1. Introduction

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

We learnt that the mean-field variational inference (VI) can be used to transform the complex inference into a high-dimensional optimization problem. There are two main shortcomings with the mean-field approximation: scalability to large data samples and strong assumptions on the choice of the variational posterior distribution. The latter one is the topic of our discussion in Lecture Notes, Part III. In this note, we are focused on the former one, scalability to large data. As we shall see, the scalability limitation of the VI inference is due to the nature of the optimization: The VI optimization involves iterating between analyzing every data point and estimating its latent structure $z$. The procedure can be inefficient for large data sets, as it requires a pass through all the data point.

In the following we describe stochastic variational inference introduced by Hoffman et al. (2013), a technique which uses the well-known stochastic optimization (Robbins and Monro, 1951) to address the scalability of the inference. Stochastic VI follows noisy estimates of the gradient of the objective, and when compared to the coordinate-ascent VI, it iterates between subsampling the data and readjusting the hidden structure based on only the subsample set as opposed to the full sample set.

We begin with briefly introducing the exponential family distributions as a general family of conjugate models, and next derive both coordinate-ascent VI and its stochastic variant. The following sections are summarized from Hoffman et al. (2013) and Blei et al. (2017). Readers are strongly encouraged to read the original materials.

2. Exponentially family

Let $\beta$ denote global latent variables with corresponding global parameters $\alpha$, and $z = \{z_i\}_{i=1}^n$ denote the set of local latent variables. Consider a generic model in the form of

$$p(x, z, \beta) = p(\beta) \prod_{i=1}^n p(z_i, x_i \mid \beta), \quad (1a)$$

with the conditional joint density over $(z_i, x_i)$ in the exponential family\footnote{1. Make yourself familiar with the basic properties of the exponential family distributions.}

$$p(z_i, x_i \mid \beta) = h(z_i, x_i) \exp\{\beta^\top t(z_i, x_i) - a(\beta)\}, \quad (1b)$$
where the scalar functions $h(\cdot)$ and $a(\cdot)$ indicate the base measure and the log-normalizer, and the vector function $t(\cdot)$ indicates the sufficient statistics. The prior on the global variables $\beta$ is taken to be the corresponding conjugate prior (Diaconis and Ylvisaker, 1979),

$$p_\alpha(\beta) = h(\beta) \exp\{\alpha^\top t(\beta) - a(\alpha)\},$$  \hspace{1cm} (1c)

where $t(\beta) = (\beta, -a(\beta))$ and $\alpha = (\alpha_1, \alpha_2)$ is the natural parameter of the prior.

**Complete conditional distributions.** For the model described in (1), the complete conditional distributions, $p_\alpha(\beta \mid x, z)$ and $p(z_i \mid x_i, \beta)$, are in the exponential family,

$$p_\alpha(\beta \mid x, z) = h(\beta) \exp\{\eta_g(x, z, \alpha)^\top t(\beta) - a(\eta_g(x, z, \alpha))\},$$  \hspace{1cm} (2a)

and

$$p(z_i \mid x_i, \beta) = p(z_i \mid x_i, \beta, z_{-i}, x_{-i})$$
$$= h(z_i) \exp\{\eta_l(x_i, \beta)^\top t(z_i) - a(\eta_l(x_i, \beta))\},$$  \hspace{1cm} (2b)

where $z_{-i} = \{z_j \mid j \neq i, \forall i, j = \{1, \ldots, n\}\}$ and the vector functions $\eta_l$ and $\eta_g$ indicate the local and global natural parameters, respectively. Using (1) and (2), these parameters can be expressed in explicit forms

$$\eta_g(x, z, \alpha) = \left(\alpha_1 + \sum_{i=1}^n t(z_i, x_i), n + \alpha_2\right),$$  \hspace{1cm} (3a)

$$\eta_l(x_i, \beta) = [\text{HW}].$$  \hspace{1cm} (3b)

These explicit forms will be useful later on in Sec. 2.2 where we derive the stochastic variational inference for the exponential family distributions.

### 2.1 Mean-field variational inference

Here, we discuss the coordinate-ascent mean-field VI for our model and discuss the scalability limitation of the inference to large data which motivates the design of a scalable inference using stochastic variational inference in Sec. 2.2.

**Mean-field variational family.** Consider a fully factorized posterior,

$$q(z, \beta) = q_{\lambda}(\beta) \prod_{i=1}^n q_{\phi_i}(z_i),$$  \hspace{1cm} (4a)

where $\{\phi_i\}_{i=1}^n$ denote the local variational parameters and $\lambda$ denotes the global variational parameter of their corresponding posterior distributions. Following the conjugacy, $q_{\lambda}(\beta)$ and $q_{\phi_i}(z_i)$ are assumed in the exponential family whose densities have the same functional form as their prior,

$$q_{\lambda}(\beta) = h(\beta) \exp\{\lambda^\top t(\beta) - a(\lambda)\},$$  \hspace{1cm} (4b)

$$q_{\phi_i}(z_i) = h(z_i) \exp\{\phi_i^\top t(z_i) - a(\phi_i)\}. \hspace{1cm} (4c)$$

**Variational lower bound.** Given our choice of prior and posterior distributions, the variational lower bound (ELBO) is expressed as,

$$\mathcal{L}(\lambda, \phi) = \mathbb{E}_q[\log p(x, z, \beta) - \log q(\beta, z)].$$  \hspace{1cm} (5)
Input: Joint model \(\log p(x, z, \beta)\), variational family \(q_\phi(z)\), \(q_\lambda(\beta)\).
Output: Optimized variational parameters \(\lambda\) and \(\phi\).
Initialization: Initialize \(\lambda\) randomly.

repeat
  for For each data point \(i\) do
    Update the corresponding local parameter
    \[
    \phi_i \leftarrow \mathbb{E}_{q_\lambda(\beta)}[\eta_l(\beta, x_i)].
    \]
  end

  Update the global parameter as
  \[
  \lambda \leftarrow \mathbb{E}_{q_\phi(z)}[\eta_g(x, z, \alpha)].
  \]
until lower bound is converged;

Algorithm 1: Mean-field variational inference using coordinate ascent updates.

The optimized variational posteriors can be computed by maximizing the ELBO. As we saw previously in (Bishop, 2006, Ch. 10), we can optimize the posteriors using the optimal solutions:

\[
q^*(\beta) \propto \exp \left\{ \mathbb{E}_{q_\phi(z)}[\log p(x, z, \beta)] \right\} \propto \exp \left\{ \mathbb{E}_{q_\phi(z)}[\log p(\beta | x, z)] \right\}
\]

\[
q^*(z_i) \propto \exp \left\{ \mathbb{E}_{q_\lambda(\beta)q_{\phi_i-z_i}}(z_i)[\log p(x, z, \beta)] \right\} \propto \exp \left\{ \mathbb{E}_{q_\lambda(\beta)}[\log p(z_i | x_i, \beta)] \right\}
\]

(6a)

(6b)

Alternatively we can optimize the posteriors by taking directly gradient of the lower bound w.r.t. the variational parameters. This approach was briefly discussed at the last paragraph of Section 10.2.2 of Bishops’ text book.

In order to make a connection with the subsequent discussion on the stochastic VI, we consider optimization using the latter case, taking gradient of the lower bound.

The gradient of lower bound and coordinate ascent inference. Let us start with writing the lower bound (5) by keeping terms depending on the variational parameters \(\lambda\), that is

\[
\mathcal{L}(\lambda) \propto \mathbb{E}_q[\log p(x, z, \beta) - \log q_\lambda(\beta)] \\
\propto \mathbb{E}_q[\log p(\beta | x, z) + \log p(x, z) - \log q_\lambda(\beta)] \\
\propto \mathbb{E}_q[\log p(\beta | x, z)] - \mathbb{E}_q[\log q_\lambda(\beta)] \\
\propto \mathbb{E}_q[\eta_g(x, z, \alpha)]^T \left[ \mathbb{E}_{q(\beta)}[t(\beta)] - \lambda^T \mathbb{E}_{q(\beta)}[t(\beta)] \right] + a(\lambda) \tag{7}
\]

where we have used the exponential family identity that the expectation of the sufficient statistics is equal to the gradient of the log normalizer, \(\mathbb{E}_{q(\beta)}[t(\beta)] = \nabla\lambda a(\lambda)\).

To derive the coordinate ascent updates, we need to take the gradient of \(\mathcal{L}(\lambda)\) w.r.t. \(\lambda\),

\[
\nabla_\lambda \mathcal{L}(\lambda) = \nabla_\lambda^2 a(\lambda) \left( \mathbb{E}_q[\eta_g(x, z, \alpha)] - \lambda \right). \tag{8a}
\]

The re-estimated global variational parameter \(\lambda\) is then obtained by setting \(\nabla_\lambda \mathcal{L}(\lambda)\) to zero,

\[
\lambda = \mathbb{E}_{q_\phi(z)}[\eta_g(x, z, \alpha)]. \tag{8b}
\]
In a similar fashion, for the local variational parameters, \( \phi = \{ \phi_i \}_{i=1}^{n} \), we obtain:

\[
\nabla_{\phi_i} L(\phi_i) = \nabla_{\phi_i}^2 a(\phi_i) \left( \mathbb{E}_q[\eta(x_i, \beta)] - \phi_i \right),
\]

which gives

\[
\phi_i = \mathbb{E}_{q_\lambda(\beta)}[\eta(x_i, \beta)].
\]

The updates (8b) and (9b) show a connection to the expectation-maximization (EM) where (9b) corresponds to the expectation step, and (8b) corresponds to the maximization step. An algorithmic presentation is shown in Algorithm 1 which guarantees to find a local optimum of the lower bound.

**Inference scalability.** The local steps in the algorithm reveal an inefficiency for large data samples, \( n \): The expectation step involves passing through all data points and evaluating \( \phi_i \) for all \( i \). This can be wasteful at times given that for large \( n \) we should be able to learn something about the global variational parameters \( \lambda \) from only a subset of the data.

### 2.2 Stochastic variational inference

The scalability of the inference to large datasets can be tackled using stochastic optimization, which leads to stochastic variational inference, an efficient algorithm that continually improves its estimate of the global parameters as it analyzes more observations (Hoffman et al., 2013). Stochastic optimization follows noisy estimates of the gradient with a decreasing step size. Beside the computational advantages of using noisy estimates of the gradient, they allow algorithms to escape shallow local optima of complex objective functions.

In the following we first discuss natural gradients: How they differ from the classical gradients and how they relate to the coordinate updates in VI. We then discuss the stochastic VI which uses noisy version of the natural gradients of the variational lower bound.

#### 2.2.1 Natural gradient of the lower bound

**Classical gradient.** Consider the lower bound in (7) expressed as a function of \( \lambda \). The classical gradient method for maximization tries to find the maximum of \( L(\lambda) \) by taking steps of size \( \rho \) in the direction of the gradient,

\[
\lambda \leftarrow \lambda + \rho \nabla_{\lambda} L(\lambda).
\]

In the classical form, the gradient points in the direction of the steepest ascent in Euclidean space, that is the gradient points to the same direction as the solution to

\[
\arg\max_{\mathbf{d}\lambda} L(\lambda + \mathbf{d}\lambda), \quad \text{s.t.} \quad ||\mathbf{d}\lambda||^2 < \epsilon^2,
\]

for sufficiently small values of \( \epsilon \). The Euclidean distance is a poor measure of similarity between two distributions \( q_{\lambda'}(\beta) \) and \( q_{\lambda''}(\beta) \) for two settings of \( \lambda' \) and \( \lambda'' \), since it only takes into account their mean estimates. As an example, consider two cases of:

(i) \( q_{\lambda'}(\beta) \sim \mathcal{N}(0, 10^{-4}) \) and \( q_{\lambda''}(\beta) \sim \mathcal{N}(0.1, 10^{-4}) \) \( \Rightarrow \) \( \text{D}_{\text{Euclidean}}[q_{\lambda'}(\beta), q_{\lambda''}(\beta)] = 0.1 \);

(ii) \( q_{\lambda'}(\beta) \sim \mathcal{N}(0, 10^4) \) and \( q_{\lambda''}(\beta) \sim \mathcal{N}(10, 10^4) \) \( \Rightarrow \) \( \text{D}_{\text{Euclidean}}[q_{\lambda'}(\beta), q_{\lambda''}(\beta)] = 10 \).

In (i) two distributions barely overlap while in (ii) the two overlap largely. This however has not been reflected in the measured distance.

We now look at the natural gradient which corrects for this issue by redefining the definition of the gradient based on symmetrized Kullback-Leibler (KL) divergence as a natural measure of dissimilarity.
**Natural gradient.** The natural gradient points in the direction of the steepest ascent in Riemannian space—the space where local distance is defined by symmetrized KL divergence rather than the $L^2$ norm. That is the gradient points to the same direction as the solution to

$$
\arg\max_{\lambda} \mathcal{L}(\lambda + d\lambda), \quad \text{s.t.} \quad D_{KL}^{\text{sym}}[q_\lambda(\beta), q_{\lambda+d\lambda}(\beta)] < \epsilon
$$

where in the general form,

$$
D_{KL}^{\text{sym}}[q_\lambda(\beta), q_\lambda(\beta)] = D_{KL}[q_\lambda(\beta)||q_\lambda(\beta)] + D_{KL}[q_\lambda(\beta)||q_\lambda(\beta)].
$$

The constraint in (12) is handled by defining the Riemannian metric $G(\lambda)$ which defines a linear transformation of $\lambda$ under which the squared Euclidean distance between $\lambda$ and $\lambda + d\lambda$ is the symmetrized KL between $q_\lambda(\beta)$ and $q_{\lambda+d\lambda}(\beta)$, that is,

$$
d\lambda^\top G(\lambda)d\lambda = D_{KL}^{\text{sym}}[q_\lambda(\beta), q_{\lambda+d\lambda}(\beta)].
$$

Amari (1998) showed that the natural gradient can be computed by premultiplying the gradient by the inverse of the Riemannian metric, $G(\lambda)^{-1}$. Thus the natural gradient of the lower bound $\mathcal{L}(\lambda)$ is given by,

$$
\hat{\nabla}_\lambda \mathcal{L}(\lambda) \triangleq G^{-1}(\lambda)\nabla_\lambda \mathcal{L}(\lambda)
$$

where $G(\lambda)$ is the Fisher information matrix of $q_\lambda(\beta)$,

$$
G(\lambda) = \mathbb{E}_{q_\lambda(\beta)} \left[ (\nabla_\lambda \log q_\lambda(\beta)) (\nabla_\lambda \log q_\lambda(\beta))^\top \right],
$$

computed from the second moment of the score function $\nabla_\lambda \log q_\lambda(\beta)$.

For our choice of the variational posterior distribution in (4b), we have

$$
\nabla_\lambda \log q_\lambda(\beta) = t(\beta) - \mathbb{E}_{q_\lambda(\beta)}[t(\beta)],
$$

which implies that $G(\lambda)$ is given by the Hessian of the log normalizer function $a(\lambda)$, that is,

$$
G(\lambda) = \mathbb{E}_{q_\lambda(\beta)} \left[ (\nabla_\lambda \log q_\lambda(\beta)) (\nabla_\lambda \log q_\lambda(\beta))^\top \right]
= \mathbb{E}_{q_\lambda(\beta)} \left[ (t(\beta) - \mathbb{E}_{q_\lambda(\beta)}[t(\beta))] (t(\beta) - \mathbb{E}_{q_\lambda(\beta)}[t(\beta)])^\top \right]
= \nabla_\lambda^2 a(\lambda).
$$

Recall the classical gradient of the lower bound w.r.t. $\lambda$ in (8a). Using (8a) and (17) in (15), the natural gradient of the lower bound w.r.t. $\lambda$ is computed by premultiplying the gradient by $G(\lambda)^{-1} = \nabla_\lambda^2 a(\lambda)$, which gives:

$$
\hat{\nabla}_\lambda \mathcal{L}(\lambda) = \mathbb{E}_{q_\lambda(x,z)}[\eta_\beta(x,z,\alpha)] - \hat{\lambda}.
$$

Similarly, the natural gradient of the lower bound w.r.t. the local variational parameters $\phi$ is given by

$$
\hat{\nabla}_{\phi_1} \mathcal{L}(\phi_1) \triangleq G^{-1}(\phi_1)\nabla_{\phi_1} \mathcal{L}(\phi_1)
$$

where $G(\phi_1)$ is the Fisher information matrix of $q_{\phi_1}(z_1)$. And similarly, for our choice of variational posterior in (4c),

$$
\hat{\nabla}_{\phi_1} \mathcal{L}(\phi_1) = \mathbb{E}_{q_{\phi_1}(\beta)}[\eta_\beta(x_1,\beta)] - \phi_1.
$$

The mean-field variational inference can be now derived using the natural gradients. An algorithmic presentation is shown in Algorithm 2. We can see that for the step size of $\rho = 1$, it becomes identical to performing coordinate ascent updates as in Algorithm 1.
Input: Joint model log $p(x, z, \beta)$, variational family $q_\phi(z), q_\lambda(\beta)$.
Output: Optimized variational parameters $\lambda$ and $\phi$.
Initialization: Initialize $\lambda$ randomly, and set the step size $\rho$.
repeat
for For each data point $i$ do
Update the corresponding local parameter
\[
\phi_i \leftarrow \mathbb{E}_{q_\lambda(\beta)}[\eta_l(\beta, x_i)].
\]
end
Update the global parameter as (using: $\lambda \leftarrow \lambda + \rho \hat{\nabla}_\lambda \mathcal{L}(\lambda)$)
\[
\lambda \leftarrow (1 - \rho)\lambda + \rho \mathbb{E}_{q_\lambda(z)}[\eta_g(x, z, \alpha)].
\]
until lower bound is converged.;
Algorithm 2: Mean-field variational inference using natural gradients.

Classical gradient vs natural gradient for the exponential family. We showed that natural gradients are more appropriate measure of dissimilarity as they take into account the geometry of the parameter space. Moreover, for the exponential family of distributions, they lead to cheaper updates. This is because in the case of classical gradients, we would need to compute the Hessian of the log normalizer functions, $\nabla^2 \alpha(\lambda)$ and $\{\nabla^2 \phi_i(\phi_i)\}_i$, while these terms disappear in the case of natural gradients.

2.2.2 Stochastic variational inference using natural gradients.

Using natural gradients, as shown in Algorithm 2, we still require going through the entire data set and re-computing the local variational parameters for each data point before re-estimating the global variational parameter. Stochastic variational inference uses stochastic optimization to fit the global variational parameter by following noisy but cheap-to-compute natural gradients to reach the optimum of an objective function.

Let us rewrite the natural gradient of the ELBO $\hat{\nabla}_\lambda \mathcal{L}(\lambda)$, in an explicit form using (3a) in (18),
\[
\hat{\nabla}_\lambda \mathcal{L}(\lambda) = \mathbb{E}_q[\eta_g(x, z, \alpha)] - \lambda = \left(\alpha_1 + \sum_{i=1}^n \mathbb{E}_{q_{\phi_i}}[t(z_i, x_i)], n + \alpha_2\right) - \lambda. \tag{20}
\]

We then construct a noisy natural gradient by
\[
j \sim \text{Uniform}(1, \ldots, n)
\frac{\hat{\nabla}_\lambda \mathcal{L}(\lambda)}{\hat{\nabla}_\lambda \mathcal{L}(\lambda)} = \mathbb{E}_{q_{\phi_j}(z_j)}[\eta_g(x_j^{(n)}, z_j^{(n)}, \alpha)] - \lambda = \left(\alpha_1 + n \mathbb{E}_{q_{\phi_j}(z_j)}[t(z_j, x_j)], 1 + \alpha_2\right) - \lambda, \tag{21}
\]
where we have first randomly sampled an index from data and replaced it $n$ times. The noisy gradient is unbiased: $\mathbb{E}_j[\hat{\nabla}_\lambda \mathcal{L}(\lambda)] = \hat{\nabla}_\lambda \mathcal{L}(\lambda)$, and it is cheap to compute—it only involves a single sampled data point and only one set of optimized local parameters.

Finally, we use the noisy natural gradients in a Robbins-Monro algorithm (Robbins and Monro, 1951) to optimize the ELBO. Let us define: $\hat{\lambda} \triangleq \left(\alpha_1 + n \mathbb{E}_{q_{\phi_j}(z_j)}[t(z_j, x_j)], 1 + \alpha_2\right)$. The global variational parameter $\hat{\lambda}$ is then updated by
\[
\lambda_t = (1 - \rho_t)\lambda_{t-1} + \rho_t \hat{\lambda}, \tag{22}
\]
**Input:** Joint model log $p(x, z, \beta)$, variational family $q_\phi(z), q_\lambda(\beta)$.

**Output:** Optimized variational parameters $\lambda$ and $\phi$.

**Initialization:** Initialize $\lambda_0$ randomly, and set the step size $\rho_0 = \rho_0$.

**repeat**

- Sample an index $j$ from data at random, $j \sim \text{Uniform}(\ldots, n)$.
- Optimize the local variational parameter for $x_j$,
  $$ \phi_j = E_{q_\phi(\beta)[\eta_l(\beta,x_j)]} $$

- Compute the intermediate global parameter $\hat{\lambda}$,
  $$ \hat{\lambda} = E_{q_{\phi_j}(z_j)}[\eta_{g}(x_j^{(n)}, z_j^{(n)}, \alpha)] $$

- Optimize the global variational parameter using
  $$ \lambda_t = (1 - \rho_t)\lambda_{t-1} + \rho_t\hat{\lambda} $$

- Update the step size $\rho_t$ appropriately such that it satisfies the Robbins and Monro conditions (23).

**until** stopping criteria are satisfied.

**Algorithm 3:** Stochastic variational inference using natural gradients.

where the step size $\rho_t$ must satisfy Robbins and Monro (1951) requirements,

$$ \sum_t \rho_t = \infty; \quad \sum_t \rho_t^2 < \infty. \quad (23) $$

There are various choices of $\rho_t$ which satisfies these conditions including,

$$ \rho_t = (t + \tau)^{-\kappa}, \quad \kappa \in (0.5, 1], \quad \tau \geq 0, \quad (24) $$

where $\kappa$ is known as the forgetting rate which controls how quickly old information is forgotten and $\tau$ is a delay term which down-weights early iterations. An algorithmic presentation of the stochastic variational inference using natural gradients is given in Algorithm 3.
3. Exercises

[HW1] Using the joint distribution and the complete conditional distributions in (1) and (2), compute the vector of natural parameters for the local parameters, $\eta_l$, and use it in (9) to derive explicit forms of the coordinate ascent updates for the local variational parameters $\phi = \{\phi_i\}_{i=1}^n$.

[HW2] In Algorithm 3, we have considered only a single observation $x_j$ sampled at random from data. This can be easily extended to mini-batches. Make the necessary changes to Algorithm 3 and extend it to the mini-batches by taking a set of samples $\{x_k\}_{k=1}^l$.

References


