Variational Methods for Bayesian Inference

An Overview of Stein Variational Gradient Descent and Mean-Field Variational Inference

Manmeet Bhabra

December 18, 2018

1 Introduction

A central, and possibly the most important, goal in Bayesian inference is the ability to accurately sample from and approximate a posterior distribution. As all inference hinges on the capability to efficiently perform this task it is no surprise that a tremendous amount of research has gone into tackling this issue.

Several approaches and algorithms exist to deal with this problem. One family of algorithms, that has had widespread use over the past decades, are Markov Chain Monte Carlo (MCMC) methods. Although quite robust, some principle gripes of these procedures is that it can be quite slow and computationally expensive, thus becoming a intractable choice for large scale problems. An alternate approach, that has been gaining popularity in the past years, is the family of techniques that fall under the umbrella of variational inference.

Variational inference reformulates the problem of approximating or sampling from the posterior as a deterministic optimization problem. The optimization is in general framed as finding a simpler distribution, within a given family of densities, that best approximates the target, where the KL divergence is used as the metric measure how well the target is approximated. Methods in variational inference can moreover be split into what are termed parametric and non-parametric methods. In this work, we investigate two approaches, with each falling in these two distinct categories: Stein variational methods for inference (a non-parametric approach) and mean-Field variational inference (a parametric approach).

In the following sections, we start by first giving an overview of Stein variational gradient descent and mean-field variational inference. The theory behind each approach is presented and details on their practical implementation is covered. Next, we consider two simple numerical cases to demonstrate both approaches on standard inference problems in order to highlight their differences and draw attention to their advantages and disadvantages.

2 Stein Variational Gradient Descent

In this section we outline the Stein variational gradient descent (SVGD) approach to inference.

Consider a target posterior distribution $\pi(x \mid D)$ that has been conditioned upon on some observed data D. In variational methods, the approach to approximate this given distribution is tackled through an optimization problem. Consider a set of distributions $\mathcal{Q} = \{q(x)\}$. We seek a density belonging to this set such that

$$q^*(x) = \arg\min_{q \in \mathcal{Q}} \big\{ D_{KL}(q \mid\mid \pi) \big\}, \tag{1}$$

where D_{KL} is the Kullback-Leibler divergence given by

$$D_{KL}(q \mid\mid \pi) = \mathbb{E}_q \bigg[\log \frac{q(x)}{\pi(x)} \bigg]. \tag{2}$$

The choice of $\mathcal{Q}$, as expected, plays a pivotal role in this optimization. A space of functions that is too large or complex will result in a more difficult optimization problem, while a space that is too simple will allow for easy optimization but generate an approximation that may not be sufficient in approximating the intractable target π.

The SVGD method provides an interesting approach to tackle this optimization problem. Rather than directly specifying a space of functions $\mathcal{Q}$, a transport map, T, is searched for that transforms an initial tractable reference distribution to the target. Consider a continuous random variable $x \in \mathbb{R}^d$, with density $q(x)$, and some invertible map $\mathcal{T} : \mathbb{R}^d \to \mathbb{R}^d$. Given this mapping, a transformation may be defined of the form $z = \mathcal{T}(x)$ where the density of the transformed random variable z is given by:

$$q_{[T]}(z) = q(\mathcal{T}^{-1}(z)) \cdot |det(\nabla_z \mathcal{T}^{-1}(z))| \tag{3}$$

where $\mathcal{T}^{-1}$ refers to the inverse map of $\mathcal{T}$ and $\nabla_z \mathcal{T}^{-1}(z)$ is the Jacobian matrix of the transformation. The SVGD approach builds off this idea in order to search for a final transformed distribution that well approximates the target with respect to the KL divergence. The final map is moreover determined sequentially through a composition of single maps of the form

$$\mathcal{T} = \mathcal{T}_1 \circ \mathcal{T}_2 \circ \mathcal{T}_3 ... \tag{4}$$

By effectively choosing each single map $\mathcal{T}_i$ such that the KL divergence of the transformed variables keeps decreasing, the composition of the maps allows for the generation of a final distribution that approximates the target with relatively good accuracy.

We begin by first specifying the space of functions that will be central to the SVGD method. Consider a continuous random variable $x \in \mathcal{X} \subset \mathbb{R}^d$. Moreover, say that we have a positive definite kernel $k(x, x') : \mathcal{X} \text{ x } \mathcal{X} \to \mathbb{R}$. The reproducing kernel Hilbert Space (RKHS) is then given as the span

$$\{f \ : \ f(x) = \sum_i^m a_i \cdot k(x, x_i), \ m \in \mathbb{N}, \ a_i \in \mathbb{R}, \ x_i \in \mathcal{X}\}. \tag{5}$$

Furthermore, the vector-valued RKHS of dimension d is given as $\mathcal{H}^d$ and consists of vector valued functions $\boldsymbol{f} = [f_1, ..., f_d]^T$, where $f_i \in \mathcal{H}$.

We now look to see how the sequence of single maps $\mathcal{T}_i$ are generated to progress to the target. Consider that a random variable x_0 with a distribution $q_0(x_0)$ was given initially. Moreover, say that a total of $i-1$ single maps ($\mathcal{T}_1,\ \mathcal{T}_2,\ ..., \mathcal{T}_{i-1}$) have been generated whose composition defines a transformed random variable x_{i-1} (which is obtained by applying the composition of the single maps on x_0) where $x_{i-1} \sim q_{i-1}(x_{i-1})$. We search now for the i^{th} map $\mathcal{T}_i$. To do this, consider a vector function $\boldsymbol{\phi}(x) \in \mathcal{H}^d$ and define a transformation to obtain a mapped random variable x_i of the form

$$x_i = \mathcal{T}_i(x_{i-1}) = x_{i-1} + \epsilon \cdot \boldsymbol{\phi}(x_{i-1}). \tag{6}$$

It can be shown that

$$-\frac{d}{d\epsilon}\Big(D_{KL}(q_{i-1} \mid\mid \pi)\Big)\Big|_{\epsilon=0} = \mathbb{E}_{q_{i-1}}\Big[trace(\mathcal{A}_\pi\ \boldsymbol{\phi}(x))\Big] \tag{7}$$

where A_π is the Stein operator and is given by

$$\mathcal{A}_\pi\ \boldsymbol{\phi}(x) = \nabla_x \log\left(\pi(x)\right) \cdot \boldsymbol{\phi}(x)^T + \nabla_x \boldsymbol{\phi}(x). \tag{8}$$

$\mathcal{T}_i$ can be seen to a mapping that applies a perturbation to the random variables in the "direction" $\boldsymbol{\phi}$. Moreover, equation (7) gives an important result for a mapping of this form as it gives the first variation of the functional (the KL divergence) along the "direction" $\boldsymbol{\phi}(x)$. Intuitively, we hope to find the direction (function $\boldsymbol{\phi}$) such that the negative of the variation is largest in order to see the greatest incremental drop in the KL divergence. It can be shown that, for the given space $\mathcal{H}^d$, a closed form expression for this optimal direction exists and is given as

$$\boldsymbol{\phi}^*_{q_{i-1},\pi}(\cdot) = \mathbb{E}_{q_{i-1}}[\nabla_x\ \log[\pi(x)] \cdot k(x, \cdot) + \nabla_x k(x, \cdot)] \tag{9}$$

Substituting $\boldsymbol{\phi}^*$ from equation (9) into equation (6) for $\boldsymbol{\phi}(x)$ we obtain the optimal transformation $\mathcal{T}_i$ that will generate the largest drop in the KL divergence. This process can then be repeated to obtain the map $\mathcal{T}_{i+1}$, $\mathcal{T}_{i+2}$, ..., until the final transformed random variable has a low KL divergence with respect to the target density. For further details we refer the reader to [1]. Moreover, it is important to note that convergence may be accelerated by using a Newton approach (rather than a steepest descent approach) for the optimization. In this approach, second derivative information may be used to approximate the Hessian which may then be used to determine a Newton direction. For further details regarding this formulation, we refer the reader to [2].

2.1 Practical Implementation

The previous subsection introduced how to generate the composite map to obtain the final distribution that is similar to the target. In practice, this process is completed using a set of particles.

We start, in the SVGD algorithm, by picking some simple reference distribution q_0 that can be easily sampled from (such as a Gaussian). A total of N particles are then independently drawn, $\{x_i\}$, with $x_i \sim q_0$. With the set of particles, we then look to compute the first function $\boldsymbol{\phi}^*_{q_0,\pi}$ to be used for the map $\mathcal{T}_1$. The optimal function, given in equation (9), is approximated with a Monte Carlo method using the particles as the samples. Once computing the optimal function, the map $\mathcal{T}_1$ is then determined and can be used to transform the particles. The resulting distribution of the particles will then follow the distribution $q_1(x_1)$, which will have a lower KL divergence with respect to the target than the previous distribution $q_0(x_0)$. The process is then repeated in order to iteratively move the particles towards a distribution that closely follows the target. This algorithm is summarized in Algorithm 1.

Algorithm 1: Updating the particles from iteration i ($\{x_k^i\}$) to i+1 in the SVGD Algorithm

Result: Distribution of particles at step $i+1$ ($\{x_k^{i+1}\}$)

for *k = 1 to N* **do**

  $\phi^*(x_k^i) = \frac{1}{N}\sum_{l=1}^{N}\left[\nabla_x \log[\pi(x_l^i)]\cdot k(x_l^i, x_k^i) + \nabla_x k(x_l^i, x_k^i)\right]$;

  $x_k^{i+1} = x_k^i + \phi^*(x_k^i)$;

end

2.2 Simple Numerical Case

We present here a simple numerical case to observe how the SVGD methods transports the set of particles to the target distribution. In this case, we consider the target distribution to be a bimodal multivariate Gaussian distribution of the form

$$\pi(x,y) = \frac{1}{3}\cdot\mathcal{N}(x\ ;\ \boldsymbol{\mu}_1, C_1) + \frac{2}{3}\cdot\mathcal{N}(y\ ;\ \boldsymbol{\mu}_2, C_2) \tag{10}$$

where

$$\begin{aligned}\boldsymbol{\mu}_1 &= \begin{bmatrix}3.5\\-1.5\end{bmatrix}, \quad C_1 = \begin{bmatrix}1.0 & -0.5\\-0.5 & 0.7\end{bmatrix},\\ \boldsymbol{\mu}_2 &= \begin{bmatrix}1.25\\1.0\end{bmatrix}, \quad C_1 = \begin{bmatrix}0.8 & 0.25\\0.5 & 0.7\end{bmatrix}\end{aligned} \tag{11}$$

The initial distribution of 400 particles was drawn from a multivariate normal with mean vector $\boldsymbol{\mu} = [0, -6]^T$ and an identity covariance matrix. A total of 2500 gradient descent iterations were completed using a step size of $\epsilon = 0.2$ and an RBF kernel with a bandwidth of 0.1. Figure (1) shows the transport of the particles to the target distribution along with the marginal probability densities. It can be seen that the particles do adequately capture the specified target.

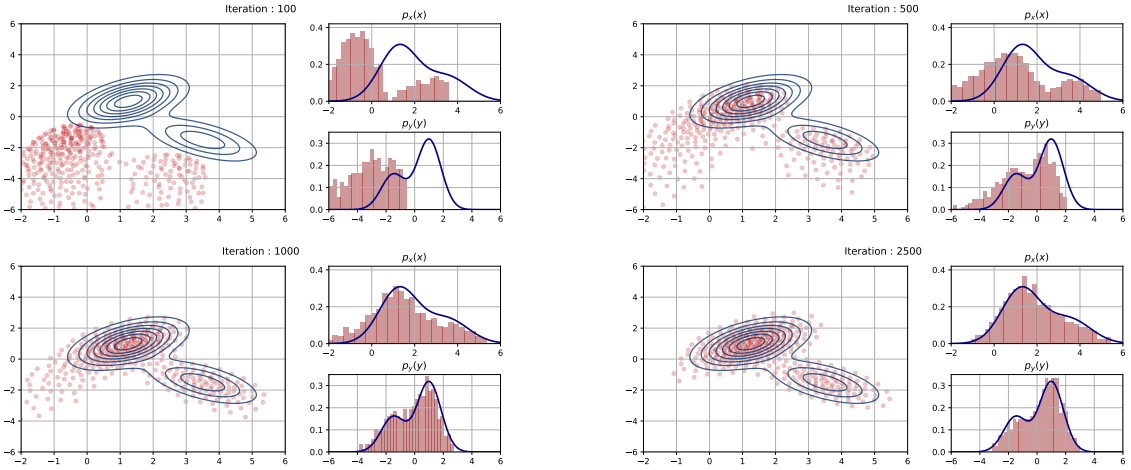


Figure 1: Convergence of the particles to the target distribution (whose contours are shown in blue) for various iterations in the Stein variational descent algorithm. In each subfigure, the marginal distributions are also shown with the target's marginal shown in dark blue.

3 Mean-Field Variational Inference

In this section we outline the fundamentals of mean-field variational inference to approximate intractable distributions.

In mean-field variational inference, we once again look to solving the optimization problem given in equation (1). The process in which we do so, however, is now through specifying a family of distributions and then searching for the parameters of the distributions such that the resulting density has a KL divergence with respect to the target π that is minimized.

In the mean-field approach, first the optimization problem is reformulated. Consider the objective of finding a a function $q(x)$ that best approximates a given posterior $\pi(x|d)$, where d is the evidence. We begin by rearranging the KL divergence between the posterior and the distribution q:

$$\begin{aligned} D_{KL}\Big[q(x) \,||\, \pi(x|d)\Big] &= \mathbb{E}_q\left[\log \frac{q(x)}{\pi(x|d)}\right] \\ &= \mathbb{E}_q[\log\, q(x)] - \mathbb{E}_q[\log\, \pi(x|d)] \\ &= \mathbb{E}_q[\log\, q(x)] - \mathbb{E}_q[\log\, \pi(x,d)] + \mathbb{E}_q[\log\, \pi(x)] \\ &= \mathbb{E}_q[\log\, \pi(x)] - ELBO(q) \end{aligned} \tag{12}$$

where $ELBO$ is the evidence lower bound and is given by

$$ELBO(q) = \mathbb{E}_q[\log\, \pi(x,d)] - \mathbb{E}_q[\log\, q(x)] \tag{13}$$

The optimization problem, of searching for the optimal distribution q^* that minimizes the KL divergence, is now be reformulated as a maximization problem in which the $ELBO$ is to be maximized.

Consider a vector of n latent variables that we hope to infer which we will denote as $\boldsymbol{x}$. In the mean-field approach, we assume a form of the distributions $q(\boldsymbol{x})$ such that the latent variables are mutually independent, with each variable described by its own variational factor:

$$q(\boldsymbol{x}) = \prod_{i=1}^{n} q_i(x_i). \tag{14}$$

The objective is then to sequentially determine the optimal factors q_i such that the $ELBO$ is maximized.

Several approaches exist to determine the optimal variational factors, but one that is more commonly used is coordinate ascent variational inference (CAVI). In this algorithm, we compute each optimal factor individually while keeping the others fixed. Consider the process of updating the i^{th} factor in the mean-field approximation. It can be shown that the optimal factor, with all others fixed, is then given by

$$q_i^*(x) \propto \exp\Big[\mathbb{E}_{\sim i}[\pi(x,d)]\Big] \tag{15}$$

where the expectation is with respect to all variation factors except for q_i. This process is repeated for all other factors, and each optimal variational factor is determined individually. Once one pass through the factors has been completed, the $ELBO$ can then be checked. If the value has not sufficiently increased, the process is repeated once again. In the case where the ELBO has increased to a desirable level, the process is terminated as a good approximation $q(x)$ has been found. It is important to note, however, that the CAVI algorithm brings the $ELBO$ to a local maximum. Pseudocode regarding the structure of this algorithm is given in Algorithm 2.

An aspect that was glossed over in the CAVI process was the details regarding to equation (15). In standard practice, each variational factor is typically taken to be part of the exponential family of distributions. When this is done, the result of the expectation is a closed form expression and whose form follows another distribution in the exponential family. In this way, we may simply recognize the form of the expression and accordingly update the parameters describing the factors in the mean-field approximation. This process will become evident in the numerical cases that will be presented. For further details on mean-field variational inference, we refer the reader to [3].

Algorithm 2: The CAVI method to determine the optimal mean-field approximation $q(x)$

Result: Optimal $q(x)$ mean-field approximation to a distribution $\pi(x|d)$

while *True* **do**
  for k = *1 to* N **do**
    $q_i \propto \exp\big[\mathbb{E}_{\sim i}[\pi(x,d)]\big]$;
  end
  ELBO = computeElbo($q(x)$, $\pi(x,d)$) ;
  if *ELBO converged to desired level* **then**
    return $q(x)$;
end

4 Numerical Cases

4.1 Inferring the Mean and Variance from Gaussian Data

In this first case, we study a simple problem from [4] where, given some data points from an unknown 1D Gaussian distribution, we wish to infer the mean and variance of the distribution.

To tackle the process of inferring the unknown mean (μ) and precision (τ) (which is simply the inverse of the variance), we start by specifying the following priors

$$\begin{aligned} \pi_0(\mu) &= \mathcal{N}(\mu \; ; \; \mu_0, (\kappa_0 \cdot \tau)^{-1}) \\ \pi_0(\tau) &= Ga(\tau \; ; \; \alpha_0, \beta_0) \end{aligned} \tag{16}$$

where Ga corresponds to the Gamma distribution. The likelihood is moreover given as (assuming a total of N observed data points x_i):

$$\pi(\boldsymbol{x} \mid \mu, \tau) = \prod_{i=1}^{N} \mathcal{N}(x_i \; ; \; \mu, \tau^{-1}) \tag{17}$$

In the following subsection, we outline the derivation of the mean-field approximation. Following this, numerical results are presented.

4.1.1 Mean-Field Approximation Derivation

In this section, we derive the closed form updates for the mean-field approximation.

We begin by first obtaining the log of the joint distribution which, using the prior and likelihood, is given by

$$\begin{aligned} \log \pi(\mu, \tau, \boldsymbol{x}) &= \log \pi(\tau) + \log \pi(\mu \mid \tau) + \log \pi(\boldsymbol{x} \mid \mu, \tau) \\ &= (\alpha_0 - 1) \log \tau - \beta_0 \, \tau + \frac{1}{2} \log \tau - \frac{\kappa_0 \, \tau}{2} \cdot (\mu - \mu_0)^2 + \frac{N}{2} \log \tau \\ &\quad - \frac{\tau}{2} \sum_{i=1}^{N} (x_i - \mu)^2 + C \end{aligned} \tag{18}$$

It should be noted that, given equation (18), the SVGD algorithm may be easily implemented using the gradient of the expression with respect to the latent variables μ and τ. Next, we determine the factors of our approximation. Assuming each variable is mutually independent, the following is used as the mean field approximation:

$$q(\mu, \tau) = q_\mu(\mu \; ; \; m_k, s_k^2) \cdot q_\tau(\tau \; ; \; \alpha_k, \beta_k) \tag{19}$$

where the factor q_μ is a Gaussian distribution with mean m_k and variance s_k^2 and q_τ is a Gamma distribution with parameters α_k and β_k.

What remains to be determined now are the variational updates for the CAVI algorithm. That is, we must determine how to update the parameters describing each variational factor such that the final approximation iteratively approaches a good approximation for the target. For each CAVI update, the optimal factors can be found using equation (15):

$$\begin{aligned} q_\mu^* &\propto \exp\Big\{ \mathbb{E}_{q_\tau}\big[\log \pi(\boldsymbol{x}, \mu, \tau)\big] \Big\} \\ q_\tau^* &\propto \exp\Big\{ \mathbb{E}_{q_\mu}\big[\log \pi(\boldsymbol{x}, \mu, \tau)\big] \Big\} \end{aligned} \tag{20}$$

All expectations may be expanded out and computed analytically. Doing so, it is found that the optimal solutions belong to the same variational factor's family. In particular, each optimal density is in the same exponential family as the original density (q_μ^* is Gaussian distributed and $q_\tau^*(\tau)$ is Gamma distributed). By observing the parametric form of each optimal density, the updates to q_μ and q_τ are given as:

$$\begin{aligned} m_k &= \frac{\kappa_0 \mu_0 + \sum_{i=1}^N x_i}{\kappa_0 + N} \\ s_k^2 &= \frac{1}{\alpha_k \cdot (\kappa_0 + N)} \\ \alpha_k &= \alpha_0 + \frac{N+1}{2} \\ \beta_k &= \beta_0 + \frac{\kappa_0}{2} \cdot \Bigg((s_k^2 + m_k^2) - 2\mu_0 m_k + \mu_0 + \\ &\quad \frac{1}{2} \sum_{i=1}^N \Big[x_i^2 - 2 m_k x_i + s_k^2 + m_k^2 \Big] \Bigg) \end{aligned} \tag{21}$$

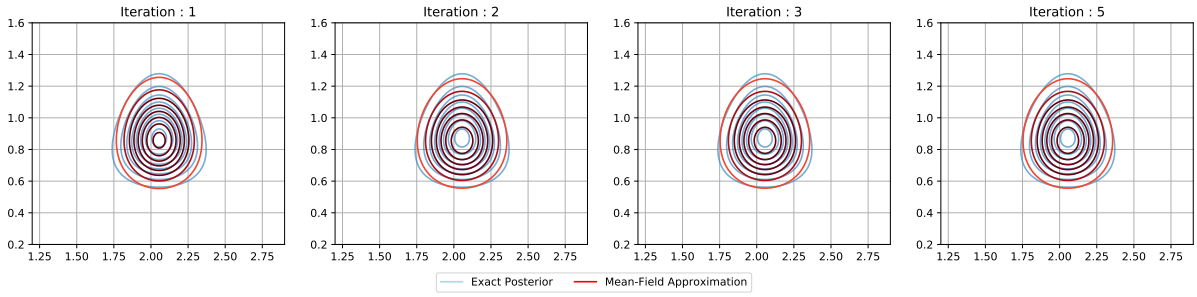
Using the update expressions, the variational factors may be iteratively updated step by step to approach the target distribution following the CAVI algorithm as described in Section 3.

4.1.2 Results

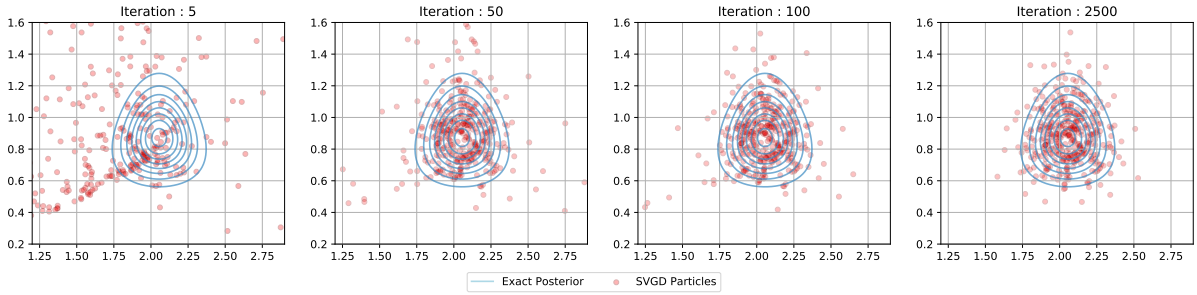
The posterior distribution was approximated using both the mean-field approximation approach and the SVGD method. The true values for the mean and variance were set to $\mu_{truth} = 2.0$ and $\sigma_{truth}^2 = 1.5$ and a total of 50 data points were used to infer the Gaussian's properties.

For the Stein variational approach a total of 400 particles were used for the approximation along with a step size of $\epsilon = 0.05$ in the negative gradient direction. These particles were initially distributed following a multivariate Gaussian distribution of the form $q \sim \mathcal{N}(\boldsymbol{\mu}, I)$, where $\boldsymbol{\mu} = [1, 1]$. An RBF kernel with a fixed bandwidth of 0.1 was used.

Running the test with the given parameters, a total of 2500 iterations were required by the SVGD approach to sufficiently converge the particles towards the target distribution. Furthermore, the mean-field variational inference approach required only 5 iterations. The difference in computational times was also significant. The mean field approach required negligible time (on the order of 0.01 seconds) while the SVGD approach required approximately 211 seconds. This large difference in computational times and the number of design iterations is magnified as a result of the simplicity of this 1D problem. The resulting approximations to the posterior are shown in Figure (2) for both approaches. The mean field approximation can be seen to rapidly converge to a very good approximation of the posterior in very few iterations and the particles in the SVGD approach also converge to the target, however in significantly more iterations. The convergence in the ELBO is also shown in Figure (3).



(a) Mean-Field Approximation Approach



(b) SVGD Approach

Figure 2: Convergence in the posterior distribution approximation using mean-field variational inference (a) and Stein variational gradient descent (b). The final approximation obtained (the approximation after the last design iteration) for each approach is shown in the right-most plots.

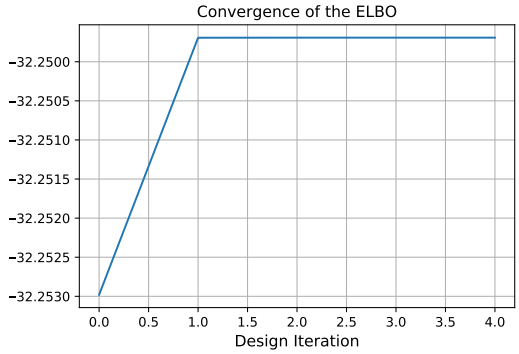


Figure 3: Convergence in the Evidence Lower Bound (ELBO) for the mean-field variational inference approach.

4.2 Bayesian Linear Regression

In this test case we consider a standard inference problem using data with observational error. We consider in particular a forward map given by

$$\mathcal{F}(x_1, x_2) = x_1 + 3 \cdot x_2 \tag{22}$$

We assume further that, for the inference problem, the true values are given by $[x_{true,1}, x_{true,2}]$. The observed data (we consider only a single data point), which consists of some noise, is obtained through the forward map and is given by

$$y = \mathcal{F}(x_{true,1}, x_{true,2}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2). \tag{23}$$

For all cases presented here we have taken the noise as $\sigma = 0.3$ and the true parameters as $x_{true,1} = 1.2$, $x_{true,2} = 2.5$. The inference problem is then to use the given data point y to infer the values of the latent variables x_1 and x_2.

To obtain the posterior distribution, we first instill priors on the parameters x_1 and x_2, which are taken to be standard normal distributions, $\pi_0(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}\ ;\ 0, I)$, where $\boldsymbol{x} = [x_1, x_2]$. The likelihood is moreover given by the forward map and is expressed as $\pi(y \mid \boldsymbol{x}) = \mathcal{N}(y\ ;\ \mathcal{F}(\boldsymbol{x}), \sigma^2)$. Finally, the (unnormalized) posterior is simply given as the product of the prior and likelihood: $\pi(\boldsymbol{x} \mid y) \propto \pi(y \mid \boldsymbol{x}) \cdot \pi_0(\boldsymbol{x})$.

4.2.1 Mean-Field Approximation Derivation

To approximate the posterior using the mean-field approximation we begin by first obtaining the log of the joint distribution $\pi(\boldsymbol{x}, y)$:

$$\log \pi(x_1, x_2, y) = -\frac{x_1^2}{2} - \frac{x_2^2}{2} - \frac{(y - x_1 - 3x_2)^2}{2\sigma^2} + C \tag{24}$$

Next, we need a mean field variational family. Each latent variable (x_1 and x_2) is assumed to be mutually independent and their distribution is governed by their own variational factor. Assuming this, the mean-field approximation is assumed to take the form

$$q(x_1, x_2) = q_1(x_1 \; ; \; m_1, s_1^2) \cdot q_2(x_2 \; ; \; m_2, s_2^2) \tag{25}$$

where the variational factors $q_1,\ q_2$ are Gaussian distributions with means $m_1,\ m_2$ and variance $s_1^2,\ s_2^2$, respectively. The optimal parameters of each variational factor can then be calculated using equation (15):

$$\begin{aligned} q_1^*(x_1) &\propto \exp\Big\{ \mathbb{E}_{q_2}\big[\pi(x_1, x_2, y)\big] \Big\} \\ q_2^*(x_2) &\propto \exp\Big\{ \mathbb{E}_{q_1}\big[\pi(x_1, x_2, y)\big] \Big\} \end{aligned} \tag{26}$$

Analytically computing each expectation, it can be shown that the coordinate optimal variational densities in equation (26) are also Gaussians whose means and variances can be used to update the parameters parametrizing each variational factor:

$$\begin{aligned} m_1 &= \frac{\frac{y}{\sigma^2} - \frac{3m_2}{\sigma^2}}{1 + \frac{1}{\sigma^2}} \quad s_1^2 = \frac{1}{1 + \frac{1}{\sigma^2}} \\ m_2 &= \frac{\frac{3y}{\sigma^2} - \frac{3m_1}{\sigma^2}}{1 + \frac{9}{\sigma^2}} \quad s_2^2 = \frac{1}{1 + \frac{9}{\sigma^2}} \end{aligned} \tag{27}$$

With the parameter updates derived, coordinate ascent variational inference (CAVI) may then be used to progressively obtain the optimal approximation.

4.2.2 Results

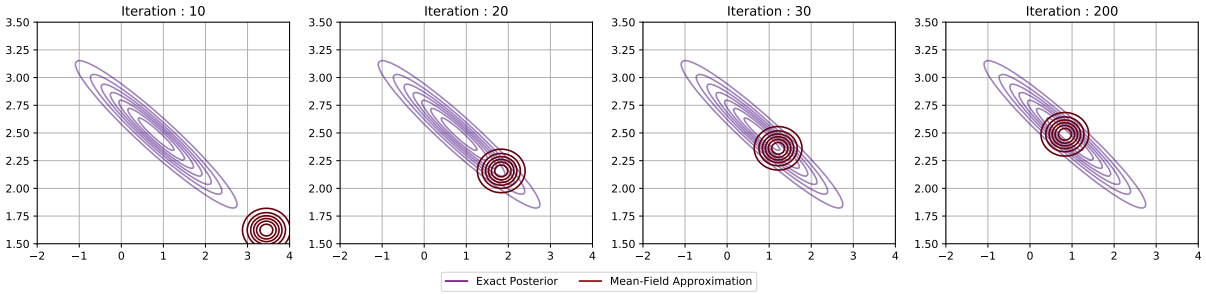
The inverse problem was solved using both mean-field variational inference and the SVGD algorithm. The mean-field solution was obtained iteratively using CAVI, with a total of 200 iteration steps. Furthermore, the SVGD solution was generated using a total of 400 particles initially distributed according to a standard multivariate normal distribution. A total of 1000 steps were taken to transport the particles in this approach. An RBF kernel with a fixed bandwidth of 0.1 was used along with a fixed gradient descent stepsize of $\epsilon = 0.05$.

The completion of both algorithms took approximately 0.01 seconds and 71.86 seconds for the mean-field inference and SVGD approach, respectively. The drastically more efficient mean-field algorithm however was not able to generate as accurate an approximation for the posterior. Figure (4) depicts the convergence in the posterior distribution approximation using both methods. It can be seen that, although an optimal solution is reached quickly for the mean-field case, it is not

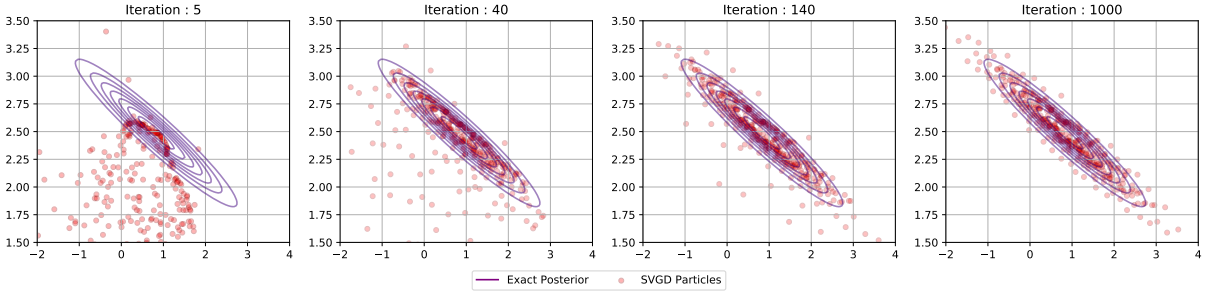
able to accurately capture the posterior. The Stein variational method, on the contrary, is able to transport particles over the whole density. The stark difference in both approximations for this case gives a clear example of one of the downfalls of the mean-field approach; by inherently assuming that all parameters are mutually independent, it becomes impossible to predict any correlation in the variables. This is evident in Figure (4) (a), as although the mean-field approximation is centered about the posterior mean, it produces a result that is uncorrelated for a posterior that clearly has some correlation. Thus, although very quick in computation, the mean-field approach may suffer significant issues in accuracy.

Moreover, the form of the optimal mean-field approximation is a result of the use of the reverse KL divergence ($\mathbb{E}_q[q \mid\mid \pi]$) rather than the forward ($\mathbb{E}_\pi[\pi \mid\mid q]$). The reverse KL divergence penalizes making q high where π is low and it doesn't have any requirement on p in regions where q is low (since we are integrating with respect to $q(x)$). As a result, the optimal mean-field approximation attempts to only be non-zero where p is non-zero, resulting the given approximation in Figure (4).

Finally, as a final note, the evidence lower bound's evolution throughout the optimization problem is shown in Figure (5), where it can be seen to steadily increase after each step in the CAVI algorithm.



(a) Mean-Field Approximation Approach



(b) SVGD Approach

Figure 4: Convergence in the posterior distribution approximation using mean-field variational inference (a) and Stein variational gradient descent (b). The final approximation obtained (the approximation after the last design iteration) for each approach is shown in the right-most plots.

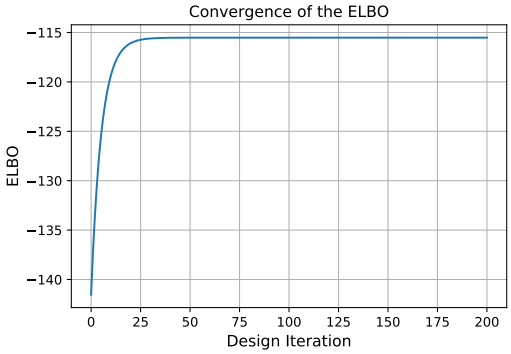


Figure 5: Convergence in the Evidence Lower Bound (ELBO) for the mean-field variational inference approach.

5 Discussion

In this work, we have investigated the use of variational approaches to inference. In particular, the Stein variational gradient descent method, a non-parametric variational approach, and the mean-field approximation, a parametric approach to inference, have been reviewed. Through the numerical cases investigated, it can be seen that both procedures possess certain advantages and disadvantages, and the choice for which to use is ultimately problem specific and dependent on the desires of the user.

Mean-field approximation was firstly found to be orders of magnitude faster than SVGD. The approach was able to effectively converge in the optimization in relatively few iterations to an optimal approximation. Although the cases studied were relatively small in nature (so both methods were relatively fast), the difference of the parametric mean-field approach in terms of speed would be non-negligible when considering problems involving large data sets or of high dimensions. In these cases, where Stein methods or MCMC prove to be too slow, mean-field variational approximation will have a significant advantage.

At the cost of high speed, however, comes issues in accuracy. Although slower than the mean-field method, the SVGD approach is able to consistently transport the particles to the target distribution to provide a sufficiently accurate approximation. This versatility however is not present using mean-field inference. Due to the central assumption that all parameters are mutually independent, the mean-field approximation has difficulty in resolving areas of high correlation in the posterior. What results is an approximation that is centered relatively well around the mean of the posterior, however that poorly approximates all other regions of the target density (as was seen in the second numerical case).

An additional complexity encountered with mean-field inference is in deriving and formulating the equations used for the inference procedure. For relatively simple problems, as investigated in this work, deriving these expressions requires additional effort that traditionally would not need to be expended in other approaches (such as MCMC or SVGD). Moreover, restrictions must also be made on the problems solved such that the conditionals of the target are in the exponential family (this ensures that the optimal variational factors are also in the exponential family so that the optimization process is simplified). This has been the case in call problems investigated in this work, however, for more complicated problems where this property is not satisfied, mean-field

variational inference proves to be a difficult task.

Furthermore, in the case of SVGD, one principle issue encountered is that the gradient of the log of the target is required. In simple problems these expressions may be analytically derived and computed in closed form, thus posing no issues. However, for more complicated problems, the gradient may not be readily available. For instance, in the case of inverse problems where a specific parameter is trying to be inferred in a complicated model (such as the case of inferring the log permeability for the case of the stochastic elliptic PDE), the gradient of the posterior requires a gradient of the likelihood which often times is not straightforward to compute. In these cases, adjoint approaches may offer an efficient alternative to compute these sensitivities.

6 Future Work

There are several avenues for investigation following this work. Firstly, it would be interesting to investigate the performance of both parametric and non-parametric approaches to larger data sets. By studying these inference problems of higher complexity, the efficacy of both approaches in tackling real world problems may be investigated. Moreover, the interface of MCMC and variational methods has started to be studied in the past years. Building on this, it would be insightful to also investigate the marriage of non-parametric and parametric variational approaches (such as SVGD and mean-field inference) to see how these algorithms may be combined to produce more robust approaches to inference.

References

[1] Q. Liu and D. Wang, "Stein variational gradient descent: A general purpose bayesian inference algorithm," in *Advances In Neural Information Processing Systems*, pp. 2378–2386, 2016.

[2] G. Detommaso, T. Cui, Y. Marzouk, A. Spantini, and R. Scheichl, "A stein variational newton method," in *Advances in Neural Information Processing Systems*, pp. 9186–9196, 2018.

[3] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[4] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.