Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)
Sister Conferences Best Papers Track

Statistical Learning with a Nuisance Component (Extended Abstract)*

Dylan J. Foster[1] and **Vasilis Syrgkanis**[2]
[1]Massachusetts Institute of Technology
[2]Microsoft Research, New England
dylanf@mit.edu, vasy@microsoft.com

Abstract

We provide excess risk guarantees for statistical learning in a setting where the population risk with respect to which we evaluate a target parameter depends on an unknown parameter that must be estimated from data (a "nuisance parameter"). We analyze a two-stage sample splitting meta-algorithm that takes as input two arbitrary estimation algorithms: one for the target parameter and one for the nuisance parameter. We show that if the population risk satisfies a condition called *Neyman orthogonality*, the impact of the nuisance estimation error on the excess risk bound achieved by the meta-algorithm is of second order. Our theorem is agnostic to the particular algorithms used for the target and nuisance and only makes an assumption on their individual performance. This enables the use of a plethora of existing results from statistical learning and machine learning literature to give new guarantees for learning with a nuisance component. Moreover, by focusing on excess risk rather than parameter estimation, we can give guarantees under weaker assumptions than in previous works and accommodate the case where the target parameter belongs to a complex nonparametric class. We characterize conditions on the metric entropy such that *oracle rates*—rates of the same order as if we knew the nuisance parameter—are achieved. We also analyze the rates achieved by specific estimation algorithms such as variance-penalized empirical risk minimization, neural network estimation and sparse high-dimensional linear model estimation. We highlight the applicability of our results in four settings of central importance in the literature: 1) heterogeneous treatment effect estimation, 2) offline policy optimization, 3) domain adaptation, and 4) learning with missing data.

*This is an extended abstract for our COLT 2019 paper [Foster and Syrgkanis, 2019].

1 Introduction

Predictive models based on modern machine learning methods are becoming increasingly widespread in policy making, with applications in health care, education, law enforcement and business decision-making. Most problems that arise in policy making are not pure prediction problems, but rather have a causal nature, such as attempting to predict counterfactual outcomes for different interventions or optimizing policies over such interventions. It is important to address the causal nature of these problems and build models that have a causal interpretation.

A common paradigm in the search of causality is that to estimate a model with a causal interpretation from observational data—e.g. data not collected via some randomized trial or via a known treatment policy—one typically needs to estimate many other quantities that are not of primary interest, but that can be used to de-bias a purely predictive ML model by formulating an appropriate loss. Examples of such *nuisance parameters* include the propensity for taking an action under the current policy, which can be used to form unbiased estimates for the reward of new policies, but is typically unknown in datasets that do not come from controlled experiments.

To make matters more concrete, let us walk through an example for which certain variants have been well-studied in machine learning (e.g., [Dudík *et al.*, 2011; Swaminathan and Joachims, 2015; Nie and Wager, 2017; Kallus and Zhou, 2018]). Suppose a decision maker wants to estimate the causal effect of some treatment $T \in \{0, 1\}$ on an outcome Y as a function of a set of observable features X; the causal effect will be denoted as $\theta(X)$. Typically, one has access to data consisting of tuples (X_i, T_i, Y_i), where X_i is the observed feature for sample i, T_i is the treatment taken, and Y_i is the observed outcome. Such settings are often referred to as having *bandit* feedback, since we only observe the outcome for the treatment that was chosen. Due to the bandit nature of the problem, one needs to create unbiased estimates of the unobserved outcome. A standard approach is to use the so-called *doubly-robust* formula, which is a combination of direct regression and inverse propensity scoring: if we let $Y_i^{(t)}$ denote the potential outcome from treatment t in sample i, and let $m_0^{(t)}(x_i) := \mathbb{E}\left[Y_i^{(t)} \mid x_i\right]$ and $p_0^{(t)} := \mathbb{E}[1\{T = t\} \mid x_i]$, then the following is an unbiased

estimator for each potential outcome:

$$\hat{Y}_i^{(t)} = m_0^{(t)}(x_i) + \frac{(Y_i - m_0^{(t)}(x_i))\,1\{T_i=t\}}{p_0^{(t)}(x_i)}. \quad (1)$$

Given such an estimator, we can estimate the treatment effect by running a regression between the unbiased estimates and the features, i.e. solve $\min_{\theta\in\Theta_n} \sum_i \left(\hat{Y}^{(1)} - \hat{Y}^{(0)} - \theta(X_i)\right)^2$ over some model class Θ_n. In the population limit, with infinite samples, this corresponds to finding a model $\theta(x)$ that minimizes the population risk $\mathbb{E}\Big[(\hat{Y}_i^{(1)} - \hat{Y}_i^{(0)} - \theta(X))^2\Big]$. Similarly, if one is interested in policy optimization rather than estimating treatment effects, they could use these unbiased estimates to solve $\min_{\theta\in\Theta_n} \sum_i (\hat{Y}_i^{(0)} - \hat{Y}_i^{(1)}) \cdot \theta(X_i)$ over a policy space Θ_n of functions mapping features to $\{0, 1\}$. When dealing with observational data, the functions m_0 and p_0 are not known, and must be estimated if we wish to evaluate the proxy labels $\hat{Y}^{(t)}$.The goal of the learner is to find a model θ that achieves good population risk when evaluated at the true nuisance functions as opposed to the estimated, since only then does the model have a causal interpretation.

This phenomenon is ubiquitous in causal inference and motivates us to formulate the abstract problem of statistical learning with a nuisance component: Given n i.i.d. examples from a distribution $\mathcal{D}$, a learner is interested in finding a *target parameter* $\widehat{\theta}_n \in \Theta_n$ so as to minimize a population risk function $L_{\mathcal{D}} : \Theta_n \times \mathcal{G}_n \to \mathbb{R}$. The population risk depends not just on the target parameter, but also on a *nuisance parameter* whose true value $g_0 \in \mathcal{G}_n$ is unknown to the learner. The goal of the learner is to produce an estimate that has small *excess risk* when evaluated at the unknown true nuisance model:

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - \inf_{\theta\in\Theta_n} L_{\mathcal{D}}(\theta, g_0). \quad (2)$$

Depending on the application, such an excess risk bound can take different interpretations. For many settings, such as treatment effect estimation, it is closely related to mean squared error, while in policy optimization problems it may correspond to regret. Following the tradition of statistical learning theory [Vapnik, 1995; Bousquet *et al.*, 2004], we make excess risk the primary focus of our work, independent of the interpretation. We develop algorithms and analysis tools that generically address (2), then apply these tools to a number of applications of interest.

The problem of statistical learning with a nuisance component is strongly connected to the problem of semiparametric inference [Robinson, 1988; Kosorok, 2008], where a true parameter θ_0 is the minimizer of a population risk that depends on unknown nuisance components. Our paper builds on a growing body of results on "double" or "debiased" machine learning in statistics and econometrics literature [Chernozhukov *et al.*, 2017; Chernozhukov *et al.*, 2018a; Chernozhukov *et al.*, 2018c; Chernozhukov *et al.*, 2018b] for addressing semiparametric inference problems. This line of research has focused on providing so-called "$\sqrt{n}$-consistent and asymptotically normal" estimates when the target parameter θ_0 is low-dimensional and nuisance parameters belong to a nonparametric class. Unlike the semiparametric inference problem, statistical learning with a nuisance component does not require a well-specified model, nor a unique minimizer of the population risk. Moreover, we do not ask for parameter recovery and asymptotic inference (i.e. asymptotically valid confidence intervals). Rather, we are content with an excess risk bound, regardless of whether there is an underlying true parameter to be identified. As a consequence, we provide guarantees even when the target parameter belongs to a large, potentially nonparametric class.

Algorithm 1 Two-Stage Estimation with Sample Splitting

Input: Sample set $S = z_1, \ldots, z_n$

1: Split S into subsets $S^{(1)} = z_1, \ldots, z_{\lfloor n/2 \rfloor}$, $S^{(2)} = S \setminus S^{(1)}$.

2: Let $\widehat{g}_n$ be the output of $\mathrm{Alg}(\mathcal{G}_n, S^{(1)})$.

3: **return** $\widehat{\theta}_n$, the output of $\mathrm{Alg}(\Theta_n, S^{(2)}\,; \widehat{g}_n)$.

The case where the target parameter belongs to an arbitrary class has not been addressed at the level of generality we consider in the present work, but we mention some prior work that goes beyond the low-dimensional/parametric setup for special cases. [Athey and Wager, 2017] and [Zhou *et al.*, 2018] give guarantees based on metric entropy of the target class for the specific problem of treatment policy learning. For estimation of treatment effects, various nonparametric classes have been used for the target class on a fairly cases by case basis, including kernels [Nie and Wager, 2017], random forests [Athey *et al.*, 2019; Oprescu *et al.*, 2019; Friedberg *et al.*, 2018], and high-dimensional linear models [Chernozhukov *et al.*, 2017; Chernozhukov *et al.*, 2018b]. Our work unifies several of these papers into a single framework and our general results have implications and improve upon each of these directions.

Our approach is to reduce the problem of statistical learning with a nuisance component to the standard formulation of statistical learning. Rather than directly analyzing particular algorithms and models from machine learning, such as regularized regression, gradient boosting, or neural network estimation, we assume a black-box guarantee for the excess in the case where a nuisance value $g \in \mathcal{G}_n$ is known. In particular, our main theorem only asks for the existence of an algorithm $\mathrm{Alg}(\Theta_n, S\,;\, g)$ that for any given nuisance parameter g and data set S, achieves good excess risk with respect to the population risk $L_D(\theta, g)$, i.e. with probability $1 - \delta$:

$$L_{\mathcal{D}}(\widehat{\theta}_n, g) - \inf_{\theta\in\Theta_n} L_{\mathcal{D}}(\theta, g) \le \mathrm{Rate}_{\mathcal{D}}(\Theta_n, S, \delta\,;\, g). \quad (3)$$

Likewise, we assume the existence of a black-box algorithm $\mathrm{Alg}(\mathcal{G}_n, S)$ to estimate the nuisance component g_0 from the data, with the required estimation guarantee varying from problem to problem.

Given access to the two black-box algorithms, we analyze a sample-splitting based simple meta-algorithm for statistical learning with a nuisance component presented as Algorithm 1. We can now state the main question addressed in this paper: *When is the excess risk achieved by sample splitting robust to nuisance component estimation error?*

In more technical terms, we seek to understand when the two-stage sample splitting estimation algorithm achieves an excess risk bound with respect to g_0, in spite of error in the estimator $\widehat{g}_n$ output by the first-stage algorithm. Robustness

to nuisance estimation error allows the learner to use more complex models for nuisance estimation and—under certain conditions on the complexity of the target and nuisance model classes—to learn a target parameter whose error is, up to lower order terms, as good as if the learner had known the true nuisance model. Such a guarantee is typically referred to as achieving an *oracle rate* in semiparametric inference.

1.1 Overview of Results

We show that *Neyman orthogonality*, which has been used to prove oracle rates for inference in semiparametric models [Neyman, 1959; Neyman, 1979; Chernozhukov *et al.*, 2018a; Chernozhukov *et al.*, 2018b], is key to providing oracle rates for statistical learning with a nuisance component. We prove that if the population risk satisfies a functional analogue of Neyman orthogonality, then the estimation error of $\widehat{g}_n$ has a second order impact on the overall excess risk (relative to g_0) achieved by $\widehat{\theta}_n$. To gain some intuition, Neyman orthogonality is weaker condition than double robustness, albeit similar in flavor, (see e.g. [Chernozhukov *et al.*, 2016]) and is satisfied by both the treatment effect loss and the policy learning loss described in the introduction. In more detail, our extension of the Neyman orthogonality condition asks that a cross-functional derivative of the loss vanish to zero, when evaluated at the optimal target and nuisance parameter. Prior work on the classical notion of Neyman orthogonality provides a number of means through which to construct orthogonal losses whenever certain moment conditions are satisfied by the data generating process [Chernozhukov *et al.*, 2018a; Chernozhukov *et al.*, 2016; Chernozhukov *et al.*, 2018b]. Indeed, orthogonal losses can be constructed in settings including treatment effect estimation, policy learning, missing data problems, estimation of structural econometric problems and game theoretic models.

We identify two regimes of excess risk behavior:

1. When the population risk is strongly convex with respect to the prediction of the target model (e.g. the treatment effect estimation loss), then typically so-called *fast rates* (e.g. rates of order of $O(1/n)$ for parametric classes) are achievable had we known the true nuisance model. Letting $R_{\mathcal{G}_n}$ denote the estimation error of the nuisance component (root-mean-squared prediction error for most of our settings), then in the fast rate setting we show that orthogonality implies that the first stage error has an impact on the excess risk of the order of $R^4_{\mathcal{G}_n}$ (e.g. $n^{-1/4}$ RMSE rates for the nuisance suffice when the target is parametric).
2. Absent any assumption on the convexity of the population risk (e.g. the treatment policy optimization loss), then typically *slow rates* (e.g. rates of order $O(1/\sqrt{n})$ for parametric classes) are achievable had we known the true nuisance model. In this case the impact is of nuisance estimation error is of the order $R^2_{\mathcal{G}_n}$ so, once again, $n^{-1/4}$ RMSE rates for the nuisance suffice when the target is parametric.

To extend the sufficient conditions above to arbitrary classes, we give conditions on the relative complexity of the target and nuisance classes—quantified via *metric entropy*—under which the sample splitting meta-algorithm achieves oracle rates (assuming the two black-box estimation algorithms are appropriately instantiated). This allows us to extend several prior works beyond the parametric regime to complex nonparametric target classes. Our technical results extends the works of [Yang and Barron, 1999; Rakhlin *et al.*, 2017], which provide minimax optimal rates without nuisance components and utilize the technique of *aggregation* in designing optimal algorithms.

The flexibility of our approach allows us to instantiate the framework with any machine learning model and algorithm of interest for both nuisance and target model estimation, and to utilize the vast literature on generalization bounds in machine learning to establish data-dependent and dimension-independent rates for several classes of interests. For instance, our approach allows us to use recent work on size-independent generalization error of neural networks. We obtain sharp guarantees for these specific model classes and more as a consequence of a new analysis for empirical risk minimization with plug-in estimation of nuisance parameters in the presence of orthogonality. Our results on plugin empirical risk minimization extend the local Rademacher complexity analysis of generalization bounds [Koltchinskii and Panchenko, 2000; Bartlett *et al.*, 2005], to account for the impact of the nuisance error. In the slow rate regime we also give a new analysis of *variance-penalized* empirical risk minimization, which allows us to recover and extend several prior results in the literature on policy learning. Our result improves upon the variance-penalized risk minimization approach of [Maurer and Pontil, 2009] by replacing the dependence on the metric entropy at a fixed approximation level with the critical radius, which is related to the entropy integral.

As a consequence of focusing on excess risk, we obtain oracle rates under weaker assumptions on the data generating process than in previous works. Notably, we obtain guarantees even when the target model is misspecified and the target parameters are not identifiable. For instance, for sparse high-dimensional linear classes, we obtain optimal prediction rates with no restricted eigenvalue assumptions. We highlight the applicability of our results to four settings of primary importance in the literature: 1) estimation of heterogeneous treatment effects from observational data, 2) offline policy optimization, 3) domain adaptation, 4) learning with missing data. For each of these applications, our general theorems allow for the use of arbitrary estimators for the nuisance and target model classes and provide robustness to the nuisance estimation error.

References

[Athey and Wager, 2017] Susan Athey and Stefan Wager. Efficient policy learning. *arXiv preprint arXiv:1702.02896*, 2017.

[Athey *et al.*, 2019] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

[Bartlett *et al.*, 2005] Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[Bousquet *et al.*, 2004] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to sta-

tistical learning theory. In *Advanced lectures on machine learning*, pages 169–207. Springer, 2004.

[Chernozhukov *et al.*, 2016] Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, and Whitney K Newey. Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*, 2016.

[Chernozhukov *et al.*, 2017] Victor Chernozhukov, Matt Goldman, Vira Semenova, and Matt Taddy. Orthogonal machine learning for demand estimation: High dimensional causal inference in dynamic panels. *arXiv preprint arXiv:1712.09988*, 2017.

[Chernozhukov *et al.*, 2018a] 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, 2018.

[Chernozhukov *et al.*, 2018b] Victor Chernozhukov, Denis Nekipelov, Vira Semenova, and Vasilis Syrgkanis. Plug-in regularized estimation of high-dimensional parameters in nonlinear semiparametric models. *arXiv preprint arXiv:1806.04823*, abs/1806.04823, 2018.

[Chernozhukov *et al.*, 2018c] Victor Chernozhukov, Whitney Newey, and James Robins. Double/de-biased machine learning using regularized riesz representers. *arXiv preprint arXiv:1802.08667*, 2018.

[Dudík *et al.*, 2011] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1097–1104. Omnipress, 2011.

[Foster and Syrgkanis, 2019] Dylan J. Foster and Vasilis Syrgkanis. Statistical learning with a nuisance component. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1346–1348, Phoenix, USA, 25–28 Jun 2019. PMLR.

[Friedberg *et al.*, 2018] Rina Friedberg, Julie Tibshirani, Susan Athey, and Stefan Wager. Local linear forests. *arXiv preprint arXiv:1807.11408*, 2018.

[Kallus and Zhou, 2018] Nathan Kallus and Angela Zhou. Policy evaluation and optimization with continuous treatments. In *International Conference on Artificial Intelligence and Statistics*, pages 1243–1251, 2018.

[Koltchinskii and Panchenko, 2000] V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. *High Dimensional Probability II*, 47:443–459, 2000.

[Kosorok, 2008] Michael R Kosorok. *Introduction to empirical processes and semiparametric inference.* Springer, 2008.

[Maurer and Pontil, 2009] Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. In *The 22nd Conference on Learning Theory (COLT*, 2009.

[Neyman, 1959] Jerzy Neyman. Optimal asymptotic tests of composite hypotheses. *Probability and statsitics*, pages 213–234, 1959.

[Neyman, 1979] Jerzy Neyman. C (α) tests and their use. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 1–21, 1979.

[Nie and Wager, 2017] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*, 2017.

[Oprescu *et al.*, 2019] Miruna Oprescu, Vasilis Syrgkanis, and Zhiwei Steven Wu. Orthogonal random forest for causal inference. In *International Conference on Machine Learning*, pages 4932–4941, 2019.

[Rakhlin *et al.*, 2017] Alexander Rakhlin, Karthik Sridharan, and Alexandre B Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2):789–824, 2017.

[Robinson, 1988] Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954, 1988.

[Swaminathan and Joachims, 2015] Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823, 2015.

[Vapnik, 1995] Vladimir N Vapnik. *The nature of statistical learning theory*. Springer, 1995.

[Yang and Barron, 1999] Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.

[Zhou *et al.*, 2018] Zhengyuan Zhou, Susan Athey, and Stefan Wager. Offline multi-action policy learning: Generalization and optimization. *arXiv preprint arXiv:arXiv:1810.04778*, 2018.