Stein's method in machine learning

Vittorio Romaniello

November 28, 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

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

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 f is a measurable function in $\mathcal{E}^{f_n}$, the set of $\mathcal{E}$-measurable functions. Furthermore, let $\Gamma(\mathcal{Y}) \equiv \{f : E \to \mathcal{Y}\}$, a map $\mathcal{S}_\mathbb{P} : \mathcal{G} \subset \Gamma(\mathcal{Y}) \to \Gamma(\mathbb{R})$ is a Stein operator over a Stein class $\mathcal{G}$ if $\int_E \mathcal{S}_\mathbb{P}[f] d\mathbb{P} = 0 \ \forall f \in \mathcal{G}$ 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

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

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. 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 $\parallel \cdot \parallel$ as follows:

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

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 $\parallel \cdot \parallel_{\mathcal{H}}$ 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 $\parallel \boldsymbol{f} \parallel_{\mathcal{H}^d} = \sqrt{\sum_{i=1}^d \parallel f_i \parallel_{\mathcal{H}}^2}$ 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 $\parallel f \parallel_{\mathcal{H}} = \sqrt{\langle f, f \rangle_{\mathcal{H}}}$) if there exists a measurable function $k : E \times E \to \mathbb{R}$ with the following properties:

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

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 $\langle k(x, \cdot), k(x', \cdot)\rangle_{\mathcal{H}} = k(x, x')$

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 variational inference and deriving the main result of the paper. Next, we briefly describe optimal transport and show 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, $\boldsymbol{g}$, Equation (4) is defined as

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

where the trace is necessary to obtain a scalar value for $E_{\mathbb{Q}} \left[\mathcal{T}_p[\boldsymbol{g}](x) \right]$ which would otherwise be a $d \times d$ matrix. 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. $\| \boldsymbol{g} \|_{\mathcal{H}^d} \leq 1$, the discrepancy in Equation (5) becomes

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

and has a closed form solution given by $\boldsymbol{g}^* = \boldsymbol{\beta} / \| \boldsymbol{\beta} \|_{\mathcal{H}^d}$ where $\boldsymbol{\beta} = E_{\mathbb{Q}} \left[\mathcal{T}_p[k(x, \cdot)] \right]$. To prove this result we need to derive some intermediate results, these can be found in the appendix of [LLJ16], here we limit to summarising such results. Firstly, for a d-dimensional vector-valued function $\boldsymbol{g} \in \mathcal{T}_p[\mathcal{G}]$ it holds by definition that

$$E_{\mathbb{P}}[\mathcal{T}_p[\boldsymbol{g}](x)] = \boldsymbol{0}_{d \times d}, \ x \in \mathbb{R}^d$$

where $\boldsymbol{0}_{d \times d}$ indicates the $d \times d$ zero matrix. Under this assumption we can show that for $\boldsymbol{g} \in \mathcal{T}_p[\mathcal{G}]$ the following holds

$$E_{\mathbb{Q}}[\mathcal{T}_p[\boldsymbol{g}]] = E_{\mathbb{Q}}[\mathcal{T}_p[\boldsymbol{g}] - \mathcal{T}_q[\boldsymbol{g}]] = E_{\mathbb{Q}}[\langle \nabla \log p, \boldsymbol{g} \rangle_{\mathcal{H}^d} + \nabla \boldsymbol{g} - \langle \nabla \log q, \boldsymbol{g} \rangle_{\mathcal{H}^d} - \nabla \boldsymbol{g}] = E_{\mathbb{Q}}[\langle \nabla \log p - \nabla \log q, \boldsymbol{g} \rangle_{\mathcal{H}^d}]$$

where $\mathcal{T}_p[\boldsymbol{g}] = \langle \nabla \log p, \boldsymbol{g} \rangle_{\mathcal{H}^d} + \nabla \boldsymbol{g}$ as in Equation (3) but with vector-valued functions and q is the density of $\mathbb{Q}$ with respect to the Lebesgue measure.

A Exercises

Show RBF is in Stein's class

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).