Contributed Discussion

Harrison Zhu[*], Xing Liu[†], Alberto Caron[‡], Ioanna Manolopoulou[§],
Seth Flaxman[¶], and François-Xavier Briol[‖]

In the Neyman-Rubin causal model, patient i is represented through a triplet (X_i, Z_i, Y_i), where X_i denotes covariates, Z_i denotes whether the patient received a treatment $(Z_i = 1)$ or not $(Z_i = 0)$, and Y_i represents the outcome (in particular $Y_i(1)$ is when the patient received the treatment and $Y_i(0)$ if they did not). Given such data for n patients, we would like to answer questions about the effect of the treatment on the outcome variables. These questions can be answered by considering certain statistics of interest (Hill, 2011). These include the (population) average treatment effect (ATE), given by $\mathbb{E}[Y(1) - Y(0)]$, the sample average treatment effect (SATE), given by $\frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))$, the population average effect of the treatment on the treated (PATT), given by $\mathbb{E}[Y(1) - Y(0)|Z = 1]$, and the sample average treatment effect of the treatment on the treated (SATT), given by $\frac{1}{n}\sum_{i:Z_i=1}(Y_i(1) - Y_i(0))$.

An interesting point is that each of these quantities can be expressed as an integral of the conditional average treatment effect (CATE), $\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$, over some distribution on covariates. As discussed in Hahn et al. (2020), this can be estimated by integrating a model $\hat{\tau}$ of τ, such as a Bayesian posterior mean. This remark, although seemingly trivial, is particularly interesting since it opens up connections with the field of probabilistic numerics and especially Bayesian probabilistic numerical integration (BPNI) (Diaconis, 1988; O'Hagan, 1991; Rasmussen and Ghahramani, 2002; Briol et al., 2019).

In BPNI, the goal is to tackle challenging problems in numerical analysis, such as the computation of an intractable integral, using tools from Bayesian nonparametrics. The motivation is that the Bayesian framework can be used to quantify uncertainty over the value of the integral. This is done through three steps: (i) a prior is placed over the integrand, (ii) this prior is conditioned on values of the function to obtain a posterior on the integrand, (iii) this posterior on the integrand implies a (one-dimensional) posterior on the value of the integral. Different prior choices allow us to encode properties of the integrand, such as smoothness or periodicity, in a straightforward manner, leading to algorithms which respect these properties. The most common model is a Gaussian process (GP); although more recent work also considers alternatives such as Bayesian additive regression trees (BART) (Zhu et al., 2020) or multi-output Gaussian processes (Xi et al., 2018; Gessner et al., 2019).

[*] Department of Mathematics, Imperial College London, hbz15@ic.ac.uk

[†] Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, liuxing971015@outlook.com

[‡] Department of Statistical Science, University College London, alberto.caron.19@ucl.ac.uk

[§] Department of Statistical Science, University College London, i.manolopoulou@ucl.ac.uk

[¶] Department of Mathematics, Imperial College London, s.flaxman@imperial.ac.uk

[‖] Department of Statistical Science, University College London and The Alan Turing Institute, f.briol@ucl.ac.uk

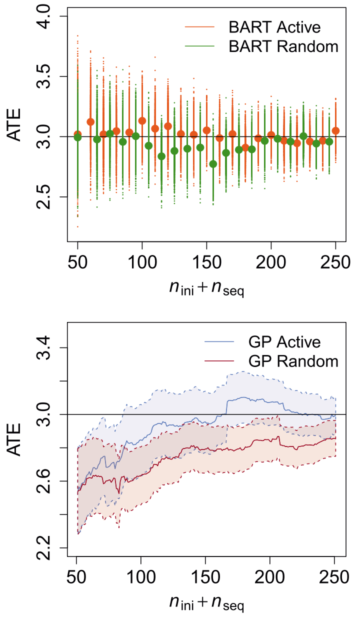

Figure 1: Estimates of the ATE for active learning and random sampling.

We can hence see the computation of ATE, PATT, SATE and SATT through integrals of the CATE as applications of BPNI to causal inference. This leads to several remarks:

1. There are a number of consistency results for BPNI methods, including refined convergence rates in a variety of scenarios depending on the model used, the domain of integration, the data generating process and the smoothness of the integrand. These could provide strong theoretical guarantees for the estimation of the ATE, SATE, SATT and PATT in a wide variety of settings; see Briol et al. (2019); Kanagawa and Hennig (2019); Kanagawa et al. (2020); Wynne et al. (2020).

2. **Active Learning:** The field of BPNI has derived a variety of experimental design schemes targeting directly the efficient approximation of integrals (rather than approximation of functions); see Osborne et al. (2012); Gunter et al. (2014); Briol et al. (2015); Jiang et al. (2019) in the case of Gaussian process models, and Zhu et al. (2020) for BART. Again, these could lead to more efficient estimates of the quantities of interest in causal inference.

To highlight the potential benefits of active learning schemes for causal inference with GPs and BART, we considered the synthetic example in Section 6.1 of Hahn et al. (2020) with homogeneous treatment effects and linear prognostic function.[1] We assume

[1] The code is available at `https://github.com/ImperialCollegeLondon/BART-Int`.

the model $Y_i = f(X_i, Z_i) + \epsilon_i$, where $\epsilon_i \sim N(0, 0.1^2)$ and the covariates X_i are i.i.d. with a known distribution Π. We consider the computation of the ATE estimated as $\Pi_m[\hat{f}(X,1) - \hat{f}(X,0)]$, where $\hat{f}$ is the fitted posterior of the Bayesian model (e.g. BART or GP) on f, and Π_m is an empirical distribution formed by samples $\{\tilde{x}_i\}_{i=1}^m$ representative of Π (and potentially different from the observed covariates $\{x_i\}_{i=1}^n$).

For the active learning algorithms (see Zhu et al., 2020 for full details), we use a candidate set of size $m = 2000$, then begin with $n_{\text{ini}} = 50$ initial random design points and acquire $n_{\text{seq}} = 200$ additional points. We use a sequential design with new point selected one at a time through the following objective: At iteration n, we select x_{n+1} and z_{n+1} as follows $\text{argmax}_{c=(x,z)} \mathbb{V}[f(x,z)\pi(x)e(z)|\{(x_i, z_i, y_i)\}_{i=1}^n]$, where π is the density of Π and e is the propensity function. We can see that active learning helped both models to obtain improved estimates of the true value of ATE (ATE $= 3$).

Overall, we conclude that the fields of causal inference and BPNI could benefit from further interactions. In this note, we pointed out how recent advances in BPNI could lead to further practical and theoretical advances in causal inference (including through a small synthetic example), but it is also clear that applications in causal inference could provide motivation for the development of novel BPNI algorithms.

References

Briol, F.-X., Oates, C. J., Girolami, M., and Osborne, M. A. (2015). "Frank-Wolfe Bayesian quadrature: Probabilistic integration with theoretical guarantees." In *Neural Information Processing Systems*, 1162–1170. 56

Briol, F.-X., Oates, C. J., Girolami, M., Osborne, M. A., and Sejdinovic, D. (2019). "Probabilistic integration: A role in statistical computation? (with discussion)." *Statistical Science*, 34(1): 1–22. MR3938958. doi: `https://doi.org/10.1214/18-STS660`. 55, 56

Diaconis, P. (1988). "Bayesian numerical analysis." *Statistical Decision Theory and Related Topics IV*, 163–175. MR0927099. 55

Gessner, A., Gonzalez, J., and Mahsereci, M. (2019). "Active multi-information source Bayesian quadrature." In *Uncertainty in Artificial Intelligence*. 55

Gunter, T., Garnett, R., Osborne, M., Hennig, P., and Roberts, S. (2014). "Sampling for inference in probabilistic models with fast Bayesian quadrature." In *Advances in Neural Information Processing Systems*, 2789–2797. 56

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). "Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects." *Bayesian Analysis*. 55, 56

Hill, J. L. (2011). "Bayesian nonparametric modeling for causal inference." *Journal of Computational and Graphical Statistics*, 20(1): 217–240. MR2816546. doi: `https://doi.org/10.1198/jcgs.2010.08162`. 55

Jiang, S., Chai, H., Gonzalez, J., and Garnett, R. (2019). "BINOCULARS for efficient, nonmyopic sequential experimental design." *arXiv:1909.04568*. 56

Kanagawa, M. and Hennig, P. (2019). "Convergence guarantees for adaptive Bayesian quadrature methods." In *Neural Information Processing Systems*, 6237–6248. 56

Kanagawa, M., Sriperumbudur, B. K., and Fukumizu, K. (2020). "Convergence analysis of deterministic kernel-based quadrature rules in misspecified settings." *Foundations of Computational Mathematics*, 20: 155–194. MR4056928. doi: https://doi.org/10.1007/s10208-018-09407-7. 56

O'Hagan, A. (1991). "Bayes–Hermite quadrature." *Journal of Statistical Planning and Inference*, 29: 245–260. MR1144171. doi: https://doi.org/10.1016/0378-3758(91)90002-V. 55

Osborne, M. A., Duvenaud, D., Garnett, R., Rasmussen, C. E., Roberts, S., and Ghahramani, Z. (2012). "Active learning of model evidence using Bayesian quadrature." In *Advances in Neural Information Processing Systems*, 46–54. 56

Rasmussen, C. and Ghahramani, Z. (2002). "Bayesian Monte Carlo." In *Advances in Neural Information Processing Systems*, 489–496. 55

Wynne, G., Briol, F.-X., and Girolami, M. (2020). "Convergence guarantees for Gaussian process means with misspecified likelihoods and smoothness." *arXiv:2001.10818*. 56

Xi, X., Briol, F.-X., and Girolami, M. (2018). "Bayesian quadrature for multiple related integrals." In *International Conference on Machine Learning*, 8533–8564. MR3577382. doi: https://doi.org/10.1214/16-BA1017A. 55

Zhu, H., Liu, X., Kang, R., Shen, Z., Flaxman, S., and Briol, F.-X. (2020). "Bayesian probabilistic numerical integration with tree-based models." *arXiv:2006.05371*. 55, 56, 57