$p(X, Z) = p(X|Z) p(Z) \quad (2)$$

and the denominator (often called evidence) is a marginal distribution

$$p(X) = \int_Z p(X, Z) dZ \quad (3)$$

Computing the posterior probability is often a very hard problem (for example, due to the exponentially large number of hidden states or because required integrations do not have a closed-form analytical solution). This is why we resort to techniques such as approximation, where the objective is to find a joint distribution  $q(Z)$  that replaces original posterior  $p(Z|X)$  in a way that enables us to find computationally tractable solutions.

## 2.1 Variational inference setup

Variational inference posits a set of densities  $\mathcal{Q}$  over the latent variables  $Z$ . We typically try to minimize the Kullback-Leibler (KL) divergence

$$KL(q(Z) || p(Z|X)) = E_q \left[ \log \frac{q(Z)}{p(Z|X)} \right] \quad (4)$$

of the approximate and the original, true joint distribution

$$q^*(Z) = \arg \min_{q(Z) \in \mathcal{Q}} KL(q(Z) || p(Z|X)) \quad (5)$$

KL divergence tells us how much information we loose by choosing the approximate distribution instead of true posterior distribution. We can expand the right hand side of (4)

$$KL(q(Z) || p(Z|X)) = E_q[q(Z)] - E_q[\log p(Z|X)] \quad (6)$$

Using (1) we get

$$KL(q(Z) || p(Z|X)) = p(X) - \underbrace{(E_q[\log p(X, Z)] - E_q[q(Z)])}_{ELBO(q)} \quad (7)$$

We want to minimize the KL divergence by varying distribution  $q(Z)$ .  $p(X)$  does not depend on  $q$  and is always positive (or equal to zero) so we need to maximize the  $ELBO(q)$ . The only assumption we are making is that  $q(Z)$  factorizes in the following way

$$q(Z) = \prod_{k=1}^K q_k(z_k) \quad (8)$$

which means that we are breaking possible dependencies between hidden variables in our approximation process.

$$ELBO(q) = \int q(Z) \log p(X, Z) dZ - \int q(Z) \log q(Z) dZ \quad (9)$$

## 3 Variational inference application

$$p(Z|\pi) = \prod_{n=1}^N \prod_{k=1}^K \pi_k^{z_{nk}} \quad (10)$$

$$p(\pi) = Dir(\pi|\alpha_0) = C(\alpha_0) \prod_{k=1}^K \pi_k^{\alpha_k - 1} \quad (11)$$

$$p(X|Z, \Theta) = \prod_{n=1}^N \prod_{k=1}^K \prod_{j=1}^J [\Theta_k^{x_n^j} (1 - \Theta_k)^{1-x_n^j}]^{z_{nk}} \quad (12)$$

$$p(\Theta|a, b) = \prod_{k=1}^K \frac{\Theta^{a-1}(1-\Theta)^{b-1}}{\text{Beta}(a, b)} \quad (13)$$

Joint distribution is given by

$$p(X, Z, \Theta, \pi, a, b) = p(X|Z, \Theta)p(Z|\pi)p(\pi)p(\Theta|a, b) \quad (14)$$

with factors on the right side of the equation as defined above. Only variables  $X = x_1, \dots, x_N$  were observed. In the next step we consider a variational distribution with the following factorization

$$q(Z, \pi, \Theta) = q(Z)q(\pi, \Theta). \quad (15)$$

We are separating the latent variables and parameters and this is actually the only assumption we are making. In order to proceed with the factorization, we use the general result given by Bishop[2006].

$$\ln q^*(Z) = E_{\pi, \Theta}[\ln p(X, Z, \Theta, \pi, a, b)] + \text{const} \quad (16)$$

For the factor of  $\mathbf{Z}$  we are only interested in those components of the decomposition 14 that depend on  $\mathbf{Z}$ . All other terms are absorbed into the additive normalization constant, giving:

$$\ln q^*(Z) = E_{\pi}[\ln p(Z|\pi)] + E_{\Theta}[\ln p(X|Z, \Theta)] + \text{const} \quad (17)$$

where

$$E_{\pi}[\ln p(Z|\pi)] = \sum_{n=1}^N \sum_{k=1}^K z_{nk} E_{\pi}[\ln \pi_k] \quad (18)$$

and

$$E_{\Theta}[\ln p(X|Z, \Theta)] = \sum_{n=1}^N \sum_{k=1}^K z_{nk} E_{\Theta}[\sum_{j=1}^J \ln \Theta_k^{x_n^j} (1 - \Theta_k)^{1-x_n^j}] \quad (19)$$

Hence

$$\ln q^*(Z) = \sum_{n=1}^N \sum_{k=1}^K z_{nk} (E_{\pi}[\ln \pi_k] + E_{\Theta}[\ln \Theta_k^{n_H} (1 - \Theta_k)^{n_T}]) + \text{const} \quad (20)$$

where  $n_H$  is a number of heads and  $n_T$  number of tails in the  $n$ -th experiment. Proceeding similarly for  $\pi$  and  $\Theta$  we get the following expression

$$\ln q^*(\pi, \Theta) = E_Z[\ln(\pi) + \ln p(Z|\pi)] + E_Z[\ln p(X|Z, \Theta) + \ln p(\Theta)] + \text{const} \quad (21)$$

Since there are no expressions that depend on both parameters at the same time we may proceed with the factorization of the parameters independently

$$q(\pi, \Theta) = q(\pi) q(\Theta) \quad (22)$$

where

$$\ln q^*(\pi) = \ln p(\pi) + E_Z[\ln p(Z|\pi)] + \text{const} \quad (23)$$

$$\ln q^*(\pi) = \sum_{k=1}^K (\alpha_k - 1) \ln \pi_k + \sum_{k=1}^K r_k \ln \pi_k + \text{const} \quad (24)$$

where  $r_k = \sum_{n=1}^N E[z_{nk}]$ . We have already derived the expectations over  $Z$  in EM document. Taking the exponential of equation 24 we get that the

$$q(\pi) = \text{Dir}(\pi|\alpha) \quad (25)$$

where  $\alpha$  is vector with  $K$  components  $\alpha_k = \alpha_k + r_k$

$$\ln q^*(\Theta) = \ln p(\Theta) + E_Z[\ln p(X|Z, \Theta)] + \text{const} \quad (26)$$

$$\ln q^*(\Theta) = \sum_{k=1}^K (\ln [\Theta_k^{a_k-1} (1 - \Theta_k)^{b_k-1}] + \sum_{n=1}^N r_{nk} \sum_{j=1}^J \ln [\Theta_k^{x_n^j} (1 - \Theta_k)^{1-x_n^j}]) + \text{const} \quad (27)$$

$$\ln q^*(\Theta) = \sum_{k=1}^K \ln [\Theta_k^{a_k-1+n_H r_k} (1 - \Theta_k)^{b_k-1+n_T r_k}] + \text{const} \quad (28)$$

where  $n_H$  equals number of the outcomes of  $x_n^j = 1$  (number of heads), and  $n_T$  number of the outcomes of  $x_n^j = 0$  (number tails). We see that the factor for  $\Theta$  is Beta distributed variable:

$$q^*(\Theta) = \prod_{k=1}^K \text{Beta}(\Theta|A, B) \quad (29)$$

where  $A$  is a vector with parameter components  $a = a_k + n_H r_k$  and  $B$  with components  $b = b_k + n_T r_k$ .

## 4 Variational lower bound

Variational lower bound  $L$  is given by

$$L = E[\ln p(X, Z, \Theta, \pi, a, b)] - E[\ln q(Z, \Theta, \pi, a, b)] \quad (30)$$

Using results from 14, 15 and 22 this simplifies to

$$\begin{aligned} L = & E[\ln p(X|Z, \Theta)] + E[\ln p(Z|\pi)] + E[\ln p(\pi)] + E[\ln p(\Theta|a, b)] \\ & - E[\ln q(Z)] - E[\ln q(\pi)] - E[\ln q(\Theta)] \end{aligned} \quad (31)$$

We write down each of the terms separately:

$$E[\ln p(\pi)] = \ln C(\alpha_0) + \sum_{k=1}^K (\alpha_k - 1) (\psi(\alpha_k) - \psi(\sum_{k=1}^K \alpha_k)) \quad (32)$$

$$E[\ln p(\Theta_k|a, b)] = \sum_{k=1}^K -\ln \text{Beta}(a_k, b_k) + (a_k - 1) \psi(a_k) - (b_k - 1) \psi(b_k) + (a_k + b_k - 2) \psi(a_k + b_k) \quad (33)$$

$$E[\ln p(Z|\pi)] = \sum_{n=1}^N \sum_{k=1}^K r_{nk} (\psi(\alpha_k) - \psi(\sum_{k=1}^K \alpha_k)) \quad (34)$$

$$E[\ln p(X|Z, \Theta)] = \sum_{n=1}^N \sum_{k=1}^K r_{nk} [n_H(\psi(a_k) - \psi(a_k + b_k)) + n_T(\psi(b_k) - \psi(a_k + b_k))] \quad (35)$$

## 5 Algorithm

*This section should contain algorithmic description of the problem*