



PROJECT MUSE®

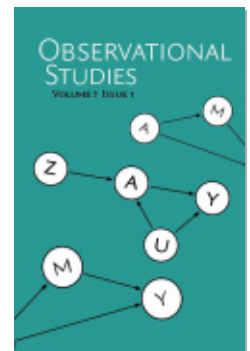
Comments on Leo Breiman's paper: "Statistical Modeling: The Two Cultures" (Statistical Science, 2001, 16(3), 199-231)

Jelena Bradic, Yinchu Zhu

Observational Studies, Volume 7, Issue 1, 2021, pp. 21-31 (Article)

Published by University of Pennsylvania Press

DOI: <https://doi.org/10.1353/obs.2021.0019>



➔ *For additional information about this article*

<https://muse.jhu.edu/article/799744>

Comments on Leo Breiman’s paper “Statistical Modeling: The Two Cultures” (Statistical Science, 2001, 16(3), 199-231)

Jelena Bradic

jbradic@ucsd.edu

*Department of Mathematics and Halicioglu Data Science Institute
University of California, San Diego
La Jolla, CA 92037, USA*

Yinchu Zhu

yinchuzhu@brandeis.edu

*Department of Economics
Brandeis University
Waltham, MA 02453, USA*

Abstract

Breiman challenged statisticians to think more broadly, to step into the unknown, model-free learning world, with him paving the way forward. Statistics community responded with slight optimism, some skepticism, and plenty of disbelief. Today, we are at the same crossroad anew. Faced with the enormous practical success of model-free, deep, and machine learning, we are naturally inclined to think that everything is resolved. A new frontier has emerged; the one where the role, impact, or stability of the *learning* algorithms is no longer measured by prediction quality, but an inferential one – asking the questions of *why* and *if* can no longer be safely ignored.

Keywords: robustness, causal inference, machine learning

1. Breiman was right

In his article “Statistical Modeling: The Two Cultures” (Breiman, 2001), Leo Breiman marveled at the possibility of empirically built models, trained solely to improve predictions. He argued for their potential impact on empirical applications. He advocated for a complete reversal of model-driven statistical work, the one that clumsily tries, often strenuously, to find the best or most appropriate model for a particular problem. Leo firmly believed that a new age was upon us. Age of models without models or that of algorithms tuned all so perfectly for a unique, individual, peculiar problem at hand. Looking back at it from today’s perspective, with deep learning dominating the success of algorithmically-driven science, we may wonder, how is it possible that the rest of the community failed to see it? Leo was a singular voice at the time; the rest surely and steadily continued the well-established statistical modeling path.

Breiman was a provocateur in the best possible terms. Without people like him, statistics would not be where it is today. One might argue that breakthroughs made, paradigms uncovered, premisses broken, only become possible when the well-established routes, like trenches, hard to remove, are challenged, deemed inappropriate, or invalid.

Models are well understood to be a poor, overly simplistic representation of nature, its complexity, and flexibility. However, models were believed to help approximate certain tasks

useful for nature; the quote of George E.P. Box, “all models are wrong but some are useful,” is cited to this day. The illusion of the success of model fitting was broken with a sequence of Peter Bickel’s seminar works, among others; e.g., Albers et al. (1976); Bickel et al. (1993). Bickel et al. (2006) established that goodness of fit tests have extremely low power unless the direction of the alternative is precisely specified. Sometimes the direction of the alternative was given by the metric implicitly used when constructing the tests. Such is the case of the one-sample Kolmogorov test for goodness of fit to the uniform $(0, 1)$ distribution. It is well known that it has power at a rate $n^{-1/2}$, notably only against alternatives where $|P(X \leq 1/2) - 1/2|$ is large. In the above, n denotes the sample size. The χ^2 tests with an increasing number of cells as $n \rightarrow \infty$, on the other hand, have trivial power in every direction at a $n^{-1/2}$ rate. As goodness of fit tests measure the usefulness of the developed models, these results implied impossibility in keeping up with the belief that all models are useful.

A new measure of success of the *fit* was needed, and Leo Breiman, in “Statistical Modeling: The Two Cultures,” argued, in a manner of speaking, for a new definition of a measure of goodness of fit – the one of predictive accuracy. A model that can predict well on a hold-out dataset was regarded as beneficial. In a way, the success of deep learning and the advent of over-parametrized neural networks are based precisely on this predictive accuracy that Leo advocated. In some communities, this measure of “usefulness” has nowadays overpowered all others. At the time of Breiman’s article, neural networks, although present as models, were not tested and used for predictive accuracy. We understand nowadays that it is only over-parametrized, overly-complex neural network designs, with many more parameters than samples, that are regarded as most powerful predictors; see Belkin et al. (2019). Twenty years ago, Leo simply advocated, in no weak terms, that Statistics needs to take this new challenge, step-up, and do more.

2. Statistics has since done more

Over the last two decades, statistics has stepped outside the model-driven molds and has since Leo’s call-outs done a lot more. Statistics has focused on achieving generalization, not through the construction of complicated models or theories, but simplification and model reduction. Among others, seminar works of (Tibshirani, 1996) on Lasso and (Fan and Li, 2001; Fan and Lv, 2008) on model selection, sparked a whole new area of interest. Since the beginning of the 21-st century, statistics has almost entirely moved away from the goodness of fit. Instead, prediction accuracy came into the front view. It has made strides in understanding theoretical aspects of random forests, (Scornet et al., 2015; Wager and Walther, 2015; Athey et al., 2019) and boosting (Bühlmann and Yu, 2003; Zhang and Yu, 2005), in understanding prediction accuracy from scratch by establishing non-asymptotic prediction error bounds (Bickel et al., 2009; Wainwright, 2019) and proposing new model-free procedures (Politis, 2013; Barber and Candès, 2015) as well as Bayesian random forest equivalents, such is Bayesian adaptive random trees (BART) (Chipman et al., 2010), among others. Intersections between algorithmic-driven learning and statistics can also be seen in new pathways in building the semi-supervised inferential tools by, for example, Lafferty and Wasserman (2007) as well as by the recently departed Larry Brown and co-authors in Zhang et al. (2019); Azriel et al. (2016), or others, such as Zhang and Bradic (2019) and

Cannings et al. (2020). Recent interests have led to building inferential (statistical) tools around algorithm-driven learning, see, e.g., works on conformal inference by Wasserman et al. (2020); Lei et al. (2018); Barber et al. (2021) for example. One can only hope that these are simple beginnings of a new statistical science era, driven and inspired by the interplay of model-free and model-driven learning methods.

3. Model-free and model-driven statistics: scrutinized and impugned

The practical success of machine and deep learning is perhaps the culmination of what Breiman advocated for in the article in question, Breiman (2001). Here, prediction accuracy is the sole driver of quality of success. This is, of course, exacerbated by a concurrence of both the availability of immense computing power as well by the access to datasets of previously unimaginable scale. While machine learning innovations were largely driven by Leo’s proposed prediction on a hold-out data, nowadays named generalization error, the deep learning’s steep success came only after a hold-out prediction fell back into a within-sample prediction. At least two paradoxes emerge when contrasting Leo’s work on random forests and the origins of deep learning success.

First, there is a sharp contrast between Leo’s work on random forests and deep learning methods in ways they re-use the data. Breiman’s work on random forests and the invention of bootstrap aggregation (bagging) was designed to avoid the pitfalls of re-using the same dataset. Bootstrapping the original data, combining trees drawn on different subsamples of the data, was a way of capturing different aspects of the same dataset – all with the intention of not reusing the same data instances repeatedly. Yet, neural networks and stochastic gradient descent (SGD) training do quite the opposite. Epochs are more related to permutations than to subsampling and bagging. With SGD, same instances of the dataset are needed, required, and re-used, usually over a hundred times.

Secondly, Leo was a firm advocate of model-free learning, but one could argue that neural networks are themselves quite the opposite, examples of extreme-model learning. Deep learning success can now theoretically also be prescribed to over-parametrized learning regimes, where the number of nodes and edges is required to be far greater than the number of samples; the number of layers seems secondary to the sheer number of the weight parameters (Ma et al., 2018; Belkin et al., 2020). With this view in mind, one has to ask whether neural networks are indeed model-free methods (Belkin et al., 2018)? In support of this hypothesis are perhaps the many papers illustrating close connections between neural networks with an infinite number of parameters and specialized, complex kernel ridge regression methods (Lee et al., 2019; Sohl-Dickstein et al., 2020).

It is clear, though, that neural-network learning is not model-driven learning and that over-parametrized neural networks achieve more than models do. However, it is unclear to what purpose. Prediction is no longer satisfactory and is not the only measure of success in practice. Stability, reproducibility, and inference, require more than mere control of the prediction error (Yu, 2013; Yu and Kumbier, 2020). For example, inferential tasks related to the average treatment effect are compelling and theoretically amenable to the prediction accuracy lens. Nevertheless, the impact of machine-learning methods on such tasks has yet to be unlocked.

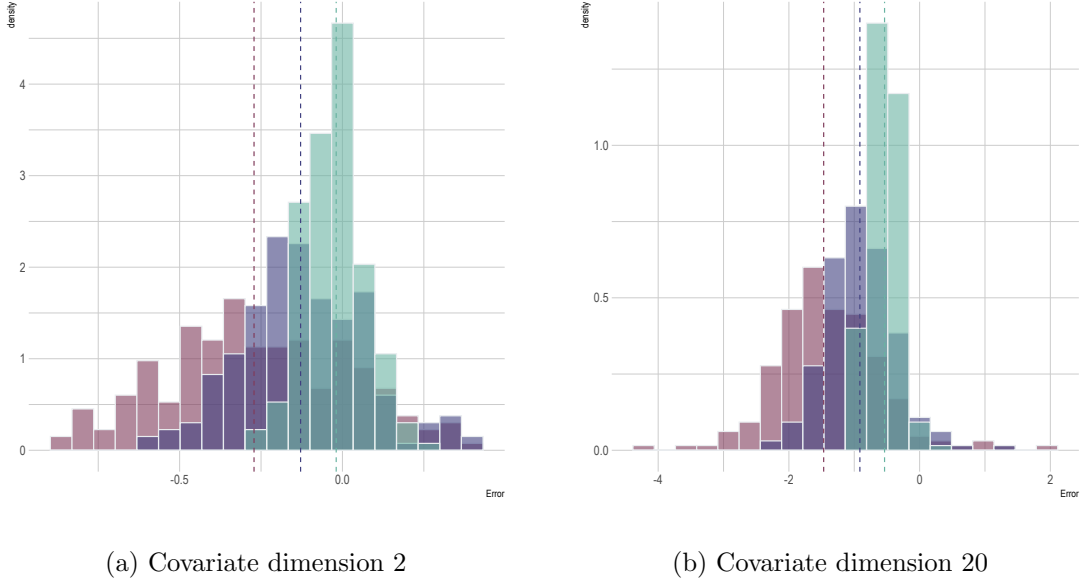


Figure 1: Histogram estimation error of 200 repeated cross-fitted Doubly Robust Average Treatment Effect estimator with twenty covariates across three sample sizes: $n = 1000$ (pink), $n = 2000$ (blue) and $n = 6000$ (green). Dashed lines represent the corresponding medians. Outcome and propensity models follow Example 1.

We discuss the average treatment effect estimation and the (unknown) impact of random forest estimates for illustrative purposes. For brevity, we focus on a simple(st) setting of independent and identically distributed observations, receiving a binary treatment under the assumption that the treatment satisfies some form of exogeneity (Rosenbaum and Rubin, 1983). Different forms of this assumption are referred to as unconfoundedness. Consider a setting with n units, for whom we observe outcomes $\{Y_i\}_{i=1}^n \in \mathbb{R}$. There is a binary treatment that varies by units, denoted with $D_i \in \{0, 1\}$ and a pair of potential outcomes $Y_i(0)$ and $Y_i(1)$ for all units. In this way, the realized or observed outcome can be represented as $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$. For each unit, we also observe a vector of potential confounders $X_i \in \mathbb{R}^p$. We are interested in estimating the average treatment effect $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$, which becomes identifiable under an additional overlap assumption $\mathbb{P}(D = 1|X) \in (0, 1)$. As we notice, for each unit i we only observe either $Y_i(0)$ or $Y_i(1)$, making the estimation of τ challenging. There are many possible ways to estimate τ , but a double robust or augmented inverse probability weighting (AIPW) estimate stands out (Robins et al., 1994). It is semi-parametrically optimal (Hahn, 1998) and asymptotically normal when either the outcome model or the treatment assignment model is correctly specified (Robins and Rotnitzky, 1995) – property also named model double-robust. There is a broad consensus that double-robustness is well understood in under-parametrized settings (Babino et al., 2019). With the work of Chernozhukov et al. (2018) double-robustness was extended to over-parametrized models, but the effect of machine-learning methods, although speculated and partially articulated, was not fully described. General conditions

such as “product-rate-condition” implied a possible validity of machine learning methods in the context of double-robust estimates.

However, upon further inspection, we observe that current theoretical understandings of random forests do not extend to the “product-rate-condition.” The slow rate of convergence of the forests might indicate that a random forest estimate of the outcome and the treatment assignment might fail in satisfying this condition. We performed two simple simulation experiments to explore the impact of sample size, parameter size, and the random forest itself on the ATE estimation. Example 1 corresponds to Heterogeneous confounding with features drawn from a mixture of multivariate normal distributions with covariances being identity and Toeplitz with correlation off-diagonal being $\rho = -0.5$ and mixing probability 0.7. Example 2 corresponds to a simple, high-dimensional, and sparse setting. Both examples are described below.

[Example 1.] Heterogeneous confounding

$$\begin{aligned} \text{Outcome: } \mathbb{E}[Y_i(a)|X_i] &= X_i^\top \beta + \log(|X_i^\top \delta|) + a, \\ \text{where } \beta &= (1, 1, \dots, 1)^\top \text{ and } \delta = 2\beta. \\ \text{Propensity: } \text{logit} \left[\mathbb{P}(D_i = 1|X_i) \right] &= \alpha X_i^\top \theta_1 + (1 - \alpha) X_i^\top \theta_2 \\ \text{where } \theta_1 &= (1, 1, \dots, 1)^\top, \theta_2 = \theta_1/2 \text{ and } \alpha = 0.8. \end{aligned}$$

[Example 2.] High-dimensional and sparse confounding

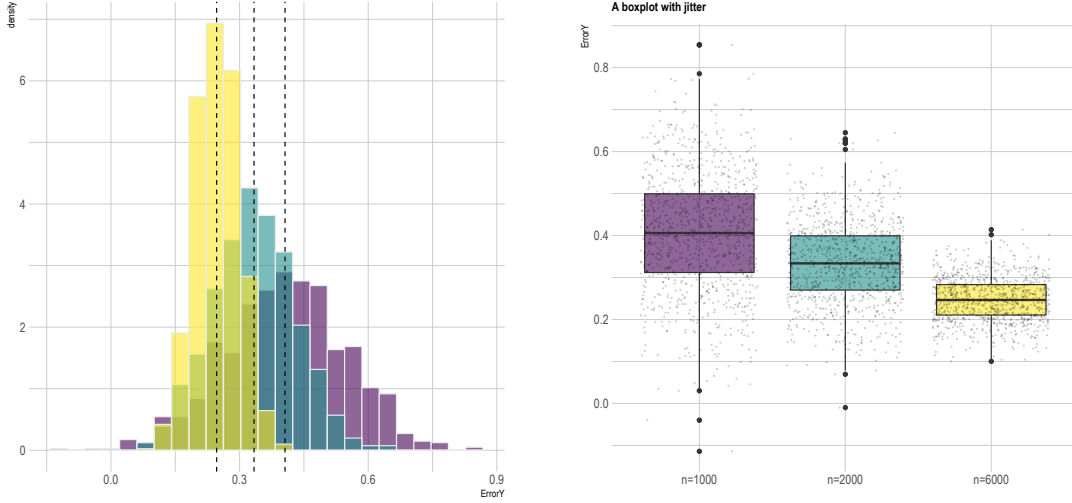
$$\begin{aligned} \text{Outcome: } \mathbb{E}[Y_i(a)|X_i] &= X_i^\top \beta_a, \\ \text{where } \beta_0 &= (1, 0, 1, 0, \dots, 0)^\top, \text{ and } \beta_1 = (1, 0, 0, 1, 0, \dots, 0)^\top. \\ \text{Propensity: } \text{logit} \left[\mathbb{P}(D_i = 1|X_i) \right] &= X_i^\top \gamma, \\ \text{where } \gamma &= (1, 1, 0, \dots, 0)^\top. \end{aligned}$$

In the above $\text{logit}(q) = q/(1 - q)$ and the dimensions of the features in both Examples varies from 2, 20 to 200, respectively. In Example 1 treatment assignment follows a mixture model with both models being logistic and the mixture probability being 0.8.

We implement a random-forest version of the cross-fitted augmented inverse propensity score estimator. The outcome and propensities are estimated in a single sample, I_1 then evaluated on a separate sample, I_2 , and averaged. Honest trees were used, whereas tuning parameters of the random forest were trained using cross-validation. Let $\widehat{\mu}_a(\cdot)$ and $\widehat{\pi}_a(\cdot)$ denote the estimated random forests corresponding to the potential outcome models $\mathbb{E}[Y_i(a)|X_i = x] = \mu_a(x)$ as well as the treatment assignment model $\widehat{\mathbb{P}}[D_i = a|X_i = x] = \pi_a(x)$. The AIPW cross-fitted estimate is then neatly represented as

$$\hat{\tau} = \frac{1}{|I_2|} \sum_{i \in I_2} \left(\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i) + \frac{D_i - \widehat{\pi}_1(X_i)}{\widehat{\pi}_1(X_i)(1 - \widehat{\pi}_1(X_i))} (Y_i - \widehat{\pi}_{D_i}(X_i)) \right).$$

We varied the sample size from $n = 1000$ to $n = 2000$ and $n = 6000$. Plots highlighting the histograms and boxplots of the error $\hat{\tau} - 1$ of the estimated average treatment effect are presented in Figures 1 and 2. Example 1 is presented in Figures 1a and 1b while Example 2 in Figures 2a and 2b.



(a) Histogram of estimation error

(b) Boxplots of estimation errors

Figure 2: Estimation error of 1000 repeated cross-fitted Doubly Robust Average Treatment Effect estimator with 200 covariates across three sample sizes: $n = 1000$ (purple), $n = 2000$ (blue) and $n = 6000$ (yellow). Dashed lines represent the corresponding medians. Outcome and propensity models are 2-sparse linear and logistic model, respectively.

The results are clear. Random forest, doubly-robust estimate fails to cover the true effect in almost all instances. Average coverage corresponding to Example 1 and Example 2 are presented in Table 1. In low-dimensional problems, with $p = 2$ the case of large sample size $n = 6000$ gets it close to covering but even with $p = 20$ the coverage is far from 95%. For covariate dimension 200, we see an awkward false concentration below zero: the estimate is getting more sure, less variable but at the wrong center.

		$n = 1000$	$n = 2000$	$n = 6000$
Example 1	$p = 2$	83.5	89.5	94
	$p = 20$	56.6	58.5	32.5
Example 2	$p = 200$	18.5	07	0.1

Table 1: Coverage of 95% confidence intervals.

The reasons for these shortcomings are unknown at the moment. We need to further our knowledge of model-free learning' effects beyond their sole predictive accuracy.

4. The next frontier: theory *with* practice instead of theory *vs.* practice

The new age of Data Science is upon us. With it comes the new challenge of addressing the questions of *why* and *if* a scientific or practical phenomenon has been discovered. This, in turn, requires new standards, formulations, definitions aimed at addressing the fundamental questions of whether there exists a phenomenon to be discovered, why the discovery was made, who influenced it, will it change drastically if we were to have observed somewhat different, distorted or data that has been intervened upon. Although models are known to be a poor representation of nature, we are now faced with questions whether our current, well-established, and natural go-to definitions are an excessively simplified depiction of practice, of the type of questions that might influence the domain of applications significantly. One might wonder if we now have to think of advancing the questions rather than models so that they advantageously drive the science?

Moreover, instinctive theoretical principles are no longer valid. Nonparametrics, an area perhaps the closest to model-free learning, traditionally utilizes different regularization methods to stabilize the estimators. Paradoxically, Tikhonov regularization (Tikhonov, 1943) characterizes much of the early nonparametrics works (Hoerl and Kennard, 1970), and it also characterizes current theoretical underpinnings of deep neural networks; e.g., Jacot et al. (2018); Arora et al. (2019). More broadly, regularization has been one of the threads underlying and connecting various research areas of statistics in the past two decades. We have made strong strides in understanding it, using it, designing it. We’ve successfully designed inferential tasks on the shoulders of those findings despite the negative impact of regularization, that is, the bias it affects; e.g., Van de Geer et al. (2014); Zhu and Bradic (2018). Regularization, albeit more implicit, is ever-present in deep learning. Much of the practical success of neural networks is attributed to various effects of the regularization. Suddenly the effect of regularization is multi-faceted: parameter, architecture, batch normalization, gradient descent, and more (Neyshabur et al., 2017; Razin and Cohen, 2020). This begs the question of whether the established notion of regularization is perhaps too wide to be useful?

Leo advocated strongly in favor of model-free learning. One of the natural bridges between model-free and model-driven learning lies perhaps in the notion of model misspecification. However, we now understand that model misspecification manifests itself differently in under- and over-parametrized settings (Bradic et al., 2019a). Classical notions reflect misspecification concerning parametrization structure, but we now understand that misspecification in the number of parameters is also possible. Fundamental limits of these impacts are largely unknown, although some progress has been made, for example, in Bradic et al. (2018), Bradic et al. (2019b) or Cai and Guo (2018) where robustness to sparsity is studied. Since the effect of the classical definition changes, new definitions are needed even for model misspecification. We see that robustness guarantees depend on specific structures which we do not understand well yet. Inferential tasks inadvertently suffer because of it, and we still do not understand the reason behind it all.

Lastly, new formulations are needed to better understand the new data world surrounding us, which uses the data to make decisions affecting millions of people, benefit science, and push the boundaries of the existing domains. Perhaps it is again the time to lean on Leo’s firm conviction that breakthroughs happen in conjunction with the practical appli-

cations and not against them. It is the time to listen and interact with practice, learn how to ask better questions, not only provide a better fitting but bridge the gap between theoretical goals and their purpose, usefulness, and scientific discovery.

Acknowledgments

We would like to acknowledge support of NSF DMS award number #1712481.

References

- Willem Albers, Peter J Bickel, and Willem R van Zwet. Asymptotic expansions for the power of distribution free tests in the one-sample problem. *The Annals of Statistics*, pages 108–156, 1976.
- Sanjeev Arora, Simon S Du, Zhiyuan Li, Ruslan Salakhutdinov, Ruosong Wang, and Dingli Yu. Harnessing the power of infinitely wide deep nets on small-data tasks. *arXiv preprint arXiv:1910.01663*, 2019.
- Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *Annals of Statistics*, 47(2):1148–1178, 2019.
- David Azriel, Lawrence D Brown, Michael Sklar, Richard Berk, Andreas Buja, and Linda Zhao. Semi-supervised linear regression. *arXiv preprint arXiv:1612.02391*, 2016.
- Lucia Babino, Andrea Rotnitzky, and James Robins. Multiple robust estimation of marginal structural mean models for unconstrained outcomes. *Biometrics*, 75(1):90–99, 2019.
- Rina Foygel Barber and Emmanuel J Candès. Controlling the false discovery rate via knockoffs. *Annals of Statistics*, 43(5):2055–2085, 2015.
- Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.
- Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pages 541–549. PMLR, 2018.
- Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.
- Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.
- Peter J Bickel, Ya’acov Ritov, and Thomas M. Stoker. Tailor-made tests for goodness of fit to semiparametric hypotheses. *The Annals of Statistics*, 34(2):721–741, 2006.
- Peter J. Bickel, Ya’acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of statistics*, 37(4):1705–1732, 2009.

- PJ Bickel, CAJ Klaassen, Y Ritov, and JA Wellner. Efficient and adaptive inference in semiparametric models, 1993.
- Jelena Bradic, Jianqing Fan, and Yinchu Zhu. Testability of high-dimensional linear models with non-sparse structures. *arXiv preprint arXiv:1802.09117*, 2018.
- Jelena Bradic, Victor Chernozhukov, Whitney K Newey, and Yinchu Zhu. Minimax semi-parametric learning with approximate sparsity. *arXiv preprint arXiv:1912.12213*, 2019a.
- Jelena Bradic, Stefan Wager, and Yinchu Zhu. Sparsity double robust inference of average treatment effects. *arXiv preprint arXiv:1905.00744*, 2019b.
- Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.
- Peter Bühlmann and Bin Yu. Boosting with the l_2 loss: regression and classification. *Journal of the American Statistical Association*, 98(462):324–339, 2003.
- T Tony Cai and Zijian Guo. Accuracy assessment for high-dimensional linear regression. *Annals of Statistics*, 46(4):1807–1836, 2018.
- Timothy I Cannings, Thomas B Berrett, and Richard J Samworth. Local nearest neighbour classification with applications to semi-supervised learning. *Annals of Statistics*, 48(3):1789–1814, 2020.
- Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.
- Hugh A Chipman, Edward I George, and Robert E McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.
- Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.
- Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.
- Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.
- Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.
- Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.
- John Lafferty and Larry Wasserman. Statistical analysis of semi-supervised regression. 2007.

- Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *arXiv preprint arXiv:1902.06720*, 2019.
- Jing Lei, Max G’Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.
- Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pages 3325–3334. PMLR, 2018.
- Behnam Neyshabur, Ryota Tomioka, Ruslan Salakhutdinov, and Nathan Srebro. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071*, 2017.
- Dimitris N Politis. Model-free model-fitting and predictive distributions. *Test*, 22(2):183–221, 2013.
- Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. *arXiv preprint arXiv:2005.06398*, 2020.
- James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.
- James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.
- Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.
- Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741, 2015.
- Jascha Sohl-Dickstein, Roman Novak, Samuel S Schoenholz, and Jaehoon Lee. On the infinite width limit of neural networks with a standard parameterization. *arXiv preprint arXiv:2001.07301*, 2020.
- Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.
- Andrey Nikolayevich Tikhonov. On the stability of inverse problems. In *Dokl. Akad. Nauk SSSR*, volume 39, pages 195–198, 1943.
- Sara Van de Geer, Peter Bühlmann, Ya’acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202, 2014.

- Stefan Wager and Guenther Walther. Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*, 2015.
- Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.
- Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.
- Bin Yu. Stability. *Bernoulli*, 19(4):1484–1500, 2013.
- Bin Yu and Karl Kumbier. Veridical data science. *Proceedings of the National Academy of Sciences*, 117(8):3920–3929, 2020.
- Anru Zhang, Lawrence D Brown, and T Tony Cai. Semi-supervised inference: General theory and estimation of means. *Annals of Statistics*, 47(5):2538–2566, 2019.
- Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4):1538–1579, 2005.
- Yuqian Zhang and Jelena Bradic. High-dimensional semi-supervised learning: in search for optimal inference of the mean. *to appear in Biometrika*, 2019.
- Yinchu Zhu and Jelena Bradic. Significance testing in non-sparse high-dimensional linear models. *Electronic Journal of Statistics*, 12(2):3312–3364, 2018.