

# SVGD

exaFLOPs (baikdenny@gmail.com)

2024

## 1 Introduction

In many fields, we often deal with intractable distributions that are too complicated to work with directly. To handle this, methods like Stein Variational Gradient Descent (SVGD) help us approximate these complex distributions. SVGD is part of a group of methods called Variational Inference (VI). VI methods approximate a complicated distribution using a simpler one, like a Gaussian. But unlike most VI methods, SVGD doesn't need to stick to a specific type of simple distribution. Instead, it uses a set of points (or particles) that move to match the target distribution, giving it more flexibility.

## 2 Algorithm

This is the Algorithm from Liu and Wang, "Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm" [1]. We can see that the initial particles following the distribution  $q$  are getting updated step-by-step. Now we need to verify how the steps recursively go on.

---

**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  $p(x)$ .

**for** iteration  $\ell$  **do**

$$x_i^{\ell+1} \leftarrow x_i^\ell + \epsilon_\ell \hat{\phi}^*(x_i^\ell) \quad \text{where} \quad \hat{\phi}^*(x) = \frac{1}{n} \sum_{j=1}^n \left[ k(x_j^\ell, x) \nabla_{x_j^\ell} \log p(x_j^\ell) + \nabla_{x_j^\ell} k(x_j^\ell, x) \right],$$

where  $\epsilon_\ell$  is the step size at the  $\ell$ -th iteration.

**end for**

---

## 3 Mathematical 3 Steps

### 3.1 Function $\phi(x)$

As seen in the algorithm, the particles are updated by the function  $\phi$ . Then we need to find a proper function  $\phi$  to approximate the target distribution. So I will define a space where we can define those functions.

#### 3.1.1 Reproducing Kernel Hilbert Space(RKHS)

**Definition 3.1** A Hilbert space  $\mathcal{H}$  is an inner product space that is a complete metric space with respect to the norm or distance function induced by the inner product.

**Definition 3.2**  $\mathcal{H}$ , a Hilbert space of  $\mathbb{R}$ -valued functions on non-empty set  $\mathcal{X}$ . A function  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is a **reproducing kernel** of  $\mathcal{H}$ , and  $\mathcal{H}$  is a **reproducing kernel Hilbert space**, if

- $\forall x \in \mathcal{X}, k(\cdot, x) \in \mathcal{H}$ ,
- $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$  (the reproducing property).

In Reproducing Kernel Hilbert Space, evaluation of function  $f$  at  $x$  is an **inner product in feature space**.

Let  $k(x, x') : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  be a positive definite kernel, and  $\phi(x) = k(\cdot, x)$  a valid feature map. In particular, for any  $x, y \in \mathcal{X}$ ,

$$\begin{aligned} k(\cdot, y) &= \phi(y), \\ k(x, y) &= \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}, \end{aligned}$$

$$k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}}. \quad (1)$$

Consider the kernel function  $k(x, y)$  which is a function of two variables. Suppose, for  $n$  points, we fix one of the variables to have  $k(x_1, y), k(x_2, y), \dots, k(x_n, y)$ . These are all functions of the variable  $y$ . RKHS is a function space which is the set of all possible **linear combinations** of these functions. Therefore, a RKHS  $\mathcal{H}$  of  $k(x, x')$  is the closure of the linear span

$$\left\{ f : f(x) = \sum_{i=1}^n a_i k(x, x_i), a_i \in \mathbb{R}, n \in \mathbb{N}, x_i \in \mathcal{X} \right\}. \quad (2)$$

$$\begin{aligned} \langle f, g \rangle_{\mathcal{H}} &\stackrel{(2)}{=} \left\langle \sum_{i=1}^n \alpha_i k(\cdot, x_i), \sum_{j=1}^n \beta_j k(\cdot, y_j) \right\rangle_{\mathcal{H}} \\ &\stackrel{(1)}{=} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \beta_j k(x_i, y_j) \end{aligned}$$

Denote by  $\mathcal{H}^d$  the space of vector functions  $f = [f_1, \dots, f_d]^\top$  with  $f_i \in \mathcal{H}$ , equipped with the inner product

$$\langle f, g \rangle_{\mathcal{H}^d} = \sum_{i=1}^d \langle f_i, g_i \rangle_{\mathcal{H}}.$$

**Remark 3.3** RKHS is a space of functions and not a space of vectors. In other words, the basis vectors of RKHS are basis functions named eigenfunctions. Because the RKHS is a space of functions rather than a space of vectors, we usually do not know the exact location of pulled points to the RKHS but we know their relation of them as a function.

At last, **“Reproducing”** got its name as functions that are elements of RKHS,  $\langle f, g \rangle_{\mathcal{H}}$  can be reduced into inner product of  $x$ 's in  $\mathcal{X}$ . Every positive definite kernel  $k$  defines a unique RKHS for which  $k$  is a reproducing kernel.

### 3.1.2 Stein's Identity and Kernel Stein Discrepancy(KSD)

I can start by introducing about Stein's Identity and KSD, but will just start and explain about those concepts when needed.

Now let's handle with the initial particles  $\{x_i^0\}_{i=1}^n$ . Initial particles start from  $q_0$  distribution, but as the algorithm goes on, distributions will be updated. Set  $\mathcal{Q}$  is the set of distributions  $q_i (i = 0, 1, 2, \dots)$  of random

variables of form  $z = T(x)$  where  $T : \mathcal{X} \rightarrow \mathcal{X}$  is a smooth one-to-one transform, and  $x$  is drawn from a tractable reference distribution  $q_0(x)$ . By the change of variables formula, the density of  $z$  is

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

where  $T^{-1}$  denotes the inverse map of  $T$  and  $\nabla_z T^{-1}$  the Jacobian matrix of  $T^{-1}$ . Such distributions are computationally tractable, in the sense that the expectation under  $q_{[T]}$  can be easily evaluated by averaging  $\{z_i\}$  when  $z_i = T(x_i)$  and  $x_i \sim q_0$ . Such  $Q$  can also, in principle, closely approximate almost arbitrary distributions.

Our goal, typically VI's goal, is to approximate the target distribution  $p(x)$  using a simpler distribution  $q^*(x)$  by minimizing KL divergence. In following

$$q^* = \arg \min_{q \in Q} \{\text{KL}(q \parallel p) \equiv \mathbb{E}_q[\log q(x)] - \mathbb{E}_q[\log \bar{p}(x)] + \log Z\}, \text{ where } p(x) = \bar{p}(x)/Z \quad (3)$$

where we do not need to calculate the constant  $\log Z$  for solving the optimization. The choice of set  $Q$  is critical and defines different types of variational inference methods. The best set  $Q$  should strike a balance between

1. **Accuracy**, broad enough to closely approximate a large class of target distributions.
2. **Tractability**, consisting of simple distributions that are easy for inference.
3. **Solvability**, so that the subsequent KL minimization problem can be efficiently solved.

In practice, however, we need to restrict the set of transforms  $T$  properly to make the corresponding variational optimization in (3) practically solvable. One approach is to consider  $T$  with certain parametric form and optimize the corresponding parameters. However, this introduces a difficult problem on selecting the proper parametric family to balance the accuracy, tractability and solvability, especially considering that  $T$  has to be an one-to-one map and has to have an efficiently computable Jacobian matrix.

Instead, this paper proposed a new algorithm that iteratively constructs incremental transforms that effectively perform steepest descent on **T in RKHS**. Our algorithm does not require to explicitly specify parametric forms, nor to calculate the Jacobian matrix, and has a particularly simple form that mimics the typical gradient descent algorithm, making it easily implementable. From now on, I will state some theorems and prove them.

**Theorem 3.4** *Let  $T(x) = x + \epsilon\phi(x)$  and  $q_{[T]}(z)$  the density of  $z = T(x)$  when  $x \sim q(x)$ , we have*

$$\nabla_{\epsilon} \text{KL}(q_{[T]} \parallel p)|_{\epsilon=0} = -\mathbb{E}_{x \sim q}[\text{trace}(\mathcal{A}_p \phi(x))], \quad (4)$$

*where  $\mathcal{A}_p \phi(x) = \nabla_x \log p(x) \phi(x)^{\top} + \nabla_x \phi(x)$  is the Stein operator.*

Theorem 3.4 gives us the rate of change of the KL as  $\epsilon$  increases, for given  $\phi$ . Now, we want to pick  $\phi$  such that

$$-\mathbb{E}_{x \sim q}[\text{trace}(\mathcal{A}_p \phi(x))]$$

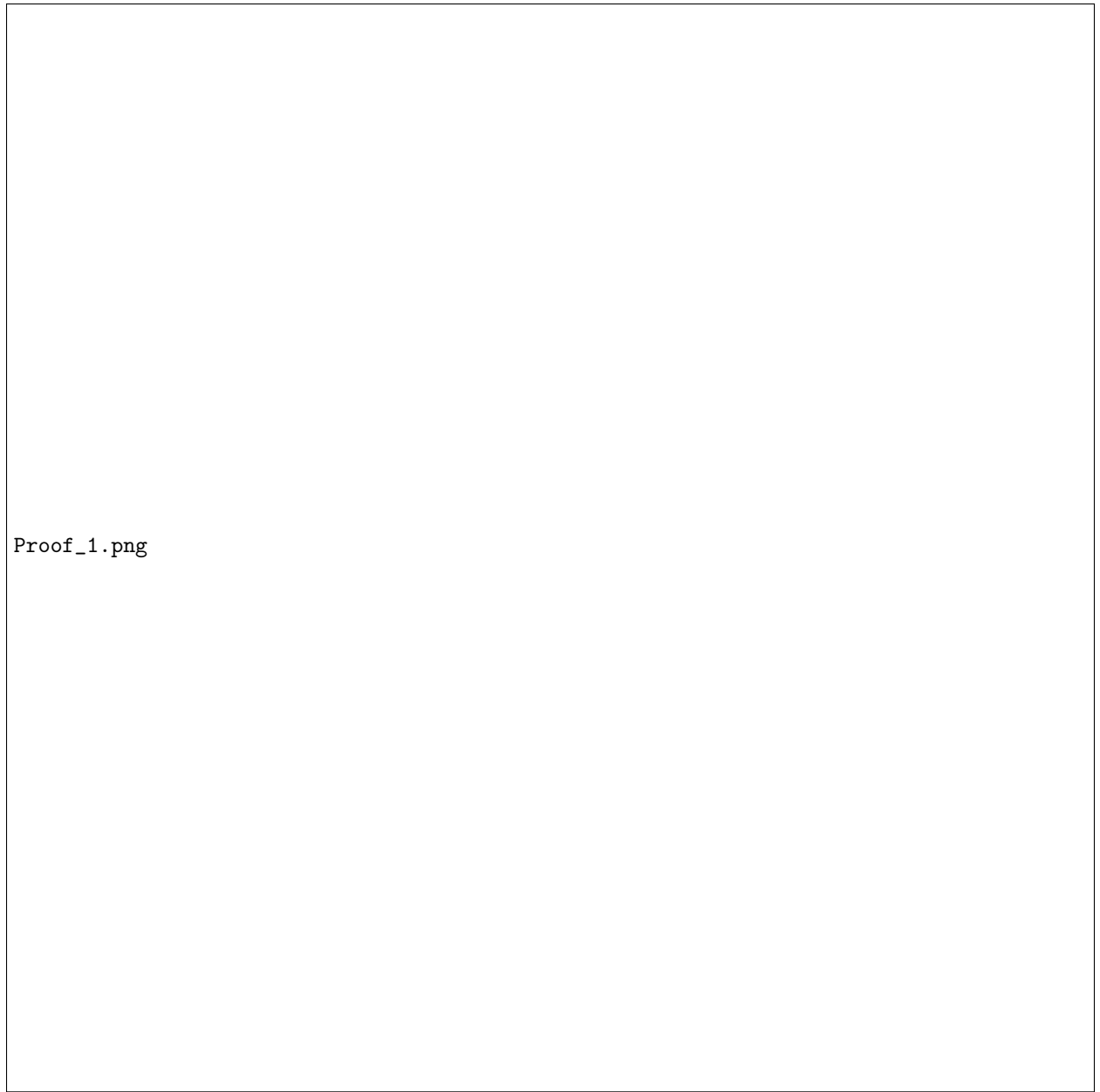
is as negative as possible. However, this minimisation has two problem

1. **Not well-defined**, because one can scale  $\phi$  by an arbitrary scalar, making the expectation unbounded.
2. **Not analytically or computationally tractable**

This issue can be resolved by considering a **constrained version of this optimization problem instead, using Reproducing Kernel Hilbert Spaces (RKHS)**.

At this point, I will give some explanation on Kernelized Stein Discrepancy(KSD).

Kernelized Stein Discrepancy (KSD) is a statistical measure used to quantify the difference between a target probability distribution  $p(x)$  and a test distribution  $q(x)$ , especially when  $p(x)$  is only known up to



Proof\_1.png

Figure 1: Proof of Theorem 3.4

a normalization constant. KSD leverages tools from Stein's method and reproducing kernel Hilbert spaces (RKHS), making it a flexible and efficient way to compare distributions. Kernelized Stein discrepancy (KSD) bypasses the difficulty mentioned above, by maximizing  $\phi$  in the unit ball of a reproducing kernel Hilbert space (RKHS) for which the optimization has a closed-form solution. KSD is defined as

$$\mathcal{D}(q, p) = \max_{\phi \in \mathcal{H}^d} \{ \mathbb{E}_{x \sim q} [\text{trace}(A_p \phi(x))] \}, \quad \text{s.t.} \quad \|\phi\|_{\mathcal{H}^d} \leq 1, \quad (5)$$

where we assume the kernel  $k(x, x')$  of RKHS  $\mathcal{H}$  is in the Stein class of  $p$  as a function of  $x$  for any fixed  $x' \in \mathcal{X}$ . I will prove this too. To give an insight of what we have done, the **gradient of KL-Divergence is the kernelized stein discrepancy (KSD)**.

**Theorem 3.5** *The optimal solution of (5) has been shown to be  $\phi(x) = \phi_{q,p}^*(x) / \|\phi_{q,p}^*\|_{\mathcal{H}^d}$ , where*

$$\phi_{q,p}^*(\cdot) = \mathbb{E}_{x \sim q} [A_p k(x, \cdot)],$$

for which we have

$$\mathcal{D}(q, p) = \|\phi_{q,p}^*\|_{\mathcal{H}^d}. \quad (6)$$

One can further show that  $\mathcal{D}(q, p)$  equals zero (and equivalently  $\phi_{q,p}^*(x) \equiv 0$ ) if and only if  $p = q$  once  $k(x, x')$  is strictly positive definite in a proper sense, which is satisfied by commonly used kernels such as the RBF kernel  $k(x, x') = \exp(-\frac{1}{h}\|x - x'\|_2^2)$ . Note that the RBF kernel is also in the Stein class of smooth densities supported in  $\mathcal{X} = \mathbb{R}^d$  because of its decaying property.

Both Stein operator and KSD depend on  $p$  only through the score function  $\nabla_x \log p(x)$ , which can be calculated without knowing the normalization constant of  $p$ , because we have  $\nabla_x \log p(x) = \nabla_x \log \tilde{p}(x)$  when  $p(x) = \tilde{p}(x)/Z$ . This property makes Stein's identity a powerful tool for handling unnormalized distributions that appear widely in machine learning and statistics.

**Lemma 3.6** *Assume the conditions in Theorem 3.4. Consider all the perturbation directions  $\phi$  in the ball  $\mathcal{B} = \{\phi \in \mathcal{H}^d : \|\phi\|_{\mathcal{H}^d} \leq \mathcal{D}(q, p)\}$  of vector-valued RKHS  $\mathcal{H}^d$ , the direction of steepest descent that maximizes the negative gradient in (4) is the  $\phi_{q,p}^*$  in (6), i.e.,*

$$\phi_{q,p}^*(\cdot) = \mathbb{E}_{x \sim q} [\nabla_x \log p(x) k(x, \cdot) + \nabla_x k(x, \cdot)], \quad (7)$$

for which (4) equals the square of KSD, that is,

$$\nabla_{\epsilon} KL(q_{[T]} \| p) \big|_{\epsilon=0} = -\mathcal{D}^2(q, p).$$

The result in Lemma (3.6) **covers the deepest direction of descent** and suggests an iterative procedure that transforms an initial reference distribution  $q_0$  to the target distribution  $p$ : we start with applying transform  $T_0^*(x) = x + \epsilon_0 \cdot \phi_{q_0,p}^*(x)$  on  $q_0$  which decreases the KL divergence by an amount of  $\epsilon_0 \cdot \mathcal{D}^2(q_0, p)$ , where  $\epsilon_0$  is a small step size; this would give a new distribution  $q_1(x) = q_0[T_0](x)$ , on which a further transform  $T_1^*(x) = x + \epsilon_1 \cdot \phi_{q_1,p}^*(x)$  can further decrease the KL divergence by  $\epsilon_1 \cdot \mathcal{D}^2(q_1, p)$ . Repeating this process one constructs a path of distributions  $\{q_\ell\}_{\ell=1}^n$  between  $q_0$  and  $p$  via

$$q_{\ell+1} = q_\ell[T_\ell^*],$$

where

$$T_\ell^*(x) = x + \epsilon_\ell \cdot \phi_{q_\ell,p}^*(x). \quad (8)$$

This would eventually converge to the target  $p$  with sufficiently small step-size  $\{\epsilon_\ell\}$ , under which  $\phi_{p,q_\infty}^*(x) \equiv 0$  and  $T_\infty^*$  reduces to the identity map. Recall that  $q_\infty = p$  if and only if  $\phi_{p,q_\infty}^*(x) \equiv 0$ .

### 3.2 Stein variational gradient descent

## 4 Related Research

### 4.1 MPPI