Proceedings of Machine Learning Research vol 99:1–3, 2019 32nd Annual Conference on Learning Theory

Statistical Learning with a Nuisance Component (Extended Abstract)

Dylan J. Foster DYLANF@MIT.EDU
Massachusetts Institute of Technology

Vasilis Syrgkanis VASY@MICROSOFT.COM
Microsoft Research, New England

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We provide excess risk guarantees for statistical learning in a setting where the population risk with respect to which we evaluate the target model depends on an unknown model that must be to be estimated from data (a "nuisance model"). We analyze a two-stage sample splitting meta-algorithm that takes as input two arbitrary estimation algorithms: one for the target model and one for the nuisance model. 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 model—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.[1]

Keywords: statistical learning, double machine learning, policy learning, treatment effects, neyman orthogonality, local rademacher complexity

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, such as attempting to predict counterfactual outcomes for different interventions or optimizing policies over such interventions, are not pure prediction problems, but rather are causal in nature. It is important to address the causal aspect 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—for example, data not collected via randomized trial or via a known treatment policy—one typically needs to estimate many other quantities that are not of primary

1. This paper is an extended abstract. The full version appears as arXiv preprint 1901.09036 v3.

interest, but that can be used to de-bias a purely predictive machine learning 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.

This motivation leads 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 model* $\widehat{\theta}_n \in \Theta_n$ 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 model, but also on a *nuisance model* 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* evaluated at the unknown true nuisance model:

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - \inf_{\theta \in \Theta_n} L_{\mathcal{D}}(\theta, g_0). \tag{1}$$

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), we make excess risk the primary focus of our work, independent of the interpretation. We develop algorithms and analysis tools that generically address (1), then apply these tools to a number of applications of interest.

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 (e.g., regularized regression, gradient boosting, or neural network estimation), we assume access to 1) an arbitrary estimation algorithm for the target class with a black-box guarantee for excess risk in the case where a nuisance value $g \in \mathcal{G}_n$ is given, and 2) an arbitrary estimation algorithm for the nuisance model. We analyze a two-stage sample splitting meta-algorithm that first estimates the nuisance model, then uses this estimator as a plug-in to estimate the target model. 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?*

Overview of results. We show that *Neyman orthogonality*, which has been used to prove oracle rates for inference in semiparametric models (Neyman, 1959, 1979; Chernozhukov et al., 2018), 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 $\widehat{\theta}_n$.
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 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.

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 extend the work 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. We also provide bounds for plug-in empirical risk minimization that extend the local Rademacher complexity analysis of generalization error (Bartlett et al., 2005) to account for the impact of the nuisance error.

References

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

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.

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

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

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

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

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