

SVGD

Justin  
Pauckert

Preliminary  
Problem Setup

Stein's  
Identity

Algorithm

References

# Stein Variational Gradient Descend

Justin Pauckert

Monte Carlo Methods in Machine Learning and Artificial Intelligence  
TU Berlin

July 30, 2020

# Overview

SVGD

Justin  
Pauckert

Preliminary

Problem Setup

Stein's  
Identity

Algorithm

References

## 1 Preliminary

## 2 Problem Setup

## 3 Stein's Identity

## 4 Algorithm

## 5 References

# What...

SVGD

Justin  
Pauckert

Preliminary

Problem Setup

Stein's  
Identity

Algorithm

References

- ... is Gradient Descend?
- ... are Kernels?
- ... is KL divergence?

# What is Gradient Descend?

SVGD

Justin  
Pauckert

Preliminary  
Problem Setup

Stein's  
Identity

Algorithm

References

Let  $F$  be a real-valued function that is differentiable in a neighborhood of a point  $a$ . Ultimately, we want to minimize  $F(x)$ . To find such a (local) minimum, we choose

$$a_{n+1} = a_n - \epsilon \cdot \nabla F(a_n),$$

where  $\epsilon \in \mathbb{R}$  is a small step size. Then, if  $\epsilon$  is small enough, we have

$$F(a_{n+1}) \leq F(a_n).$$

# What is Gradient Descend?

SVGD

Justin  
Pauckert

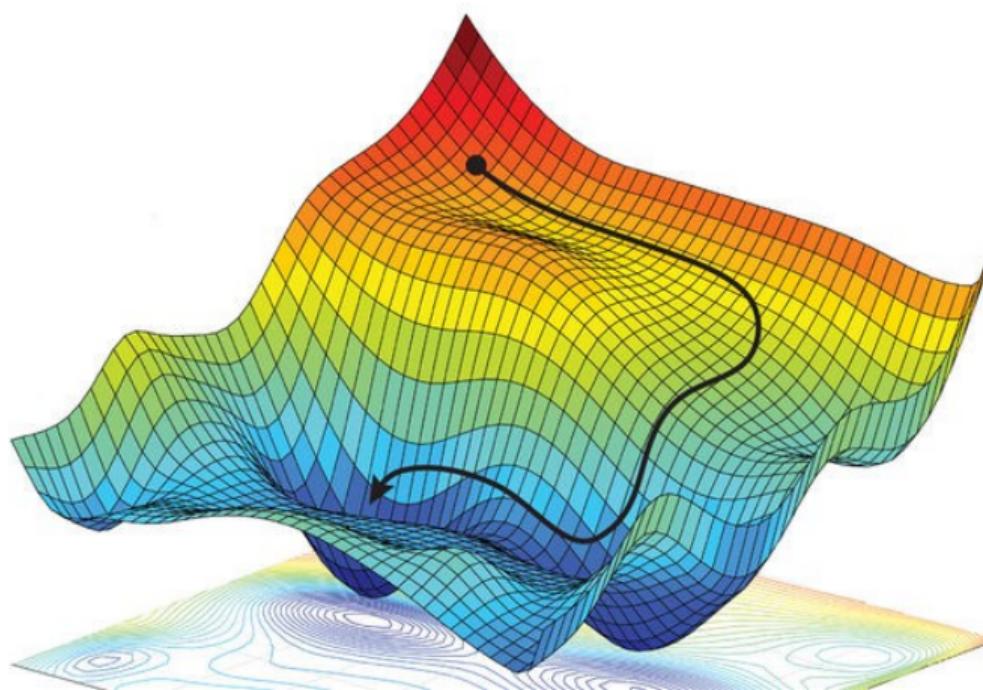
Preliminary

Problem Setup

Stein's  
Identity

Algorithm

References



Source: <https://www.sciencemag.org/news/2018/05/ai-researchers-allege-machine-learning-alchemy>

# What are Kernels?

SVGD

Justin  
Pauckert

Preliminary

Problem Setup

Stein's  
Identity

Algorithm

References

Let  $X \subseteq \mathbb{R}^d$  be an input space. A function  $k : X \times X \rightarrow \mathbb{R}$  is called a **kernel**, if there exists an inner product space  $(F, \langle \cdot, \cdot \rangle)$  and a function  $\varphi : X \rightarrow F$  which satisfies

$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle$$

for all  $x, x' \in X$ . This is used to find linear solutions for Problems in  $F$  which would not be linearly solvable in  $X$ .

# Example: Gaussian RBF Kernel

SVGD

Justin  
Pauckert

Preliminary

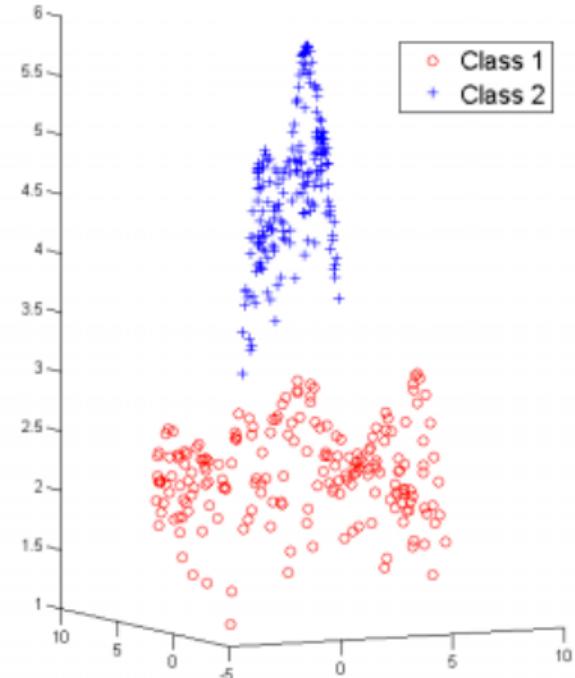
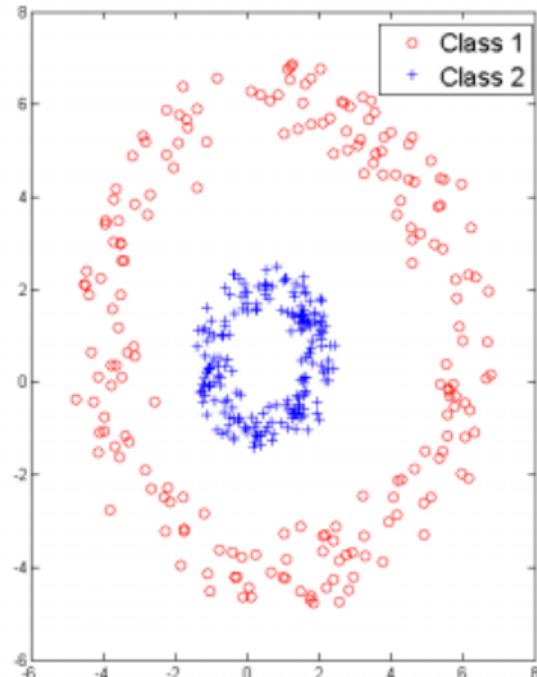
Problem Setup

Stein's  
Identity

Algorithm

References

$$k(x, x') = \exp\left(-\frac{\|x-x'\|_2^2}{h}\right) \text{ for some } h \in \mathbb{R}.$$



# What is KL divergence?

SVGD

Justin  
Pauckert

Preliminary

Problem Setup

Stein's  
Identity

Algorithm

References

The Kullback-Leibler (KL) divergence is a measure of how different one continuous probability distribution is from another.

$$\begin{aligned} \text{KL}(q||p) &:= \int_X p(x) \log \frac{q(x)}{p(x)} dx \\ &= \int_{x \sim q} \log \frac{q(x)}{p(x)} dx = \mathbb{E}_q[\log q(x) - \log p(x)] \end{aligned}$$

# Problem Setup

SVGD

Justin  
Pauckert

Preliminary

Problem Setup

Stein's  
Identity

Algorithm

References

Let  $X$  be a continuous random variable taking values in  $\mathcal{X} \subseteq \mathbb{R}^d$ ,  $\{D_k\}$  a set of  $N$  i.i.d. observations. Then with a prior distribution  $p_0(x)$  we have

$$p(x|\{D_k\}) = \frac{p_0(x) \prod_{k=1}^N p(D_k|x)}{\int p_0(x) \prod_{k=1}^N p(D_k|x) dx}.$$

# Stein's Identity

SVGD

Justin  
Pauckert

Preliminary

Problem Setup

Stein's  
Identity

Algorithm

References

Let  $p(x)$  be a smooth density on  $\mathcal{X}$ ,  $\phi(x) = [\phi_1(x), \dots, \phi_d(x)]^T$  a smooth vector function. Then if  $\phi$  is sufficiently regular we have

$$\mathbb{E}_p[\mathcal{A}_p\phi(x)] = 0, \tag{1}$$

$$\text{where } \mathcal{A}_p\phi(x) = \phi(x)\nabla_x \log p(x)^T + \nabla_x \phi(x).$$

$\mathcal{A}_p$  is called the **Stein Operator** and we say that  $\phi$  is in the Stein class of  $p$  if (1) holds.

# Stein Discrepancy

SVGD

Justin  
Pauckert

Preliminary

Problem Setup

Stein's  
Identity

Algorithm

References

Consider another smooth density  $q(x)$ . Then we get a discrepancy measure via

$$\mathbb{S}(q, p) = \max_{\substack{\phi \in \mathcal{H}^d \\ \|\phi\| \leq 1}} \{\mathbb{E}_q[\text{tr}(\mathcal{A}_p \phi(x))]^2\},$$

# Stein Discrepancy

SVGD

Justin  
Pauckert

Preliminary

Problem Setup

Stein's  
Identity

Algorithm

References

Consider another smooth density  $q(x)$ . Then we get a discrepancy measure via

$$\mathbb{S}(q, p) = \max_{\substack{\phi \in \mathcal{H}^d \\ \|\phi\| \leq 1}} \{\mathbb{E}_q[\text{tr}(\mathcal{A}_p \phi(x))]^2\},$$

$\mathcal{H}^d$  being the reproducing kernel Hilbert space (RKHS) for some kernel  $k$ . The optimal solution has been shown to be

$$\phi_{q,p}^*(\cdot) = \mathbb{E}_q[\mathcal{A}_p k(x, \cdot)],$$

normalized to  $\phi(x) = \phi_{q,p}^*(x)/\|\phi_{q,p}^*\|$ .

# The *variational* Part

SVGD

Justin  
Pauckert

Preliminary

Problem Setup

Stein's  
Identity

Algorithm

References

Given a set of distributions  $\{\mathcal{Q}\}$ , we want to approximate the target distribution  $p(x)$  by minimizing the KL divergence.

$$q^* = \min_{q \in \mathcal{Q}} \{\text{KL}(q||p)\}$$

# The *variational* Part

SVGD

Justin  
Pauckert

Preliminary

Problem Setup

Stein's  
Identity

Algorithm

References

Given a set of distributions  $\{\mathcal{Q}\}$ , we want to approximate the target distribution  $p(x)$  by minimizing the KL divergence.

$$q^* = \min_{q \in \mathcal{Q}} \{\text{KL}(q||p)\}$$

Focus on sets  $\mathcal{Q}$  consisting of distributions obtained by transforming a reference distribution. If  $T : \mathcal{X} \rightarrow \mathcal{X}$  is a transformation, consider  $x \sim q_0$  and  $z = T(x)$  with density

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

# Stein Operator and KL Divergence

SVGD

Justin  
Pauckert

Preliminary

Problem Setup

Stein's  
Identity

Algorithm

References

Let  $T(x) = x + \epsilon\phi(x)$  be a transformation,  $x \sim q(x)$  and  $z = T(x)$  with density  $q_{[T]}(z)$ . Then

$$\nabla_\epsilon \text{KL}(q_{[T]} || p) \Big|_{\epsilon=0} = -\mathbb{E}_q[\text{tr}(\mathcal{A}_p \phi(x))].$$

# Stein Operator and KL Divergence

SVGD

Justin  
Pauckert

Preliminary

Problem Setup

Stein's  
Identity

Algorithm

References

Let  $T(x) = x + \epsilon\phi(x)$  be a transformation,  $x \sim q(x)$  and  $z = T(x)$  with density  $q_{[T]}(z)$ . Then

$$\nabla_\epsilon \text{KL}(q_{[T]} || p) \Big|_{\epsilon=0} = -\mathbb{E}_q[\text{tr}(\mathcal{A}_p \phi(x))].$$

We know from earlier

$$\phi_{q,p}^* = \operatorname*{argmax}_{\substack{\phi \in \mathcal{H}^d \\ \|\phi\| \leq 1}} \{\mathbb{E}_q[\text{tr}(\mathcal{A}_p \phi(x))]^2\}.$$

# Stein Operator and KL Divergence

SVGD

Justin  
Pauckert

Preliminary

Problem Setup

Stein's  
Identity

Algorithm

References

Let  $T(x) = x + \epsilon\phi(x)$  be a transformation,  $x \sim q(x)$  and  $z = T(x)$  with density  $q_{[T]}(z)$ . Then

$$\nabla_\epsilon \text{KL}(q_{[T]} || p) \Big|_{\epsilon=0} = -\mathbb{E}_q[\text{tr}(\mathcal{A}_p \phi(x))].$$

We know from earlier

$$\phi_{q,p}^* = \underset{\substack{\phi \in \mathcal{H}^d \\ \|\phi\| \leq 1}}{\text{argmax}} \{ \mathbb{E}_q[\text{tr}(\mathcal{A}_p \phi(x))]^2 \}.$$

So for  $T(x) = x + f(x)$ ,  $f \in \mathcal{H}^d$  we have

$$\nabla_f \text{KL}(q_{[T]} || p) \Big|_{f=0} = -\phi_{q,p}^*(x).$$

# The Algorithm

$T^*(x) = x + \epsilon \cdot \phi_{q,p}^*(x)$  is equal to a step of functional gradient descend in RKHS.  
Additionally, we can express  $\phi_{q,p}^*$  with respect to our kernel  $k$ :

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

SVGD

Justin  
Pauckert

Preliminary

Problem Setup

Stein's  
Identity

Algorithm

References

# The Algorithm

SVGD

Justin  
Pauckert

Preliminary

Problem Setup

Stein's  
Identity

Algorithm

References

$T^*(x) = x + \epsilon \cdot \phi_{q,p}^*(x)$  is equal to a step of functional gradient descend in RKHS.  
Additionally, we can express  $\phi_{q,p}^*$  with respect to our kernel  $k$ :

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

---

## Algorithm Stein Variational Gradient Descend

---

**Input:** Target distribution density  $p(x)$ , set of initial particles  $\{x_i\}_{i=1}^n$ .

**Output:** Set of particles  $\{x_i\}_{i=1}^n$  that approximate the target distribution.

```
1: for  $\ell = 1$  to max_iterations do
2:   for  $i = 1$  to  $n$  do
3:      $x_i^{\ell+1} = x_i^\ell + \epsilon \cdot \hat{\phi}^*(x_i^\ell),$ 
       where  $\hat{\phi}^*(x) = \frac{1}{n} \sum_{j=1}^n (k(x_j^\ell, x) \nabla_{x_j^\ell} \log p(x_j^\ell) + \nabla_{x_j^\ell} k(x_j^\ell, x))$ 
4:   end for
5: end for
```

# Examples

SVGD

Justin  
Pauckert

Preliminary

Problem Setup

Stein's  
Identity

Algorithm

References

# Examples

SVGD

Justin  
Pauckert

Preliminary

Problem Setup

Stein's  
Identity

Algorithm

References

$$\hat{\phi}^*(x) = \frac{1}{n} \sum_{j=1}^n (k(x_j^\ell, x) \nabla_{x_j^\ell} \log p(x_j^\ell) + \nabla_{x_j^\ell} k(x_j^\ell, x))$$

# The Algorithm

SVGD

Justin  
Pauckert

Preliminary

Problem Setup

Stein's  
Identity

Algorithm

References

Because the update only depends on  $\nabla \log p(x)$  and

$$\nabla \log \frac{p(x)}{Z} = \nabla(\log p(x) - \log Z) = \nabla \log p(x),$$

we can completely ignore the normalization constant. Suddenly,

$$p(x|\{D_k\}) = p_o(x) \prod_{k=1}^N p(D_k|x)$$

is good enough as an estimator.

# References

SVGD

Justin  
Pauckert

Preliminary

Problem Setup

Stein's  
Identity

Algorithm

References



Liu, Qiang and Wang, Dilin (2016)

Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm

*arXiv* 1608.04471

[github.com/DartML/Stein-Variational-Gradient-Descend](https://github.com/DartML/Stein-Variational-Gradient-Descend)