Stein's method in machine learning

Vittorio Romaniello

December 10, 2019

1 Background

Stein's method, introduced by Stein [Ste72], is a popular technique used in probability theory to prove approximation and limit theorems. Applications span several fields, from network analysis [FM06] to sequence analysis in genetics [RSW00] and the study of epidemic models [BB90], for more examples of fields of application see [Rei11]. Despite the large variety of fields of application, Stein's method has been, until recently, solely used in theoretical statistics. However, the work of [OGC17; Oat+19; GM15; LLJ16] related ideas from Stein's method to problems in computational statistics and machine learning, motivating a large body of literature in the area. Examples of applications of Stein's method in machine learning include goodness of fit tests [LLJ16; CSG16; YRN19; Kan+19], variational inference [LW16; Zhu+17; WZL17; HL18] and Markov chain Monte Carlo [SG18; Che+19].

This paper presents the work of Liu and Wang [LW16] on Stein Variational Gradient Descent (SVGD). First, we introduce Stein's method and concepts necessary to understand SVGD. Next, in Section 2, we detail SVGD and discuss future research directions.

1.1 Stein's method

Ross et al. [Ros+11] describe Stein's method as based on two components: the first, a framework to convert the problem of bounding the error in the approximation of one distribution of interest by another into a problem of bounding the expectation of a certain functional of the random variable of interest. The second component of Stein's method is a collection of techniques to bound the expectation appearing in the first component [Ros+11].

More formally, and here we borrow some of the notation from [Bar+19], let $(\Omega, \mathcal{A})$ be a measurable space and denote by $\mathcal{P}_\Omega$ the set of probability measures on $(\Omega, \mathcal{A})$. Further let $\mathcal{P}_E \subset \mathcal{P}_\Omega$ be the set of probability measures on the measurable space $(E, \mathcal{E})$ with $E \subset \Omega$ and $\mathcal{E} \subset \mathcal{A}$. Define $D : \mathcal{P}_E \times \mathcal{P}_E \to \mathbb{R}_+$ as

$$D_{\mathcal{E}}(\mathbb{P}, \mathbb{Q}) = \sup_{g \in \mathcal{E}^{g_n}} \left| \int_E g(x) d\mathbb{P} - \int_E g(x) d\mathbb{Q} \right|, \ x \in E \tag{1}$$

a measurable function that quantifies the discrepancy between two probability measures $\mathbb{Q}, \mathbb{P} \in \mathcal{P}_E$, with $D_{\mathcal{E}}(\mathbb{P}, \mathbb{Q}) = 0 \iff \mathbb{Q} = \mathbb{P}$, where g is a measurable function in $\mathcal{E}^{g_n}$, the set of $\mathcal{E}$-measurable functions. Furthermore, let $\Gamma(\mathcal{Y}) \equiv \{g : E \to \mathcal{Y}\}$, a map $\mathcal{S}_{\mathbb{P}} : \mathcal{G} \subset \Gamma(E) \to \Gamma(\mathbb{R})$ is a Stein operator over a Stein class $\mathcal{G}$ if $\int_E \mathcal{S}_{\mathbb{P}}[g] d\mathbb{P} = 0 \ \forall g \in \mathcal{G}$ for any $\mathbb{P}$. Using the definition in Equation (1), the Stein discrepancy (SD) can be defined as

$$SD_{\mathcal{S}_{\mathbb{P}}[\mathcal{G}]}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{S}_{\mathbb{P}}[\mathcal{G}]} \left| \int_E f(x) d\mathbb{P} - \int_E f(x) d\mathbb{Q} \right| = \sup_{g \in \mathcal{G}} \left| \int_E \mathcal{S}_{\mathbb{P}}[g] d\mathbb{Q} \right|, \ x \in E \tag{2}$$

To simplify notation in the rest of the paper, we write $\int_E f(x) d\mathbb{P} = E_{\mathbb{P}}[f(x)]$ where $E[\cdot]$ is the expectation operator and the subscript is used to specify the measure with respect to which we compute the expectation. The first component of Stein's method transforms the problem of bounding Equation (1) to the problem of bounding Equation (2). The second component of Stein's method corresponds to a set of techniques to bound Equation (2). Depending on the measurable space chosen, the form of the Stein operator and the techniques needed to bound SD differ. The discussion of such techniques is not necessary for an understanding of the rest of the paper, hence, we refer the reader to [Ros+11] for some examples.

Given the definitions above, there are three remarks necessary for clarification:

Remark 1: There are several ways of formulating Stein's method, for instance using an exchangeable pair of random variables [Ste86]. However, here we choose the formulation based on Stein discrepancy and the Stein operator because it relates directly to the topic of this paper.

Remark 2: In Equation (1) we defined the discrepancy metric in general terms, without specifying the measurable space nor the probability measures. Depending on the measurable space, different discrepancies can be constructed (e.g. Kolmogorov, Wasserstein, total variation [Ros+11]), explaining the broad spectrum of applications of Stein's method. While Stein's method can be used under the class of discrepancies defined in Equation (1), known in the literature as integral probability metrics (IPM), attempts have been made to relate IPMs to other classes of divergences commonly used in the statistics literature, which would enable further uses of Stein's method. Such attempts did not, however, prove successful as an equivalence between metrics is difficult to find unless under very strong assumptions are. An example relating the class of ϕ-divergences (e.g. the Kullback-Liebler (KL) divergence) to IPMs can be found in [Sri+09].

Remark 3: If the probability measure $\mathbb{P}$ has a $\mathcal{C}^1$ density p, with respect to the Lebesgue measure, then we can consider the Stein operator of the form

$$\mathcal{T}_p[g] = \langle \nabla \log p, g \rangle + \langle \nabla, g \rangle \tag{3}$$

where $\mathcal{C}^1$ denotes the space of 1-time continuously differentiable functions, $\langle \cdot, \cdot \rangle$ denotes the inner product operation and ∇ denotes the differential operator. Note that because of the $\nabla \log p$, the Stein operator in Equation (3) does not depend on the normalising constant of p. In the rest of the paper we will focus on the measurable space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, hence we will use the Stein operator of Equation (3).

In addition to Stein's method, we provide background on reproducing kernel Hilbert spaces (RKHS) as the concepts introduced here will be at the basis of SVGD. We start by defining a Hilbert space and then move on by describing the RKHS.

1.2 Hilbert spaces

This definition is taken from [HN01]. Here, for simplicity, we define the notion of Hilbert space for a real space, note however that a Hilbert space can also be defined on complex spaces. Let $\Gamma(\mathcal{Y})$ be a linear functional space as defined in Section 1.1. The function

$$\langle \cdot, \cdot \rangle : \Gamma(\mathcal{Y}) \times \Gamma(\mathcal{Y}) \to \mathbb{R}$$

is an inner product if the following properties hold:

1. $\langle f, f \rangle = 0$ if and only if $f = 0$ (positive definite)
2. $\langle f, f \rangle \geq 0$ (nonnegative)
3. $\langle f, g \rangle = \langle g, f \rangle$ (symmetric)
4. $\langle f, \alpha_1 g_1 + \alpha_2 g_2 \rangle = \alpha_1 \langle f, g_1 \rangle + \alpha_2 \langle f, g_2 \rangle$ (linearity in the second argument)

for all $f, g_1, g_2 \in \Gamma(\mathcal{Y})$ and $\alpha_1, \alpha_2 \in \mathbb{R}$. Note that while an inner product in the real space is bilinear, i.e. the linearity can be in either the first or second argument, for a complex space this is not the case. Every inner product gives rise to a norm $\| \cdot \|$ as follows:

$$\| f \| = \sqrt{\langle f, f \rangle}$$

A linear space with an inner product is called inner product space and any inner product space is also a normed linear space. A Hilbert space is a complete[1] inner product space and is denoted as $\mathcal{H} = (\Gamma(\mathcal{Y}), \langle \cdot, \cdot \rangle)$ with $\Gamma(\mathcal{Y})$ complete under $\langle \cdot, \cdot \rangle$. In the following we use $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $\| \cdot \|_{\mathcal{H}}$ to denote the inner product and norm on Hilbert space $\mathcal{H}$, respectively. The definition of Hilbert space above is general and holds for functions on any space. In this paper we deal with real functions, therefore, we will use $\Gamma(\mathcal{Y}) = \Gamma(\mathbb{R})$.

Throughout the paper, we denote by $\mathcal{H}^d = \mathcal{H} \times \ldots \times \mathcal{H}$ the space of $d \times 1$ vector functions $\boldsymbol{f} = \{f_i : f_i \in \mathcal{H}\}_{i \in \{1,\ldots,d\}}$ with an inner product $\langle \boldsymbol{f}, \boldsymbol{g} \rangle_{\mathcal{H}^d} = \sum_{i=1}^d \langle f_i, g_i \rangle_{\mathcal{H}}$ for $\boldsymbol{f}$ and $\boldsymbol{g} = \{g_i\}_{i \in \{1,\ldots,d\}}$, and norm $\| \boldsymbol{f} \|_{\mathcal{H}^d} = \sqrt{\sum_{i=1}^d \| f_i \|_{\mathcal{H}}^2}$ as in [LLJ16].

1.3 Reproducing kernel Hilbert Spaces

Having defined a Hilbert space, we can now define a RKHS (in this we use the definition of [RW06]). Let $\mathcal{H}$ be a Hilbert space of real functions $f \in \Gamma(\mathbb{R})$. Then $\mathcal{H}$ is a reproducing kernel Hilbert space endowed with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ (and norm $\| f \|_{\mathcal{H}} = \sqrt{\langle f, f \rangle_{\mathcal{H}}}$) if there exists a measurable function $k : E \times E \to \mathbb{R}$ with the following properties:

1. for every fixed $x \in E$, $k(x, x')$ as a function of $x' \in E$ belongs to $\mathcal{H}$
2. k has the reproducing property $\langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$ and $\langle k(x, \cdot), k(x', \cdot) \rangle_{\mathcal{H}} = k(x, x')$

[1] Completeness means that every Cauchy sequence of elements of the space converges to an element in the space.

2 Open questions and research directions

This section discusses the details of SVGD [LW16] and future directions for research applying Stein's method to machine learning. We start by presenting the kernelized Stein discrepancy (KSD) and deriving an essential result for SVGD. We continue introducing optimal transport and variational inference. Next, derive the main result of the paper and show how its importance for SVGD. We conclude with a discussion of methodological improvements and topics for further research.
In the remaining of the paper we will use the measurable space $(E, \mathcal{E}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.

2.1 Kernelized Stein discrepancy

Given two measures $\mathbb{P}, \mathbb{Q}$ absolutely continuous with respect to the Lebesgue measure, the Stein discrepancy as defined in Equation (2) can be rewritten in the form

$$SD_{\mathcal{T}_p[\mathcal{G}]}(\mathbb{P}, \mathbb{Q}) = \max_{g \in \mathcal{G}} \left\{ E_{\mathbb{Q}} \left[\mathcal{T}_p[g](x)\right]^2 \right\}, \ x \in \mathbb{R}^d \tag{4}$$

Computation of the discrepancy in Equation (4) is not tractable, therefore, its use in machine learning has been limited. [GM15] derived a computationally tractable version of the discrepancy under some constraints that transformed the optimisation into a linear programming problem. A method to compute Stein discrepancy in a closed form appeared for the first time in [LLJ16]. The authors derived kernelized Stein discrepancy using vector-valued functions $\boldsymbol{g}$ in the RKHS $\mathcal{H}^d$. For a vector-valued function, we assume $\boldsymbol{g}$ is a column vector, the Stein operator can be written as

$$S_{\mathbb{P}}[\boldsymbol{g}] = (\nabla \log p)\boldsymbol{g}^T + \nabla \boldsymbol{g}$$

which results in a $d \times d$ matrix. However, in order to obtain a scalar value from the expectation in the Stein discrepancy, the trace of $S_{\mathbb{P}}[\boldsymbol{g}]$ is used in practice. Computing the elements on the diagonal of $S_{\mathbb{P}}[\boldsymbol{g}]$, it can be shown that

$$\text{trace}(S_{\mathbb{P}}[\boldsymbol{g}]) = \mathcal{T}_p[\boldsymbol{g}]$$

with $\mathcal{T}_p[\boldsymbol{g}]$ as in Equation (3). Therefore the Stein discrepancy can simply be rewritten in the following way

$$SD_{\mathcal{T}_p[\mathcal{G}]}(\mathbb{P}, \mathbb{Q}) = \max_{\boldsymbol{g} \in \mathcal{G}} \left\{ E_{\mathbb{Q}} \left[\mathcal{T}_p[\boldsymbol{g}](x)\right]^2 \right\}, \ x \in \mathbb{R}^d \tag{5}$$

Furthermore, for RKHS $\mathcal{H}$ with positive definite kernel $k(x, x')$ in $\mathcal{T}_p[\mathcal{G}]$, the Stein class $\mathcal{G}$ of p (see Definition 3.4 in [LLJ16] for the conditions), if we restrict $\boldsymbol{g}$ to the unit ball of $\mathcal{H}^d$, i.e. $\parallel \boldsymbol{g} \parallel_{\mathcal{H}^d} \leq 1$, the discrepancy in Equation (5) becomes

$$SD_{\mathcal{T}_p[\mathcal{H}^d]}(\mathbb{P}, \mathbb{Q}) = \max_{\boldsymbol{g} \in \mathcal{H}^d} \left\{ E_{\mathbb{Q}} \left[\mathcal{T}_p[\boldsymbol{g}](x)\right]^2, \ s.t. \parallel \boldsymbol{g} \parallel_{\mathcal{H}^d} \leq 1 \right\}, \ x \in \mathbb{R}^d \tag{6}$$

and has a closed form solution given by $\boldsymbol{g}^* = \boldsymbol{\beta}/ \parallel \boldsymbol{\beta} \parallel_{\mathcal{H}^d}$ where $\boldsymbol{\beta} = E_{\mathbb{Q}}\left[\mathcal{T}_p[k(x, \cdot)]\right]$. To prove this result we can use Definition 3.2 of [LLJ16] together with the reproducing property for $k(x, x')$ (condition 2 for an RKHS)

$$\begin{aligned}
E_{\mathbb{Q}}[\mathcal{T}_p[\boldsymbol{g}]]^2 &= E_{\mathbb{Q}}[(\nabla \log p - \nabla \log q)^T k(x, x')(\nabla \log p - \nabla \log q)] \\
&= E_{\mathbb{Q}}[(\nabla \log p - \nabla \log q)^T \langle k(x, \cdot), k(x', \cdot)\rangle_{\mathcal{H}} (\nabla \log p - \nabla \log q)] \\
&= \sum_{i=1}^{d} \langle E_{\mathbb{Q}}[(\nabla_i \log p - \nabla_i \log q) k(x, \cdot)], E_{\mathbb{Q}}[k(x', \cdot)(\nabla_i \log p - \nabla_i \log q)]\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{d} \langle \boldsymbol{\beta}_i, \boldsymbol{\beta}_i \rangle_{\mathcal{H}} = \parallel \boldsymbol{\beta} \parallel^2_{\mathcal{H}^d}, \ x, x' \in \mathbb{R}^d
\end{aligned}$$

from which $\boldsymbol{g}^*$ follows. We used q, the density of $\mathbb{Q}$ with respect to the Lebesgue measure, ∇_i, the derivative with respect to the i-th variable and $E_{\mathbb{Q}}\left[\mathcal{T}_p[k(x, \cdot)]\right] = E_{\mathbb{Q}}\left[(\nabla \log p - \nabla \log q) k(x, \cdot)\right]$ which can be derived using the fact that $k(x, x')$ is in the Stein class $\mathcal{G}$ of p. Using samples from q the optimal solution to $\boldsymbol{g}^*$ can be computed approximately using Equation (3).

2.2 Optimal transport

Before introducing how the results in the previous section can be used to design SVGD, we need to briefly introduce optimal transport as this will be essential in the next section.

Optimal transport deals with the problem of transporting mass from one distribution to another at the minimum cost possible and preserving total mass. Eventually, optimal transport tries to map points in the support of one distribution to the other. In probability terms, given two measures μ, ν on measurable spaces $(E, \mathcal{E})$ and $(D, \mathcal{D})$, respectively, optimal transport aims at finding a measurable function $T : E \to D$ such that a distance metric, d_{OT}, is minimised while total mass is preserved. The distance metric d_{OT} can be chosen depending on the measurable spaces considered. If we assume that $\mu \ll \nu$, $\nu \ll \mu$ and T is a diffeomorphism then, optimal transport can be simply seen as the change of variables given by the Radon-Nikodym theorem [PC+19]. Hence we have

$$\int_A g d\nu = \int_{T^{-1}(A)} (g \circ T) \frac{d\nu}{d\mu} d\mu, \ A \subset D \tag{7}$$

2.3 Variational inference

Having defined the kernelized Stein discrepancy and the core idea of optimal transport we can now analyse how the two concepts are used to design SVGD and perform variational inference. First we introduce variational inference and then we move on to a detailed description of SVGD. In the following discussion we assume that all measures are absolutely continuous with respect to the Lebesgue measure and indicate with a lower case letter the density of the corresponding measure e.g. p is the density of $\mathbb{P}$.

The aim of variational inference is to approximate a difficult distribution p, usually the posterior distribution of a Bayesian model, with an easier distribution q, from which we can easily obtain samples. Such approximation is carried out selecting a family of probability measures $\mathcal{Q} = \{\mathbb{Q}_\theta : \theta \in \Theta\}$ with $\mathbb{Q}_\theta \ll \mathbb{P}$ for all $\mathbb{Q}_\theta \in \mathcal{Q}$ and finding the optimal q_θ such that a divergence between p and q_θ is minimised. The KL divergence is usually chosen for this purpose, meaning that the problem of variational inference can be summarised as

$$q_\theta^* = \arg\min_{q_\theta \in \mathcal{Q}} \{KL(q_\theta \parallel p)\} = \arg\min_{q_\theta \in \mathcal{Q}} \left\{ \int_{\mathbb{R}^d} \log \frac{q_\theta}{p} d\mathbb{Q}_\theta \right\} = \arg\min_{q_\theta \in \mathcal{Q}} \left\{ E_{\mathbb{Q}_\theta} \left[\log \frac{q_\theta(x)}{p(x)} \right] \right\}, \ x \in \mathbb{R}^d \tag{8}$$

The choice of distribution families $\mathcal{Q}$ needs to strike a balance between accuracy of the approximation and tractability. Choosing a family of distributions that is too simple might not be able to capture complex characteristics of the distribution that we are trying to approximate despite making the minimisation problem easy. On the other hand, a too complex distribution family could provide a very accurate approximation, at the cost of increasing the computational cost of the minimisation. Therefore, choosing $\mathcal{Q}$ requires thought. Next, we show how [LW16] found a way to achieve the balance between computation and complexity by choosing $\mathcal{Q}$ such that the KL divergence can be related to Stein's method.

Examining the definition of the KL divergence, it is not clear how Stein's method can be related to variational inference. In fact, the KL divergence does not fall into the family of divergences that can be directly used to apply Stein's method (compare the KL divergence to Equation (1)). The work of [Sri+09] tried to relate IPMs to family of divergences of which the KL is a special case, however, theoretical results showed that the KL divergence cannot be directly related to IPMs therefore not allowing for a straightforward application of Stein's method to variational inference. Despite such theoretical results, [LW16] showed that by applying the basic idea of optimal transport from Section 2.2, together with a restriction to functions in the RKHS, the KL divergence can be related to KSD. In particular, the relation derived by [LW16] connects the gradient of the KL divergence to KSD. Although this is not a direct relation such as the one [Sri+09] tried to find, the connection between the gradient of the KL and KSD is sufficient to perform variational inference. In practice, the technique used to minimize Equation (8) is stochastic gradient descent, which only requires computation of the gradient of the KL divergence.

2.4 Stein Variational Gradient Descent

SVGD follows from the main result of [LW16] which we derive here. Let $T : \mathbb{R}^d \to \mathbb{R}^d$ define a smooth diffeomorphism that maps a random variable X with tractable distribution q_θ^0 to a random variable $Z = T(X)$ with distribution $q_\theta^{[T]}$, then using Equation (7) the distribution of Z can be written as

$$q_\theta^{[T]}(z) = q_\theta^0(T^{-1}(z))|\det(\nabla_z T^{-1}(z))| \tag{9}$$

with T^{-1} the inverse map of T and ∇T^{-1} the Jacobian of T^{-1}. Note that expectations with respect to $q_\theta^{[T]}$ can be easily evaluated averaging over $\{z_i\}_{i=1}^N$ with $z_i = T(x_i)$ and $x_i \sim q_\theta^0$. However, in this setting, a tractable and accurate parametric family $\mathcal{Q}$ needs to be specified and the transform T needs to have an efficiently computable Jacobian. SVGD, however, does not require a parametric family specification nor the Jacobian computation. Defining T as a small perturbation of the identity map, i.e. $T(x) = x + \epsilon g(x)$ with g a smooth function in the RKHS $\mathcal{H}$ indicating the perturbation direction and ϵ determining its magnitude, it is possible to relate the KL divergence to Stein discrepancy.

Given the transformation $T(x) = x + \epsilon g(x)$, it can be shown that

$$\nabla_\epsilon KL\left(q_\theta^{[T]} \parallel p\right)\Big|_{\epsilon=0} = -E_{\mathbb{Q}_\theta}\left[\mathcal{T}_p[g](x)\right] \tag{10}$$

Proof (Theorem 3.1 [LW16]): We first introduce a Lemma that is needed in the proof:

$$\nabla_\epsilon KL\left(q_\theta^{[T]} \parallel p\right) = E_{\mathbb{Q}_\theta}\left[(\langle\nabla \log p, \nabla_\epsilon T\rangle + \operatorname{trace}\left((\nabla_x T)^{-1}\nabla_\epsilon \nabla_x T\right))(x)\right], \ x \in \mathbb{R}^d \tag{11}$$

next, in the same way as for Equation (9), we define

$$p^{[T]}(z) = p(T^{-1}(z))|\det \nabla_z T^{-1}(z)| \tag{12}$$

and from the definitions in Equation (9) and Equation (12) and using the change of variables $x = T^{-1}(z)$ we can show that

$$\begin{aligned} KL\left(q_\theta^T \parallel p\right) &= E_{\mathbb{Q}_\theta^T}\left[\log \frac{q_\theta^T(z)}{p(T^{-1}(z))}\right] = E_{\mathbb{Q}_\theta^T}\left[\log q_\theta(T^{-1}(z)) - \log \frac{p^{[T]}(z)}{|\det \nabla_z T^{-1}(z)|}\right] \\ &= E_{\mathbb{Q}_\theta}\left[\log \frac{q_\theta(x)}{p^{[T]}(T^(x))}\right] = KL(q_\theta \parallel p^{[T]}) \end{aligned}$$

from which it follows

$$\nabla_\epsilon KL(q_\theta^T \parallel p) = -E_{\mathbb{Q}_\theta}\left[\nabla_\epsilon \log p^{[T]}(T(x))\right]$$

with

$$\nabla_\epsilon \log p^{[T]}(T(x)) = (\langle\nabla \log p, \nabla_\epsilon T\rangle + \operatorname{trace}\left((\nabla_x T)^{-1}\nabla_\epsilon \nabla_x T\right))(x)$$

following from Equation (11). Therefore, assuming $T(x) = x + \epsilon g(x)$ and $\epsilon = 0$ we have

$$T(x) = x, \quad \nabla_\epsilon T(x) = g(x), \quad \nabla_x T(x) = I, \quad \nabla_\epsilon \nabla_x T(x) = \nabla_x g(x)$$

with I the identity matrix. The result in Equation (10) follows substituting the elements above into Equation (11). □

Furthermore, if we restrict the functions g to the ball $\mathcal{B} = \{\boldsymbol{g} \in \mathcal{H}^d : \parallel \boldsymbol{g} \parallel_{\mathcal{H}^d}^2 \leq SD_{\mathcal{T}_{p[\mathcal{H}^d]}}(\mathbb{P}, \mathbb{Q}_\theta)\}$, the direction of steepest descent that maximises the negative gradient of the KL divergence is given by $\boldsymbol{\beta}_{\mathbb{Q}_\theta} = E_{\mathbb{Q}_\theta}\left[\mathcal{T}_p[k(x,\cdot)]\right]$, for which $\nabla_\epsilon KL(q_\theta^T \parallel p)\big|_{\epsilon=0} = -SD_{\mathcal{T}_{p[\mathcal{H}^d]}}(\mathbb{P}, \mathbb{Q}_\theta)$. The relation between the gradient of the KL divergence and Stein discrepancy can further be generalized to be seen as a step of functional gradient descent in RKHS (see [LW16] for more details on this), however, the results above suffice to define SVGD.

2.4.1 SVGD Algorithm

From the knowledge that $\boldsymbol{\beta}_{\mathbb{Q}_\theta}$ is the direction of steepest descent that maximises the negative gradient of the KL divergence, we can develop a modification to stochastic gradient descent (SGD) to incorporate such information. In particular, we can substitute the computation of the gradient of the KL divergence in each SGD iteration with the computation of $\boldsymbol{\beta}_{\mathbb{Q}_\theta}$.

The iterative procedure consists of two steps. The first step applies the transform $T_l(x) = x + \epsilon_l \boldsymbol{\beta}_{\mathbb{Q}_\theta^l}(x)$, which minimises the gradient of the KL divergence at step l. The second step updates the distribution $q_\theta^{l+1} = q_\theta^{[T_l]}$ such that the relation between the gradient of the KL divergence and Stein discrepancy is preserved. In this procedure, ϵ_l is a step size parameter, usually updated automatically by the optimisation method used, that when sufficiently small makes the Jacobian of T full rank (close to the identity matrix) [LW16]. In addition to this, it is necessary to notice that $\boldsymbol{\beta}_{\mathbb{Q}_\theta}$ requires the computation of an expectation, which can be approximated using a set of particles $\{x_i\}_{i=1}^N$. The particles used in the approximation can be sampled from q_θ^0 in the first iteration of the algorithm and updated at each iteration, using the transform T, to preserve all the properties derived. Algorithm 1, taken from [LW16], summarises SVGD, in practice, the RBF kernel $k(x, x') = \exp\left(-\frac{1}{h} \parallel x - x' \parallel^2\right)$ is used as it falls in the Stein class of smooth densities (see the solution Exercise 1 for a proof).

```
Algorithm 1: Bayesian Inference via Variational Gradient Descent
Input: A target distribution with density function $p(x)$ and a set of initial particles $\{x_i^0\}_{i=1}^N$
Output: A set of particles $\{x_i\}_{i=1}^N$ that approximates the target distribution
for iteration $l$ do
    $x_i^{l+1} \leftarrow x_i^l + \epsilon_l \hat{f}(x_i^l) \quad \text{where} \quad \hat{f}(x) = \frac{1}{N} \sum_{j=1}^{N} \left[ k(x_j^l, x) \nabla_{x_j^l} \log p(x_j^l) + \nabla_{x_j^l} k(x_j^l, x) \right]$
    where $\epsilon_l$ is the step size at the $l$-th iteration.
end
```

In Algorithm 1 the particles $\{x_i\}_{i=1}^N$ do not depend on the initial distribution q_θ^0, indicating that SVGD can be applied using particles generated by other procedures that could provide better starting samples. Furthermore, we expect that as N increases, we obtain a better approximation of the expectation and therefore a better optimisation result, some theoretical results for SVGD have been derived in [Liu17; LW18].

2.4.2 Computational challenges

Despite SVGD performed better than existing methods in experimental results (see [LW16]), computational issues arise in two parts of the method. Firstly, in each iteration, the gradient of $\log p(x)$ has to be computed for each particle used in the approximation. The authors propose to compute the gradient using mini-batches of the data and parallelizing the gradient computation on multiple cores. Secondly, although $\hat{f}(x)$ only requires the computation of some entries of the kernel matrix, computing the approximation for each particle in the approximation eventually requires the entire kernel matrix. The kernel matrix contains N^2 entries, its computation could therefore be highly expensive in situations where a large number of particles is used in the approximation. Although the results on toy examples in [LW16] show that 100 particles are enough to obtain a good approximation, for harder distributions more particles might be needed. To overcome this issue the authors propose to subsample particles resulting in a smaller kernel matrix. We believe that methods from the Gaussian processes literature could also be used to efficiently compute the kernel matrix.

2.5 Discussion and future work

In this paper we reformulated Stein variational gradient descent by [LW16] in more theoretical terms. After introducing Stein's method, RKHS, we derived the relation between the gradient of the KL divergence and Stein discrepancy, which allowed to formalise SVGD. Since the introduction of SVGD in 2016, the literature on Stein's method and machine learning has developed rapidly with theoretical work as well as applications. SVGD has been generalised to graphical models [WZL17], reinforcement learning [Liu+17] and mixture models [WL19]. Furthermore, improvements to the original version of SVGD presented in this paper have beed proposed, for instance, [HL17] introduced ideas from importance sampling into SVGD and [HL18] proposed a gradient-free version of the algorithm. More recently, [Wan+19] generalised SVGD replacing the scalar valued kernels with matrix valued kernels and were able to use second order elements, such as the Hessian, in the KL minimisation.

Despite the advances in the literature using Stein's method for machine learning, there are some research directions that future work could explore. Firstly, in the derivation of Stein discrepancy in Equation (5), the literature uses the trace of the kernel matrix to ensure the expectation results in a scalar value. In [LLJ16], the authors do not provide insight into why the trace is chosen over other functions such as the determinant for example. We saw that the solution to Stein discrepancy does not simplify computations in practice as the full kernel matrix needs to be computed for SVGD. Exploring other methods to enforce the expectation to be a scalar in Stein discrepancy would not cause increased computational cost but might lead to functions that provide better approximations. Redefining Stein discrepancy could lead however, to a loss in the relation between KL and Stein discrepancy.

Furthermore, the paper did not investigate other one-to-one transformations that would leave the derivations invariant. The added value of investigating different transform function lies in the fact that approximations of the target distribution could converge faster. Instead of limiting the set of transforms to a single function, we propose to use neural network architectures especially designed to provide reversible transformations with a trivial Jacobian. These architectures, such as the NICE network of [DKB14] or normalizing flows, could be optimised during the iterations of SVGD so as to provide better transformations that preserve the SVGD properties. Although the design of these architectures would require careful study, the flexibility added to the algorithm could be substantial and impact the convergence speed. Obviously, adding neural network architectures within SVGD would likely increase the time per iteration but this might be balanced out by a faster convergence.

Lastly, it would be interesting to investigate the performance of different kernels, with the important restriction to those in the Stein class.

A Exercises

A.1 Exercise 1

In practical applications of SVGD, the RBF kernel is used with the never explained argument that the RBF kernel is in the Stein class of p. In this exercise we will give more thought to the reasons why the RBF kernel can be used in SVGD.
A kernel $k(x, x')$ is said to be in the Stein class of p, with p the density of a probability measure $\mathbb{P}$ with respect to the Lebesgue measure, if $k(x, x')$ has continuous second order partial derivatives, and both $k(x, \cdot)$ and $k(\cdot, x)$ are in the Stein class of p for any fixed x [LLJ16].

Verify that the RBF kernel $k(x, x') = \exp\left(-\frac{1}{2h^2} \parallel x - x' \parallel_2^2\right)$ is in the Stein class for smooth densities with support on $\mathbb{R}^d$.
Proof: The proof of this result is trivial and hinges on the Taylor series expansion of exponential functions

$$\exp(x) = \sum_{i=0}^{\infty} \frac{x^i}{i!}$$

From this it can easily be observed that for $k(x, x')$ the second order partial derivatives are continuous and that $k(x, \cdot)$ and $k(\cdot, x)$ have continuous second order derivatives, therefore the RBF kernel is in the Stein class of p.

A.2 Exercise 2

The work in the paper introducing SVGD has been motivated and enabled by the possibility of efficiently computing the kernelized Stein discrepancy. In this exercise we will derive the closed form solution of the kernelized Stein discrepancy. This exercise consists in a series of results from [LLJ16].

1. For a vector-valued function $\boldsymbol{g}$ to be in the Stein class of p the following must hold:

   $$E_{\mathbb{P}}[\mathcal{S}_{\mathbb{P}}[\boldsymbol{g}]] = E_{\mathbb{P}}[(\nabla \log p)\boldsymbol{g}^T + \nabla \boldsymbol{g}] = 0$$

   Using this knowledge, prove that for densities p and q of $\mathbb{P}$ and $\mathbb{Q}$, respectively, for which $\mathbb{Q} \ll \mathbb{P}$ and $\boldsymbol{g}$ in the Stein class of p we have

   $$E_{\mathbb{P}}[\mathcal{S}_{\mathbb{Q}}[\boldsymbol{g}]] = E_{\mathbb{P}}[(\nabla \log q - \nabla \log p)\boldsymbol{g}^T]$$

   Proof: The proof is straighforward using the fact that $\boldsymbol{g}$ is in the Stein class of p. Since $E_{\mathbb{P}}[\mathcal{S}_{\mathbb{P}}[\boldsymbol{g}]] = 0$,

   $$E_{\mathbb{P}}[\mathcal{S}_{\mathbb{Q}}[\boldsymbol{g}]] = E_{\mathbb{P}}[\mathcal{S}_{\mathbb{Q}}[\boldsymbol{g}] - \mathcal{S}_{\mathbb{P}}[\boldsymbol{g}]] = E_{\mathbb{P}}[(\nabla \log q - \nabla \log p)\boldsymbol{g}^T]$$

2. If $k(x, x')$ is in the Stein class of p, so is any $f \in \mathcal{H}$. Why is this claim true?
   Solution: To show that this is true we can use the reproducing property of the kernel k. We know that $k(x, x') = \langle k(x, \cdot), k(x', \cdot)\rangle_{\mathcal{H}}$ with both $k(x, \cdot)$ and $k(x', \cdot)$ belonging to $\mathcal{H}$. Furthermore, from the reproducing property of the kernel we know that $f(x) = \langle f(\cdot), k(\cdot, x)\rangle_{\mathcal{H}}$ and $f(\cdot)$ is a function belonging to $\mathcal{H}$. For $f(\cdot) = k(x, \cdot)$ we know that the claim holds, therefore, it must hold also for any other function $f(\cdot) \in \mathcal{H}$.

3. Using the results in the previous sub-questions and denoting

   $$\begin{aligned} u_q(x, x') =& (\nabla \log q(x))^T k(x, x') \nabla \log q(x') + (\nabla \log q(x))^T \nabla_{x'} k(x, x') \\ &+ (\nabla_x k(x, x'))^T \nabla \log q(x') + \operatorname{trace}(\nabla_{x,x'} k(x, x')) \end{aligned}$$

   show that

   $$SD_{\mathcal{S}_{\mathbb{P}}[\mathcal{G}]}(\mathbb{Q}, \mathbb{P}) = E_{\mathbb{P}}[u_q(x, x')]$$

where

$$SD_{\mathcal{S}_{\mathbb{P}}[\mathcal{G}]}(\mathbb{Q},\mathbb{P}) = E_{\mathbb{P}}[(\nabla \log q(x) - \nabla \log p(x))^T k(x,x')(\nabla \log q(x') - \nabla \log p(x'))]$$

Proof: Proving this result can be done applying the result in part 1. twice, first on $k(\cdot, x')$ for fixed x' and then with fixed x. We start by denoting $v(x,x') = k(x,x')\nabla \log q(x') + \nabla_{x'} k(x,x') = \mathcal{S}_{\mathbb{Q}}[k(x,x')]$. Applying the result in part 1. on $k(x,\cdot)$ with fixed x we have

$$\begin{aligned} SD_{\mathcal{S}_{\mathbb{P}}[\mathcal{G}]}(\mathbb{Q},\mathbb{P}) &= E_{\mathbb{P}}[(\nabla \log q(x) - \nabla \log p(x))^T k(x,x')(\nabla \log q(x') - \nabla \log p(x'))] \\ &= E_{\mathbb{P}}[(\nabla \log q(x) - \nabla \log p(x))^T v(x,x')] \end{aligned}$$

follows by expanding the product, adding and subtracting $\nabla_{x'} k(x,x')$ in the expectation and simplifying the expression based on the fact that $E_{\mathbb{P}}[\mathcal{S}_{\mathbb{P}}[k(x,x')]] = 0$. The latter result holds because $k(\cdot,x')$ is in the Stein class of p and so is $\nabla_{x'} k(\cdot,x')$ and therefore also $v(\cdot,x')$. Applying the result in part 1. on $v(\cdot,x')$ with fixed x' we obtain

$$\begin{aligned} SD_{\mathcal{S}_{\mathbb{P}}[\mathcal{G}]}(\mathbb{Q},\mathbb{P}) &= E_{\mathbb{P}}[(\nabla \log q(x) - \nabla \log p(x))^T (v(x,x') + \mathrm{trace}(\nabla_x v(x,x')) - \mathrm{trace}(\nabla_x v(x,x')))] \\ &= E_{\mathbb{P}}[(\nabla \log q(x)^T v(x,x') + \mathrm{trace}(\nabla_x v(x,x'))] \end{aligned}$$

where we used the trace to ensure that the result is a scalar, depending on the definition of Stein discrepancy other functions could be used. Lastly, noting that $\nabla_x v(x,x') = \nabla_x k(x,x') s_q(x')^T + \nabla_{x,x'} k(x,x')$ we obtain the desired result.

References

[Ste72] C. Stein. "A bound for the error in the normal approximation to the distribution of a sum of dependent random variables". In: *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory.* The Regents of the University of California. 1972.

[FM06] M. Franceschetti and R. Meester. "Critical Node Lifetimes in Random Networks via the Chen-Stein Method". In: *IEEE/ACM Trans. Netw.* 14.SI (June 2006), pp. 2831–2837. ISSN: 1063-6692. DOI: 10.1109/TIT.2006.874545. URL: https://doi.org/10.1109/TIT.2006.874545.

[RSW00] G. Reinert, S. Schbath, and M. S. Waterman. "Probabilistic and statistical properties of words: an overview". In: *Journal of Computational Biology* 7.1-2 (2000), pp. 1–46.

[BB90] F. Ball and A. Barbour. "Poisson approximation for some epidemic models". In: *Journal of applied probability* 27.3 (1990), pp. 479–490.

[Rei11] G. Reinert. "A short introduction to Stein's method". In: *Lecture Notes* (2011).

[OGC17] C. J. Oates, M. Girolami, and N. Chopin. "Control functionals for Monte Carlo integration". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.3 (2017), pp. 695–718.

[Oat+19] C. J. Oates et al. "Convergence rates for a class of estimators based on Stein's method". In: *Bernoulli* 25.2 (2019), pp. 1141–1159.

[GM15] J. Gorham and L. Mackey. "Measuring sample quality with Stein's method". In: *Advances in Neural Information Processing Systems.* 2015, pp. 226–234.

[LLJ16] Q. Liu, J. Lee, and M. Jordan. "A kernelized Stein discrepancy for goodness-of-fit tests". In: *International conference on machine learning.* 2016, pp. 276–284.

[CSG16] K. Chwialkowski, H. Strathmann, and A. Gretton. "A kernel test of goodness of fit". In: JMLR: Workshop and Conference Proceedings. 2016.

[YRN19] J. Yang, V. Rao, and J. Neville. "A Stein–Papangelou Goodness-of-Fit Test for Point Processes". In: *The 22nd International Conference on Artificial Intelligence and Statistics.* 2019, pp. 226–235.

[Kan+19] H. Kanagawa et al. "A Kernel Stein Test for Comparing Latent Variable Models". In: *arXiv preprint arXiv:1907.00586* (2019).

[LW16] Q. Liu and D. Wang. "Stein variational gradient descent: A general purpose bayesian inference algorithm". In: *Advances in neural information processing systems.* 2016, pp. 2378–2386.

[Zhu+17] J. Zhuo et al. "Message passing stein variational gradient descent". In: *arXiv preprint arXiv:1711.04425* (2017).

[WZL17] D. Wang, Z. Zeng, and Q. Liu. "Stein variational message passing for continuous graphical models". In: *arXiv preprint arXiv:1711.07168* (2017).

[HL18] J. Han and Q. Liu. "Stein variational gradient descent without gradient". In: *arXiv preprint arXiv:1806.02775* (2018).

[SG18] K. Shaloudegi and A. György. "Adaptive MCMC via Combining Local Samplers". In: *arXiv preprint arXiv:1806.03816* (2018).

[Che+19] W. Y. Chen et al. "Stein Point Markov Chain Monte Carlo". In: *arXiv preprint arXiv:1905.03673* (2019).

[Ros+11] N. Ross et al. "Fundamentals of Stein's method". In: *Probability Surveys* 8 (2011), pp. 210–293.

[Bar+19] A. Barp et al. "Minimum Stein Discrepancy Estimators". In: *arXiv preprint arXiv:1906.08283* (2019).

[Ste86] C. Stein. "Approximate computation of expectations". In: IMS. 1986.

[Sri+09] B. K. Sriperumbudur et al. "On integral probability metrics,\phi-divergences and binary classification". In: *arXiv preprint arXiv:0901.2698* (2009).

[HN01] J. K. Hunter and B. Nachtergaele. *Applied analysis.* World Scientific Publishing Company, 2001.

[RW06] C. E. Rasmussen and C. K. Williams. "Gaussian Processes for Machine Learning". In: *Gaussian Processes for Machine Learning, by CE Rasmussen and CKI Williams. ISBN-13 978-0-262-18253-9* (2006).

[PC+19] G. Peyré, M. Cuturi, et al. "Computational optimal transport". In: *Foundations and Trends® in Machine Learning* 11.5-6 (2019), pp. 355–607.

[Liu17] Q. Liu. "Stein variational gradient descent as gradient flow". In: *Advances in neural information processing systems.* 2017, pp. 3115–3123.

[LW18] Q. Liu and D. Wang. "Stein variational gradient descent as moment matching". In: *Advances in Neural Information Processing Systems.* 2018, pp. 8854–8863.

[Liu+17] Y. Liu et al. "Stein variational policy gradient". In: *arXiv preprint arXiv:1704.02399* (2017).

[WL19] D. Wang and Q. Liu. "Nonlinear Stein Variational Gradient Descent for Learning Diversified Mixture Models". In: *International Conference on Machine Learning.* 2019, pp. 6576–6585.

[HL17] J. Han and Q. Liu. "Stein variational adaptive importance sampling". In: *arXiv preprint arXiv:1704.05201* (2017).

[Wan+19] D. Wang et al. "Stein Variational Gradient Descent With Matrix-Valued Kernels". In: *Advances in Neural Information Processing Systems.* 2019, pp. 7834–7844.

[DKB14] L. Dinh, D. Krueger, and Y. Bengio. "Nice: Non-linear independent components estimation". In: *arXiv preprint arXiv:1410.8516* (2014).