Stein's method in machine learning

Vittorio Romaniello

November 24, 2019

1 Background

Stein's method, introduced by Stein et al. [Ste+72], 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 $(\Omega, \mathcal{H})$ be a measurable space and denote by $\mathcal{P}_\Omega$ the set of probability measures on $(\Omega, \mathcal{H})$. 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 $\mathcal{E} \subset \mathcal{H}$. Define $D : \mathcal{P}_E \times \mathcal{P}_E \to \mathbb{R}_+$ as

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

a measurable function that quantifies the discrepancy between two probability measures $\mathbb{Q}, \mathbb{P} \in \mathcal{P}_E$, where f is a measurable function in $\mathcal{E}_+^{f_n}$, the set of positive $\mathcal{E}$-measurable functions. Furthermore, let $\Gamma(\mathcal{Y}) \equiv \{f : E \to \mathcal{Y}\}$, a map $\mathcal{S}_{\mathbb{P}} : \mathcal{G} \subset \Gamma(\mathbb{R}^d) \to \Gamma(\mathbb{R})$ 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

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

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 and probability measure $\mathbb{Q}$ 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

2 Open questions and research directions

A Exercises

References

[Ste+72] C. Stein et al. "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).