Lecture Notes
Part III: Black-Box Variational Inference

Jalil Taghia jalil.taghia@it.uu.se
Department of Information Technology
Uppsala University
Uppsala, Sweden

1. Introduction

Let $x = \{x_1, \dots, x_n\}$ denote a set of observed variables and $z = \{z_1, \dots, z_n\}$ denote a set of latent variables which govern the distribution of data. Consider a generic probabilistic model in the form of: $p(x, z) = p(x \mid z)p(z)$. Our objective is to find an approximation to the intractable posterior distribution, $p(z \mid x)$.

For conjugate model classes such as the exponential family, we showed that the mean-field variational inference (VI) can be used to approximate the posterior distribution. Among other limitations associated with the mean-field assumption, the algorithm explicitly requires iterating through the entire data set at each iteration which limits its application to large data. For the same family of models, we discussed that we can use stochastic optimization together with natural gradients to obtain a scalable variational inference, referred to as the stochastic variational inference (refer to Lecture Notes – PART II). Stochastic optimization using noisy (natural) gradients enables application of the variational inference to large data.

The variational inference studied so far was limited to conditionally conjugate models, for which the evidence lower bound (ELBO) could be computed analytically. For *non-conjugate models* and a *less restrictive choice of variational family of distributions*, computing the ELBO may remain intractable due to the required expectations which are not in general computable in analytically closed forms. Although local variational inference methods (Bishop, 2006, Ch. 10.5) can be used to derive tractable bounds—for example in Bayesian logistic regression (Bishop, 2006, Ch. 10.6)—but in general that requires *model-specific analysis* which may still be intractable for more complex models—such as, deep Bayesian generative models—or require outstanding mathematical expertise.

In this note, we discuss automated variational inference methods that relax this requirement and simplify the inference. The idea is to develop a generic inference algorithm for which only the generative process of the data has to be specified. We shall see here that central to this idea are *stochastic gradient estimators of the ELBO*. Here we use the *black-box variational inference* (BBVI) as an umbrella term to refer to the techniques which rely on this idea. The goal in BBVI is to obtain Monte Carlo estimates of the gradient of the ELBO and to use stochastic optimization to fit the variational parameters.

2. Stochastic gradient of the evidence lower bound

Let $q_\phi(z)$ denote the variational posterior distribution parametrized with ϕ. For simplicity of discussion, also let us define

$$g_\phi(z) = \log p(x, z) - \log q_\phi(z). \tag{1}$$

As before, we start with writing the ELBO,

$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi(z)}[\log p(x,z) - \log q_\phi(z)] = \mathbb{E}_{q_\phi(z)}[g_\phi(z)]. \tag{2}$$

Recall that we are now interested in situations where it is no longer possible to analytically compute required expectations w.r.t. q for our generic model.

Our objective is to maximize the ELBO. The recipe is simple: We first directly take the gradient of the ELBO w.r.t. the variational parameter. The resulting gradient is represented as an expectation which is then evaluated using Monte Carlo techniques—by sampling from the variational distribution and forming the corresponding Monte Carlo estimates of the gradient.

We now take the gradient of $\mathcal{L}$, w.r.t. ϕ,

$$\nabla_\phi \mathcal{L}(\phi) = \nabla_\phi \mathbb{E}_{q_\phi(z)}\left[g_\phi(z)\right]. \tag{3}$$

We shall represent the above gradient of expectation as the expectation of the gradient. Computing the gradient w.r.t. ϕ needs extra attention since the expectation is being taken w.r.t. q and thus we cannot simply swap the gradient with the expectation. We will evaluate the gradient in two different ways resulting in two different estimators:

1. **Score function estimator:** The gradient of the ELBO is expressed as an expectation with respect to the variational distribution using the log-derivative trick, also known as REINFORCE or score function method, under which the Lebesgue's dominated convergence theorem becomes applicable so that we can take the gradient of the expectation in (3) by moving the gradient inside the expectation. It then takes samples from the variational distribution to calculate noisy gradients.

2. **Pathwise gradient estimator:** Based on the *reparameterization trick.* This approach reparameterizes the latent variable z in terms of a set of auxiliary variables ϵ with a *known distribution* $p(\epsilon)$ which does not depend on the variational parameters. This means that crucially the expectation in (3) can be now taken w.r.t the known distribution $p(\epsilon)$ as opposed to $q_\phi(z)$ which then enables us to simply move the gradient inside the expectation. The procedure only requires drawing samples from $p(\epsilon)$.

In the following we discuss each technique in more details.

3. Black-box variational inference using score function estimator

Recall our gradient of the lower bound, $\nabla_\phi \mathcal{L}(\phi)$. The gradient can be expressed using the Lebesgue's dominated convergence theorem (Appendix A) and the log-derivative identity $\nabla_\phi q_\phi(z) = q_\phi(z)\nabla_\phi \log q_\phi(z)$, as:

$$\begin{aligned}
\nabla_\phi \mathcal{L} &= \nabla_\phi \int q_\phi(z) g_\phi(z) \mathrm{d}z && \text{(4a)} \\
&= \int \nabla_\phi (q_\phi(z) g_\phi(z)) \mathrm{d}z && \text{(4b)} \\
&= \int g_\phi(z) \nabla_\phi q_\phi(z)\ \mathrm{d}z + \int q_\phi(z) \nabla_\phi g_\phi(z)\ \mathrm{d}z && \text{(4c)} \\
&= \mathbb{E}_{q_\phi(z)}[g_\phi(z)\nabla_\phi \log q_\phi(z)] + \mathbb{E}_{q_\phi(z)}[\nabla_\phi g_\phi(z)], && \text{(4d)}
\end{aligned}$$

where $\nabla_\phi \log q_\phi(z)$ is the *score function*[1] whose expectation is zero for any q,

$$\mathbb{E}_q[\nabla_\phi \log q_\phi(z)] = \int \frac{\nabla_\phi q_\phi(z)}{q_\phi(z)} q_\phi(z) \mathrm{d}z = \nabla_\phi \int q_\phi(z) \mathrm{d}z = \nabla_\phi 1 = 0. \tag{5}$$

[1] The gradient of the log of a probability distribution is often called score function.

```
Input: Joint model $\log p(x,z)$ and variational family $q_\phi(z)$
Output: Optimized variational parameter $\phi$
Initialization: Initialize $\phi$ randomly, set the initial step-size $\rho_t = \rho_0$, and set $t = 1$
repeat
    (1) Draw $L$ samples from $q_\phi(z)$, $z_l \sim q_\phi(z)$;
    (2) Update the variational parameter using the collected samples,
        $\phi = \phi + \rho_t \frac{1}{L} \sum_{l=1}^{L} \nabla_\phi \log q_\phi(z_l) \left( \log p(x, z_l) - \log q_\phi(z_l) \right);$
    (3) Set $t = t + 1$ and update $\rho_t$ accordingly to satisfy the Robbins and Monro conditions:
        $\sum_t \rho_t = \infty$ and $\sum_t \rho_t^2 < \infty$;
until convergence criteria is satisfied;
```

Algorithm 1: A basic black-box variational inference using the score function estimator.

Using this property, further we have: $\mathbb{E}_{q_\phi(z)}[\nabla_\phi \log q_\phi(z)] = \mathbb{E}_{q_\phi(z)}[\nabla_\phi g_\phi(z)] = 0$, and thus the gradient of the lower bound, in (4a), can be rewritten in terms of the score function as

$$\nabla_\phi \mathcal{L} = \mathbb{E}_{q_\phi(z)} \left[\nabla_\phi \log q_\phi(z) g_\phi(z) \right]. \tag{6}$$

We now can compute noisy unbiased gradient estimates using L Monte Carlo samples from the variational distribution,

$$\boxed{\nabla_\phi \mathcal{L} \simeq \frac{1}{L} \sum_{l=1}^{L} \nabla_\phi \log q_\phi(z_l) g_\phi(z_l), \quad \text{where } z_l \sim q_\phi(z).} \tag{7}$$

The requirements are minimal. We need to:

- sample from $q_\phi(z)$, i.e., $z_l \sim q_\phi(z)$,
- evaluate the score function at z_l, $\nabla_\phi \log q_\phi(z_l)$,
- and evaluate $\log p(x, z_l) - \log q_\phi(z_l)$.

An algorithmic representation of a basic black-box variational inference using the score function estimator is given in Algorithm 1. The algorithm can readily be extended to take advantage of the natural gradients by pre-multiplying the (classical) gradient with the inverse of the fisher information matrix (refer to Lecture Notes–Part II).

In practice, however, the variance of the estimator of the gradient under the Monte Carlo estimates can be very large—sampling rare values can lead to large scores and thus high variance. There is a body of research dedicated to addressing this shortcoming using *variance reduction* techniques. Two common methods are discussed in the following: Rao-Blackwellization (Ranganath et al., 2014) and control variate, (Paisley et al., 2012; Ranganath et al., 2014).

Rao-Blackwellization[2] Rao-Blackwellization reduces the variance of a random variable by replacing it with its conditional expectation w.r.t. a subset of the variables. Consider two random variables X and Y, and a function $J(X,Y)$. Now consider a function $\widehat{J}(X) = \mathbb{E}[J(X,Y) \mid X]$. Indeed $\widehat{J}(X)$ has the same expectation as $J(X)$, that is $\mathbb{E}[\widehat{J}(X)] = \mathbb{E}[J(X,Y)]$. Importantly that means in

[2] This section is based on (Ranganath et al., 2014). Refer to the original material for more details.

our Monte Carlo approximation, we can use $\widehat{J}(X)$ in place of $\mathbb{E}[J(X,Y)]$ with the desirable property that $\widehat{J}(X)$ is a lower (or equal when $J(X,Y)=\widehat{J}(X)$) variance estimator than $J(X,Y)$,

$$\mathrm{Var}\left[\widehat{J}(X)\right] = \mathrm{Var}\left[J(X,Y)\right] - \mathbb{E}\left[\left(J(X,Y)-\widehat{J}(X)\right)^2\right].$$

Returning to our model, let $z_{(i)}$ be the Markov blanket of z_i, and $p_i(x, z_{(i)})$ be the terms in the joint that depend on those variables. Given a fully factorized mean-field family of variational distributions, $q_\phi(z) = \prod_i q_{\phi_i}(z_i)$, the gradient w.r.t. ϕ_i simplifies to

$$\nabla_{\phi_i}\mathcal{L} = \mathbb{E}_{q_{\phi_{(i)}}(z)}\left[\nabla_{\phi_i}\log q_{\phi_i}(z_i)\left(\log p_i(x, z_{(i)}) - \log q_{\phi_i}(z_i)\right)\right] \tag{8a}$$

$$\simeq \frac{1}{L}\sum_{l=1}^{L}\nabla_{\phi_i}\log q_{\phi_i}(z_i)\left(\log p_i(x, z_l) - \log q_{\phi_i}(z_l)\right), \quad \text{where } z_l \sim q_{\phi_{(i)}(z)} \tag{8b}$$

where $q_{\phi_{(i)}}(z)$ is the variational family of distributions corresponding to the Markov blanket of z_i[3].

Control variate[4] The key idea behind the variance reduction is to replace the target function, whose expectation is being approximated by Monte Carlo, with an axillary function that has *the same expectation* but a *smaller variance.*

Let $f(z)$ be the target function whose expectation is being estimated using Monte Carlo,

$$\nabla_\phi\mathcal{L} = \mathbb{E}_{q_\phi(z)}[\underbrace{\nabla_\phi\log q_\phi(z)\left(\log p(x,z) - \log q_\phi(z)\right)}_{f(z)}]. \tag{9}$$

Consider $h(z)$ as a control variate which approximates $f(z)$ in the space of $q_\phi(z)$. The control variate is chosen such that the axillary function $\widehat{f}(z)$ satisfies:

$$\mathbb{E}_q[\widehat{f}(z)] = \mathbb{E}_q[f(z)], \tag{10}$$

$$\mathrm{Var}_q[\widehat{f}(z)] < \mathrm{Var}_q[f(z)]. \tag{11}$$

Using h and a scalar $a \in \mathbb{R}$, a general class $\widehat{f}$ is defined as

$$\widehat{f} = f(z) - a(h(z) - \mathbb{E}_q[h(z)]). \tag{12}$$

This function has the same expectation as f and therefore can replace it in the lower bound. The scalar variable a is set such that the variance of $\widehat{f}$ is minimized. The variance of $\widehat{f}$ is given by

$$\mathrm{Var}[\widehat{f}] = \mathrm{Var}[f] - 2a\mathrm{Cov}[f,h] + a^2\mathrm{Var}[h], \tag{13}$$

from which the optimal value of a^* is given by the ratio of the empirical covariance to the variance,

$$a^* = \frac{\mathrm{Cov}[f,h]}{\mathrm{Var}[h]}. \tag{14}$$

How should we select the control variate h? To gain some insights into the potential reduction in variance, we can compute the ratio of the two variances by inserting (14) into (13),

$$\frac{\mathrm{Var}(\widehat{f})}{\mathrm{Var}(f)} = 1 - \left(\mathrm{Corr}[f,h]\right)^2, \quad \mathrm{Corr}[f,h] = \frac{\mathrm{Cov}[f,h]}{\sqrt{\mathrm{Var}(f)\mathrm{Var}(\widehat{f})}}. \tag{15}$$

[3] As an exercise, first show equation (8a) under the mean field assumption. Next, choose a structured mean-field family of distributions and compute the Rao-Blackwellized estimator for the gradient.

[4] This section is based on (Paisley et al., 2012; Ranganath et al., 2014). Refer to the original materials for more details.

Thus, the greater the correlation between f and h, the greater the variance reduction. One possible choice is to choose h such that $\widehat{f}$ is a tight lower bound or an upper bound of f. In that case by construction, there will be high correlation between f and h, (Paisley et al., 2012). Approximating functions that do not bound f can be also used but there will not be theoretical guarantees on their effectiveness, for example, the score function, $h(z) = \nabla_\phi \log q_\phi(z)$. The advantage of using the score function is that its expectation is always zero, $\mathbb{E}_q[\nabla_\phi \log q_\phi(z)] = 0$, for all q.

Using our choice for the control variate $h(z)$, the stochastic approximation of the gradient can be computed using

$$\begin{aligned}\widehat{\nabla}_\phi \mathcal{L} &= \mathbb{E}_{q_\phi(z)}[\underbrace{f(z) - a^* h(z)}_{\widehat{f}(z)}] \qquad (16)\\ &\simeq \frac{1}{L}\sum_{l=1}^{L} \nabla_\phi \log q_\phi(z_l)\left(\log p(x, z_l) - \log q_\phi(z_l) - a^*\right), \quad \text{where } z_l \sim q_\phi(z).\end{aligned}$$

The idea of using control variates can be always combined with Rao-Blackwellization to further reduce the variance. Refer to (Ranganath et al., 2014) for more details.

4. Black-box variational inference using pathwise gradient estimator

We saw that the gradient of the expectation can be represented using score functions. We found that using score functions results in estimators with prohibitively large variance. Although variance reduction techniques could be useful, the large variance could still be problematic. Under some additional assumptions, we can represent $\nabla_\phi \mathcal{L}$ using pathwise gradient estimators which generally have *lower variance* compared to the score function representation. The pathwise gradient estimator is also known as the *reparameterization gradient estimator* as it is computed using the so-called *reparameterization trick* (Price, 1958), introduced by (Kingma and Welling, 2014; Rezende et al., 2014) in the context of variational autoencoders (VAEs).

Pathwise gradient estimator using the reparametrization trick. Under some mild assumptions, for an approximate posterior $q_\phi(z)$, we can choose a transformation $t_\phi(\epsilon)$ such that $z = t_\phi(\epsilon)$ is distributed according to $q_\phi(z)$ where ϵ is an auxiliary variable with a known distribution $\epsilon = p(\epsilon)$.

Given our deterministic mapping, $t_\phi(\epsilon)$, and the fact that the differential area (mass of the distribution) is invariant under the change of variables, $q_\phi(z)\mathrm{d}z = p(\epsilon)\mathrm{d}\epsilon$, we can rewrite $\nabla_\phi \mathcal{L}$ as,

$$\begin{aligned}\nabla_\phi \mathcal{L} &= \nabla_\phi \int q_\phi(z) g_\phi(z)\mathrm{d}z\\ &= \nabla_\phi \int p(\epsilon) g_\phi(z)\mathrm{d}\epsilon\\ &= \int p(\epsilon)\nabla_\phi g_\phi(z)\mathrm{d}\epsilon\\ &= \int p(\epsilon)\nabla_z g_\phi(z)\nabla_\phi t_\phi(\epsilon)\mathrm{d}\epsilon\\ &= \mathbb{E}_{p(\epsilon)}[\nabla_z g_\phi(z)\nabla_\phi t_\phi(\epsilon)].\end{aligned}$$

We now can compute noisy unbiased gradients using Monte Carlo samples:

$$\boxed{\nabla_\phi \mathcal{L} \simeq \frac{1}{L}\sum_{l=1}^{L} \nabla_z g_\phi(z_l)\nabla_\phi t_\phi(\epsilon_l), \quad \text{where} \quad z_l = t_\phi(\epsilon_l) \text{ and } \epsilon_l \sim p(\epsilon).} \qquad (17)$$

Our assumptions throughout this procedure are minimal:

(i) The variable z should be a continuous random variable with a known reparameterization, $z = t_\phi(\epsilon)$;

(ii) We can easily generate samples from the base distribution $p(\epsilon)$;

(iii) $g_\phi(z)$ is differentiable with respect to z.

Regarding the assumption (iii), we can use automatic differentiation tools to compute $\nabla_z g_\phi(z)$ effectively. See (Baydin et al., 2015) for a survey on automatic differentiation tools.

An example of the transformation in assumption (i) is provided by the univariate Gaussian case: $z \sim \mathcal{N}(\mu, \sigma^2)$. In this case, a valid reparametrization is $z = \mu + \sigma\epsilon$ where ϵ is an auxiliary noise variable $\epsilon \sim \mathcal{N}(0, 1)$. More generally, if the variational distribution is a multivariate Gaussian distribution with mean μ and covariance Σ, the transformation consists of standardizing the random variable z as: $z = t_{\mu,\Sigma}(\epsilon) = \Sigma^{\frac{1}{2}}\epsilon + \mu$. Kingma and Welling (2014, Section 2.4) discuss a large class of $q_\phi(z)$ for which we can choose such a differentiable transformation $t_\phi(\cdot)$ and auxiliary variable ϵ, namely:

- **Tractable inverse CDF**. In this case a valid transformation $t_\phi(\epsilon)$ is the inverse cumulative density function (CDF) of $q_\phi(z)$ where $\epsilon \sim \text{Unif}(0, I)$. Examples: Exponential, Cauchy, Logistic, Rayleigh, Pareto, Weibull, Reciprocal, Gompertz, Gumbel and Erlang distributions.

- **Location-scale family**. In this case a valid transformation is $t_\phi(\epsilon) = \text{location} + \text{scale} \cdot \epsilon$, where ϵ is chosen as the standard normal distribution with location $= 0$ and scale $= 1$. Examples include: Laplace, Elliptical, Student's-t, Logistic, Uniform, Triangular and Gaussian distributions.

- **Composition**. In this case a valid transformation is obtained by expressing random variables as different transformations of auxiliary variables. Examples are Log-Normal (exponentiation of normally distributed variable), Gamma (a sum over exponentially distributed variables), Dirichlet (weighted sum of Gamma variates), Beta, Chi-Squared, and F distributions.

If we cannot find a differentiable transformation $t_\phi(\cdot)$ that falls into the above three categories, we can use approximations to the inverse CDF (Kingma and Welling, 2014).

Pathwise estimator vs score function estimator. When compared to the score function estimator, (7), the pathwise gradient estimator generally has better-behaved variance (Titsias and Lázaro-Gredilla, 2014; Rezende et al., 2014). Recall the solutions from the score-function estimator and the pathwise-gradient estimator given by (7) and (17). In the case of the pathwise-gradient estimator, $\nabla_z g_\phi(z)$ can be computed effectively by estimating the gradient through a backpropagation algorithm in the context of stochastic computation graphs (Schulman et al., 2015)—the gradient is taken w.t.t. z which is treated as a deterministic node (given our deterministic mapping $t_\phi(\epsilon)$), and that would allow backpropagation. The gradient of the second term $\nabla_\phi t_\phi(\epsilon)$ generally is well-behaved as ϵ has a known distribution. In the case of the score-function gradient estimator, the gradient of the score function, $\nabla_\phi \log q_\phi(z)$, is taken w.r.t. ϕ which does not allow backpropagation in a natural form—this is because ϕ is a random variable thus treated as a random node on the graph while backpropagation can not flow effectively through a random node. Furthermore, Monte Carlo samples are taken from $q_\phi(z)$ as opposed to the known distribution $p(\epsilon)$. Altogether, in the evaluation of (7) there is a greater chance of sampling rare values that contribute to the high variance.

Finally, in comparison to the score-function estimator, the pathwise-graident estimator is less generic since it only covers differentiable models—$g_\phi(z)$ needs to be differentiable w.r.t. z.

Example: An important example of the use of the reparametrization trick and the pathwise gradient estimator is given by the popular models of variational autoencoders (Kingma and Welling, 2014).

Appendix A. Dominated convergence theorem

Let $f(x,t)$ be defined for $t \in \mathbb{R}$. Recall Leibniz's rule for differentiation under the integral sign

$$\nabla_x \int_{a(x)}^{b(x)} f(x,t)\mathrm{d}t = f(x,b(x))\nabla_x b(x) - f(x,a(x))\nabla_x a(x) + \int_{a(x)}^{b(x)} \nabla_x f(x,t)\mathrm{d}t,$$

where $a(x) > -\infty$ and $b(x) < \infty$. If $a(x)$ and $b(x)$ are not a function of x, then

$$\nabla_x \int_a^b f(x,t)\mathrm{d}t = \int_a^b \nabla_x f(x,t)\mathrm{d}t.$$

Assume that:

- for all $x \in I = [a,b]$, $f(x,t)$ is measurable and integrable in x, that is $\int_{\mathbb{R}} f(x,t)\mathrm{d}t$ is well defined.
- $\nabla_x f(x,t)$ exists and is integrable over $t \in \mathbb{R}$ for each $x \in I$.
- There exists an integrable dominating function $\mathfrak{F} : \mathbb{R} \to \mathbb{R}$ where $|\nabla_x f(x,t)| \leq \mathfrak{F}(t), \forall x \in I, t \in \mathbb{R}$. This property means $\nabla_x f(x,t)$ is bounded.

Then based on dominated convergence theorem, $\nabla_x \int_{\mathbb{R}} f(x,t)\mathrm{d}t$ exists and

$$\nabla_x \int_{\mathbb{R}} f(x,t)\mathrm{d}t = \int_{\mathbb{R}} \nabla_x f(x,t)\mathrm{d}t.$$

In our case, the argument inside the integral in (4a) satisfies all the conditions—as the score function exists and bounded—which allows application of the theorem.

References

A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind. Automatic differentiation in machine learning: a survey. *arXiv preprint arXiv:1502.05767*, 2015.

C. M. Bishop. *Pattern recognition and machine learning.* Springer, New York, NY, USA, 2006.

Samuel Gershman and Noah D. Goodman. Amortized inference in probabilistic reasoning. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society, CogSci 2014, Quebec City, Canada, July 23-26, 2014*, 2014.

Peter W. Glynn. Likelihood ratio gradient estimation for stochastic systems. *Commun. ACM*, 33 (10):75–84, October 1990. ISSN 0001-0782.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *ICLR*, April 2014.

John William Paisley, David M. Blei, and Michael I. Jordan. Variational bayesian inference with stochastic search. In *ICML*, 2012.

R. Price. A useful theorem for nonlinear devices having Gaussian inputs. *IRE Trans. on Information Theory*, IT-4:69–72, June 1958.

Rajesh Ranganath, Sean Gerrish, and David M. Blei. Black box variational inference. In *AISTATS*, pages 814–822, 2014.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1278–1286, 2014.

John Schulman, Nicolas Heess, Theophane Weber, and Pieter Abbeel. Gradient estimation using stochastic computation graphs. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 3528–3536, 2015.

Michalis K. Titsias and Miguel Lázaro-Gredilla. Doubly stochastic variational Bayes for non-conjugate inference. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1971–1979, 2014.