Advanced Variational Inference

Caterina De Bacco and Isabel Valera

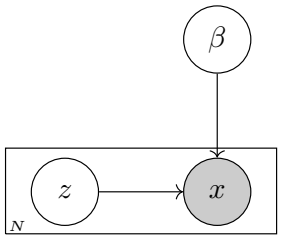

Figure 1: Graphical model.

1 Gradient-ascent for Variational Inference

Following [1], we consider the fairly simple —but general— graphical model depicted in Figure 1. Thus, the joint distribution over the observed variables (the data) $X = \{x_n\}_{n=1}^N$, the set of local latent variables $Z = \{z_n\}_{n=1}^N$, and the set of global latent variables β, can be written as

$$p(X, Z, \beta) = p(\beta) \prod_{n=1}^{N} p(x_n, z_n|\beta). \tag{1}$$

Here, the distinction between local and global hidden variables is determined by the conditional dependencies. In particular, the n-th observation x_n and the n-th local variable z_n are conditionally independent, given global variables β, of all other observations and local hidden variables, i.e.,

$$p(x_n, z_n|X_{-n}, Z_{-n}, \beta) = p(x_n, z_n|\beta).$$

In order to approximate the posterior distribution of the latent variables, $p(Z, \beta|X)$, we rely on Variational Inference using a mean-field variational distribution family of the form:

$$q(Z, \beta) = q_\gamma(\beta) \prod_{n=1}^{N} q_{\phi_n}(z_n),$$

where γ and $\phi = \{\phi_n\}_{n=1}^N$ are the global and local variational parameters, respectively. More in detail, VI aims to find the set of variational parameters γ^* and ϕ^* that maximize the evidence lower bound (ELBO),[1] which is given by

$$\begin{aligned}
\mathcal{L}(x, \gamma, \phi) &= \mathbb{E}_{q_{\gamma,\phi}(Z,\beta)}[\log p(X, Z, \beta)] - \mathbb{E}_{q_{\gamma,\phi}(Z,\beta)}\left[\log q_{\gamma,\phi}(Z, \beta)\right] && (2)\\
&= \mathbb{E}_{q_{\gamma,\phi}(Z,\beta)}[\log p(X|Z, \beta)] - \mathrm{KL}(q_\phi(Z) \| \, p(Z)) - \mathrm{KL}(q_\gamma(\beta) \| \, p(\beta)) && (3)\\
&= \sum_n \Big(\mathbb{E}_{q_{\gamma,\phi_n}(z_n,\beta)}[\log p(x_n|z_n, \beta)] - \mathrm{KL}(q_{\phi_n}(z_n) \| \, p(z_n)) \Big) - \mathrm{KL}(q_\gamma(\beta) \| \, p(\beta)) && (4)
\end{aligned}$$

[1] Or equivalently, that minimise the Kullback-Leibler divergence from $q_{\gamma,\phi}(Z, \beta)$ to $p(Z, \beta|X)$ [4].

In order to find the variational parameters γ^* and ϕ^* that maximize the evidence lower bound (ELBO), we may make use of the gradient ascent algorithm. At each iteration of the gradient ascent algorithm, the variational global parameters are updated as

$$\gamma^t = \gamma^{t-1} - \alpha \nabla_\gamma \mathscr{L}(x, \gamma, \phi),$$

and the variational local parameters as

$$\phi_n^t = \phi_n^{t-1} - \alpha \nabla_{\phi_n} \mathscr{L}(x, \gamma, \phi),$$

where $\alpha \in \Re^+$ is the learning rate.

Obs1: Note that in gradient ascent VI, similarly as in CAVI (see a previous lecture), computing the ELBO amounts to analytically solving the expectation over q, i.e., $\mathbb{E}_{q_{\gamma,\phi_n}(z_n,\beta)}[\log p(x_n|z_n,\beta)]$. In this context, it is common to assume that all the conditional distributions in the model (i.e., $p(x_n|z_n,\beta)$, $p(z_n|x_n,\beta)$ and $p(\beta|x_n,z_n)$) are in the exponential family [1] and, thus, are of the form:

$$p(x|\eta) = h(x)\exp\left[\eta^\top T(x) - A(\eta)\right],$$

where x is the variable, the vector function η corresponds to *natural parameters*, the vector function $T(x)$ corresponds to the *sufficient statistic*, and the scalar functions $h(x)$ and $A(\eta)$ are respectively the base measure and the log-normalizer . The exponential family includes many of the most common distributions. Among many others, it includes distributions such as the normal, exponential gamma, Dirichlet, Bernoulli, categorical and Poisson distributions. As an example, let us assume a random variable x, which is distributed according to the normal distribution with mean μ and standard deviation σ, i.e., $x \sim \mathcal{N}(x; \mu, \sigma^2)$. In such case, the natural parameters are given by $\eta = (\eta_1, \eta_2)^\top$, with $\eta_1 = \mu/\sigma^2$ and $\eta_2 = 1/(2\sigma^2)$; the sufficient statistics by $T(x) = (x, x^2)^\top$; $h(x) = 1/\sqrt{2\pi}$; and $A(\eta) = \mu^2/(2\sigma^2) + \log(\sigma)$. For more details on the exponential family in VI, please refer to [1].

Obs2: Every iteration of the gradient ascent algorithm scales with the number of observations N, and therefore it is very expensive for large datasets. In order o overcome this limitation, the authors in [1] introduced *Stochastic Variational Inference* (SVI).

1.1 Stochastic VI (SVI)

Stochastic Variational Inference (SVI) solve the VI optimization problem using stochastic gradient ascent (SGA). Specifically, in every iteration, one randomly selects a mini-batch os size S to obtain a stochastic estimate of the ELBLO:

$$\hat{\mathscr{L}}(x, \gamma, \phi) = \frac{N}{S} \sum_s \left(\mathbb{E}_{q_{\gamma,\phi_{i_s}}(z_{i_s},\beta)}\left[\log p(x_{i_s}|z_{i_s},\beta)\right] - \mathrm{KL}(q_{\phi_{i_s}}(z_{i_s}) \| \, p(z_{i_s})) \right) - \mathrm{KL}(q_\gamma(\beta) \| \, p(\beta)), \quad (5)$$

where i_s is the variable index from the mini-batch. Then, taking the gradient of (5) yields to a noisy estimator of the direction of the steepest ascent to the true ELBO.

SVI requires the same conditions for convergences as regular SGA. First, the mini-batch indices i_s must be sampled uniformly at random. The size S of the mini-batch should be $1 \leq S < N$. Note that SVI reduces to standard VI, when $S = N$. While larger values of S reduce the variance of the stochastic estimate of the gradient, the computational saving are only obtained when $S \ll N$. The optimal choice of S emerges from a trade-off between the computational overhead associated with processing a mini-batch, such as performing inference over global parameters (favoring larger mini-batches which have lower gradient noise, allowing larger learning rates), and the cost of iterating over local parameters in the mini-batch (favoring small mini-batches). Additionally, this tradeoff is also affected by memory structures in modern hardware such as GPUs.

Second, the learning rate α^t needs to decrease with iterations t, satisfying the Robbins-Monro conditions, i.e., $\sum_{t=1}^\infty \alpha^t = \infty$ and $\sum_{t=1}^\infty (\alpha^t)^2 < \infty$. This guarantees that every point in the parameter space can be reached, while the gradient noise decreases quickly enough to ensure convergence [3].

Importantly, the convergence speed of SGA, forming the basis of SVI, depends on the variance of the gradient estimates. Smaller gradient noise allows for larger learning rates and leads to faster convergence. As a result, we can easily fin several works in the literature proposing different methods to select the mini-batch size and reduce the variance in order for faster convergence. Refer to [4] for an overview.

2 Black Box Variational Inference (BBVI)

Next, we focus on techniques which aim at making VI more generic. This includes making VI applicable to a broader class of models, and also to make VI more automatic, eliminating the need for model specific calculations. This makes VI more accessible and easier to use.

In classical VI, the ELBO is first derived analytically, and then optimized. This procedure is usually restricted to models in the conditionally conjugate exponential family [1]. For many models, including Bayesian deep learning architectures or complex hierarchical models, the ELBO contains intractable expectations with no known or simple analytical solution. Even if an analytic solution is available, the analytical derivation of the ELBO often requires time and mathematical expertise. In contrast, BBVI proposes a generic inference algorithm for which only the generative process of the data has to be specified. The main idea is to represent the gradient as an expectation, and to use Monte Carlo techniques to estimate this expectation.

As discussed in the previous section, in general VI aims at maximizing the ELBO, which is equivalent to minimizing the KL divergence between the variational posterior and target distribution. To maximize the ELBO, one needs to follow the gradient or stochastic gradient of the variational parameters. The key insight of BBVI is that one can obtain an unbiased gradient estimator by sampling from the variational distribution without having to compute the ELBO analytically [2].

2.1 REINFORCE gradients and score function

As shown in previous section, for a broad class of models, the gradients of the ELBO can be expressed as an expectation with respect to the variational distribution:

$$\nabla_{\gamma,\phi}\mathcal{L}(x,\gamma,\phi) = \nabla_{\gamma,\phi}\left(\mathbb{E}_{q_{\gamma,\phi}(Z,\beta)}\left[\log p(X,Z,\beta) - \log q_{\gamma,\phi}(Z,\beta)\right)\right]. \tag{6}$$

Thus, the question that arise is how to compute the gradient of the ELBO with respect to the variational parameters γ, ϕ , since they appear in the expectation. Extending the formulation of the expectation and using the equality $\nabla_{\gamma,\phi}\log q_{\gamma,\phi}(Z,\beta) = \frac{\nabla_{\gamma,\phi} q_{\gamma,\phi}(Z,\beta)}{q_{\gamma,\phi}(Z,\beta)}$, we can write:

$$\begin{aligned}
&\nabla_{\gamma,\phi}\mathcal{L}(x,\gamma,\phi) = \nabla_{\gamma,\phi}\left(\mathbb{E}_{q_{\gamma,\phi}(Z,\beta)}\left[\log p(X,Z,\beta) - \log q_{\gamma,\phi}(Z,\beta)\right)\right] \\
&= \nabla_{\gamma,\phi}\int q_{\gamma,\phi}(Z,\beta)(\log p(X,Z,\beta) - \log q_{\gamma,\phi}(Z,\beta))dZd\beta \\
&= \int \left(\nabla_{\gamma,\phi} q_{\gamma,\phi}(Z,\beta)\right)(\log p(X,Z,\beta) - \log q_{\gamma,\phi}(Z,\beta))dZd\beta \\
&\quad + \int q_{\gamma,\phi}\nabla_{\gamma,\phi}(Z,\beta)(\log p(X,Z,\beta) - \log q_{\gamma,\phi}(Z,\beta))dZd\beta \\
&= \int q_{\gamma,\phi}\left(\nabla_{\gamma,\phi}\log q_{\gamma,\phi}(Z,\beta)\right)(\log p(X,Z,\beta) - \log q_{\gamma,\phi}(Z,\beta))dZd\beta \\
&\quad + \int q_{\gamma,\phi}\nabla_{\gamma,\phi}(\log p(X,Z,\beta) - \log q_{\gamma,\phi}(Z,\beta))dZd\beta \\
&= \int q_{\gamma,\phi}\left(\nabla_{\gamma,\phi}\log q_{\gamma,\phi}(Z,\beta)\right)(\log p(X,Z,\beta) - \log q_{\gamma,\phi}(Z,\beta))dZd\beta + \nabla_{\gamma,\phi}\int q_{\gamma,\phi}(Z,\beta)dZd\beta \\
&= \mathbb{E}_{q_{\gamma,\phi}(Z,\beta)}\left[\left(\nabla_{\gamma,\phi}\log q_{\gamma,\phi}(Z,\beta)\right)(\log p(X,Z,\beta) - \log q_{\gamma,\phi}(Z,\beta))\right].
\end{aligned} \tag{7}$$

As a result, we can obtain a Monte-Carlo unbiased estimate of the gradient of the ELBO with respect to the variational parameters as long as:

- we are able to compute the gradient of the logarithm of the variational distribution, i.e., $\nabla_{\gamma,\phi} \log q_{\gamma,\phi}(Z,\beta)$, often called the *score function*;
- sample from the variational distribution, i.e., $Z,\beta \sim q_{\gamma,\phi}(Z,\beta)$.

In other words, the full gradient $\nabla_{\gamma,\phi}\mathcal{L}(x,\gamma,\phi)$ can now be approximated by a stochastic gradient estimation $\nabla_{\gamma,\phi}\hat{\mathcal{L}}(x,\gamma,\phi)$ by sampling from q as:

$$\nabla_{\gamma,\phi}\hat{\mathcal{L}}(x,\gamma,\phi) = \frac{1}{K}\sum_k \big(\nabla_{\gamma,\phi}\log q_{\gamma,\phi}(Z^{(k)},\beta^{(k)})\big)(\log p(X,Z^{(k)},\beta^{(k)}) - \log q_{\gamma,\phi}(Z^{(k)},\beta^{(k)})), \quad (8)$$

where $Z^{(k)},\beta^{(k)} \sim q_{\gamma,\phi}(Z,\beta)$.

Obs1: note that the score function and sampling algorithms depend only on the variational distribution, not the underlying model: this is why is called "black-box".

Obs2: one can build up a collection of these score functions for various variational approximations and reuse them in a package for a broad class of models.

Obs3: it only requires the practitioner to provide the joint distribution $p(X,Z^{(k)},\beta^{(k)})$ of observations and latent variables without the need to derive the gradient of the ELBO explicitly.

Obs4: main drawback: a direct implementation of stochastic gradient ascent based on (8) suffers from high variances of the estimated gradients. Much of the success of BBVI can be attributed to variance reduction through Rao-Blackwellization and control variates [2].

2.2 Reparameterization of the gradients

An alternative, to the reinforce gradients introduced before, relies on the *reparameterization trick* to reparameterize the gradients. In this case, the gradients are obtained by representing the variational distribution as a deterministic parametric transformation of a noise distribution.

More in detail, let us assume that each local variable z_n is given by a parameterized, deterministic function of random noise ϵ, such that $z_n = g(\epsilon,\phi_n)$, with $\epsilon \sim p(\epsilon)$. Importantly, the noise distribution $p(\epsilon)$ is considered independent of the variational parameters of $q_{\phi_n}(z_n)$, and therefore $q_{\phi_n}(z_n)$ and $g(\epsilon,\phi_n)$ share the same parameters ϕ_n. Note that a similar trick may also be applied to the global variables, such that $\beta = g(\epsilon,\gamma)$.

As an example, consider a random variable z to be distributed according to $q_{\mu,\sigma} = \mathcal{N}(z;\mu,\sigma^2)$. Then, we can apply the reparameterization trick by assuming $z = \mu + \sigma\epsilon$ and $\epsilon \sim \mathcal{N}(\epsilon;0,1)$.

This allows to compute any expectation over the latent variables Z,β as an expectation over the noise variables ϵ by the theory behind the change of variables in integrals. We can now build a stochastic gradient estimator of the ELBO by pulling the gradient into the expectation, and approximating it by samples from the noise distribution:

$$\nabla_{\gamma,\phi}\hat{\mathcal{L}}(x,\gamma,\phi) = \frac{1}{K}\sum_k \nabla_{\gamma,\phi}(\log p(X,g(\epsilon^{(k)},\gamma,\phi)) - \log q_{\gamma,\phi}(g(\epsilon^{(k)},\gamma,\phi)), \quad (9)$$

where we have used the fact that $Z,\beta = g(\epsilon^{(k)},\gamma,\phi)$ and $\epsilon^{(k)} \sim p(\epsilon)$.

Obs1: empirically, reparameterization gradients are often found to have lower variance than REINFORCE gradients, however, a theoretical analysis can show that this is not guaranteed.

Obs2: the reparameterization trick does not trivially extend to many distributions, in particular to discrete ones. Even if a reparameterization function exists, it may not be differentiable. In order for

the reparameterization trick to apply to discrete distributions, the variational distributions require further approximations. Several works have addressed this problem. For example, the categorical distribution is often approximated by replacing the argmax operation with a softmax operator.

References

[1] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[2] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. *Artificial Intelligence and Statistics*, pages 814–822, 2014.

[3] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[4] Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.