Northwestern | INSTITUTE FOR POLICY RESEARCH

RESEARCH EXCELLENCE • POLICY IMPACT

IPR Working Paper Series

WP-24-10

Comprehensive OOS Evaluation of Predictive Algorithms with Statistical Decision Theory

Jeff Dominitz
NORC at the University of Chicago

Charles F. Manski
Northwestern University and IPR

Version: March 25, 2024

DRAFT

Abstract

Dominitz and Manski argue that comprehensive out-of-sample (OOS) evaluation using statistical decision theory (SDT) should replace the current practice of K-fold and Common Task Framework validation in machine learning (ML) research. SDT provides a formal framework for performing comprehensive OOS evaluation across all possible (1) training samples, (2) populations that may generate training data, and (3) populations of prediction interest. Regarding feature (3), the researchers emphasize that SDT requires the practitioner to directly confront the possibility that the future may not look like the past and to account for a possible need to extrapolate from one population to another when building a predictive algorithm. SDT is simple in abstraction, but it is often computationally demanding to implement. They discuss progress in tractable implementation of SDT when prediction accuracy is measured by mean square error or by misclassification rate. They summarize research studying settings in which the training data will be generated from a subpopulation of the population of prediction interest. They also consider conditional prediction with alternative restrictions on the state space of possible populations that may generate training data. They conclude by calling on ML researchers to join with econometricians and statisticians in expanding the domain within which implementation of SDT is tractable.

The authors have benefitted from the opportunity to present this work in a seminar at Northwestern University.

1. Introduction

As hopes and fears related to artificial intelligence (AI) continue to grow, it is important to recognize that the impact of any AI system will always be tied to the quality of the predictive algorithms on which the AI is based. Throughout the 20th century, evaluation of predictive algorithms trained on sample data was primarily the domain of statisticians and econometricians, who use frequentist or Bayesian statistical theory to propose and evaluate prediction methods. In the 21st century, prediction is increasingly performed by machine learning (ML) researchers who view prediction methods as computational algorithms and who do not use statistical theory to assess the algorithms. Instead, ML researchers perform *out-of-sample* (OOS) tests of predictive accuracy.

In an influential early article, Breiman (2001) argued for this approach, writing (p. 201): "Predictive accuracy on test sets is the criterion for how good the model is." He asserted that OOS evaluation is close to theory-free, stating (p. 205): "the one assumption made in the theory is that the data is drawn i.i.d. from an unknown multivariate distribution." Although Breiman did not state it explicitly, we conjecture that he had in mind settings where the data are drawn from the distribution of prediction interest, not from just *any* unknown distribution.

Over the past two decades, OOS evaluation has often been applied even without this basic assumption. Taddy (2019) remarked on the subsequent wide adoption of the approach, writing (p.70): "Machine learning's wholesale adoption of OOS validation as the arbitrator of model quality has freed the ML engineer from the need to *theorize* about model quality." Taddy described OOS evaluation this way (p. 70):

> "Out-of-sample validation is a basic idea: you choose the best model specification by comparing predictions from models estimated on data that was not used during the model 'training' (fitting). This can be formalized as a cross-validation routine: you split the data into K 'folds,' and then K times fit the model on all data but the K^{th} fold and evaluate its predictive performance (e.g., mean squared error or misclassification rate) on the left-out fold. The model with optimal average OOS performance (e.g., minimum error rate) is then deployed in practice."

Rather than use a K-fold split of the training data to form validation samples, OOS evaluation is sometimes performed in the so-called 'Common Task Framework' (CTF). Here, predictive accuracy is measured in some pre-specified test data that may differ systematically from the training sample. Donoho (2017) described the CTF as follows:

> "An instance of the CTF has these ingredients: (a) A publicly available training dataset involving, for each observation, a list of (possibly many) feature measurements, and a class label for that observation. (b) A set of enrolled competitors whose common task is to infer a class prediction rule from the training data. (c) A scoring referee, to which competitors can submit their prediction rule. The referee runs the prediction rule against a testing dataset, which is sequestered behind a Chinese wall. The referee objectively and automatically reports the score (prediction accuracy) achieved by the submitted rule. All the competitors share the *common task* of training a prediction rule which will receive a good score; hence the phase *common task framework*. A famous recent example is the Netflix Challenge, where the common task was to predict Netflix user movie selections."

K-fold and CTF OOS evaluation may appear attractive to persons who lack expertise in statistical theory, but who feel that they can appraise prediction accuracy heuristically. However, it should be obvious that these types of OOS evaluation cannot yield generalizable lessons. This was recognized early on by Cox (2001), whose published Comment on Breiman (2001) distinguished between cases where (p. 216-17): "prediction is localized to situations directly similar to those applying to the *[training]* data" and those where "prediction is under quite different conditions from the data." On the former, Cox concluded that the algorithmic "empirical black-box approach" may be preferable to explicit modelling of the data generating process, whereas, in the latter case "prediction, always hazardous, without some understanding of underlying process and linking with other sources of information, becomes more and more tentative."

More recently, Taddy (2019) wrote (p. 62): "Machine learning can do fantastic things, but it is basically limited to predicting a future that looks mostly like the past." Poggio and Fraser (2024) caution that the popular ML approach of fitting training data to *deep neural networks* should be expected to yield good predictions of an outcome y conditional on covariates x only if the true best predictor of interest (often the conditional mean function) is "compositionally sparse;" that is, if it has a structure similar to that of deep neural network functions. We share the important concerns of Cox, Taddy, and Poggio and Fraser.

Efron (2020), in an article contrasting the perspectives of statisticians and computer scientists, wrote (p. S49): "In place of theoretical criteria, various prediction competitions have been used to grade algorithms in the so-called 'Common Task Framework.' . . . None of this is a good substitute for a so-far nonexistent theory of optimal prediction." We agree with Efron that the prediction competitions of the CTF are not a satisfactory way to evaluate prediction methods. However, we sharply disagree with Efron's second statement. To reiterate what Manski (2023) observed recently (p. 648):

> "[Efron] was not correct when he stated that a theory of optimal prediction is 'so-far nonexistent'. Wald (1939, 1945, 1950) considered the general problem of using sample data to make decisions. He posed the task as choice of a *statistical decision function*, which maps potentially available data into a choice among the feasible actions. His development of statistical decision theory provides a broad framework for decision making with sample data, yielding optimal decisions when these are well-defined and proposing criteria for reasonable decision making more generally."

Wald recommended ex ante (frequentist) evaluation of statistical decision functions as *procedures* applied as a sampling process is engaged repeatedly to draw independent data samples. Statistical decision theory (SDT) measures the mean performance of a prediction method across all possible training samples, when the objective is to predict outcomes in a population that may possibly differ from the one generating the data in the training sample.

SDT is remote from the types of OOS evaluation currently practiced by ML researchers, but it is not remote conceptually. Indeed, SDT provides a formal framework for performing what we shall term *comprehensive OOS* evaluation. SDT performs OOS evaluation across all possible (1) training samples, (2) populations that may generate training data, and (3) populations of prediction interest. We think that feature (3) is particularly important. SDT requires the practitioner to directly confront the possibility that the future may not look like the past and to account for a possible need to extrapolate from one population to another when building a predictive algorithm.

Even in the best-case scenario where the future will credibly look like the past—i.e., no extrapolation problem—SDT provides a framework for OOS evaluation that is scientifically more rigorous than the current standard practice of K-fold and CTF validation. When the training sample is known to be drawn i.

i. d. from the population distribution of prediction interest, one must recognize that this sample is just one draw among all possible samples from this population distribution. With comprehensive OOS evaluation, the performance of the predictive algorithm is assessed across all samples on which it could be trained, rather than just the one sample on which it actually is trained.

ML researchers often motivate their versions of OOS evaluation by stating that it protects against drawing misleading conclusions from prediction performance on the training sample, which may be unrealistically high due to so-called "overfitting" the data. Concern with overfitting does not arise in the Wald framework because it evaluates performance across all feasible samples, not a particular training sample. Whereas some may believe that the big data revolution renders attention to sampling variation unnecessary, the desire to build ML models with high-dimensional covariates gives rise to the same finite sample concerns that SDT was developed to address. Even if the total sample size is orders of magnitude larger than Wald may have envisioned three-quarters of a century ago, statistical imprecision is an important consideration when conditioning on covariates whose distribution has sufficiently large support.

In this paper, we argue that comprehensive OOS evaluation performed using SDT should replace the current practice of K-fold and CTF validation in ML research. The argument is simple to make in abstraction. As we explain in Section 2, the Wald framework is remarkably general, transparent, and intellectually attractive. In principle, it enables comparison of essentially all predictive algorithms, the only caveat being satisfaction of weak mathematical regularity conditions. It enables comparison of alternative sampling processes generating training data. It uses no asymptotic approximations when interpreting the training data. Further, the population of interest in prediction may differ from that generating the training data.

The argument is more challenging to make in practice. Application of SDT is easy in some important settings, but it is often computationally demanding. Computation commonly requires numerical methods to find approximate solutions.

SDT requires the practitioner to specify a decision criterion. Among the criteria that have been studied, we find minimax regret particularly appealing. A statistical decision function (SDF) with small maximum

regret is uniformly near-optimal across all possible populations generating training data and populations of prediction interest. In the terminology of SDT, a possible pair of such populations is a *state of nature*. The set of all possible populations is the *state space*. Maximum regret is computed across the state space.

Illustration: Consider the familiar problem of using sample data ψ to make a point prediction $p(\psi)$ of a binary outcome y. Analysis commonly supposes that (ψ, y) are generated by some joint probability distribution $P(\psi, y)$. If P were known, one might evaluate the accuracy of prediction function $p(\cdot)$ by its mean square error (MSE), $E[y - p(\psi)]^2$, or by its misclassification rate (MCR), $P[y \neq p(\psi)]$. In practice, however, P generally is not known. Suppose one knows that P lies in a specified set of distributions $(P_s, s \in S]$. This is the state space. Then one does not know the true MSE or MCR of $p(\cdot)$. However, one can compute their possible values $\{E_s[y - p(\psi)]^2, s \in S\}$ and $\{P_s[y \neq p(\psi)], s \in S\}$. We show in Section 2 that, when accuracy is measured by MSE, the regret of $p(\cdot)$ in state of nature s is

$$P_s[p(\psi) = 1][1 - P_s(y = 1)] + P_s(y = 1)\{1 - P_s[p(\psi) = 1]\} - P_s(y = 1)[1 - P_s(y = 1)].$$

When accuracy is measured by MCR, the regret of $p(\cdot)$ in state s is

$$P_s[p(\psi) = 1][1 - P_s(y = 1)] + P_s(y = 1)\{1 - P_s[p(\psi) = 1]\} - \min[P_s(y = 1), 1 - P_s(y = 1)].$$

In both cases, if $P_s(y = 1)$ were known, the optimal predictor would be the data-invariant function $p(\psi) = 1$ for all ψ if $P_s(y = 1) \geq ½$ and $p(\psi) = 0$ for all ψ if $P_s(y = 1) \leq ½$. The problem in practice is to predict y with knowledge of the data ψ but without knowledge of $P_s(y = 1)$.

Computation of maximum regret across the state space may be easy or challenging, depending on the state space. Using either MSE or MCR to measure accuracy, regret is a function of the two probabilities $P_s[p(\psi) = 1]$ and $P_s(y = 1)$, which jointly lie in the unit square. Hence, regret must be maximized over a subset of the unit square. The nature of this subset is determined by the state space. ■

Section 3 summarizes progress in tractable implementation of SDT when prediction accuracy is measured by MSE and the population that generates the training data is a subset of the population of

prediction interest. Thus, one wants to learn an outcome distribution P(y), but the training data are drawn from a sub-population P(y|δ = 1) for some binary indicator δ. Important applications arise in settings with missing outcome data, where a random sample of outcomes is drawn from the population of prediction interest, but only a subset of the drawn outcomes are observed. Dominitz and Manski (2017) evaluated the maximum regret of various tractable SDFs that researchers may use in practice, when a researcher lacks credible assumptions to model the distribution of missing data. We summarize the findings, which are applicable not only to settings with missing data but more generally when the population generating the training data is a subset of the population of prediction interest.

Section 4 considers prediction of an outcome y conditional on a specified covariate value x . To focus on the problem of statistical imprecision, we suppose that the population of prediction interest is the same as that generating the training data. The central issue is specification of cross-covariate restrictions on conditional outcome distributions, which enable outcome data associated with covariate values x' ≠ x to be informative about P(y|x).

Statistical and econometric theory has long studied settings in which cross-covariate restrictions are imposed by assuming all conditional distributions, or their means, lie in a specified finite-dimensional set (parametric regression) or a specified space of suitable smooth functions (nonparametric regression). The ML literature using deep neural networks as predictor functions has recently favored assumptions that conditional probability distributions or means are compositionally sparse, which Poggio and Fraser (2024) described as (p. 1): “a key principle underlying successful learning architectures.” Yet other cross-covariate restrictions on conditional distributions are specified in the bounded-variation assumptions studied by Manski (2023). We do not take a stand favoring any one class of restrictions on the state space over another—the choice should be application specific. Our theme is that SDT may in principle be used to perform comprehensive OOS evaluation however a researcher chooses to restrict the state space.

The concluding Section 5 calls on ML researchers to join with econometricians and statisticians in expanding the domain within which implementation of comprehensive OOS evaluation is tractable.

2. The Statistical Decision Theory Perspective on Prediction

When Abraham Wald formally considered the general problem of using sample data to make decisions, this included the use of data to make predictions. His first major contribution on the way to development of statistical decision theory was to recast the hypothesis-testing problem previously posed by Neyman and Pearson (1928, 1933) as a decision problem in which one must choose between two feasible actions. ML researchers now refer to this type of decision as a binary *classification problem*. We discuss Wald's formalization of the classification problem in Section 2.2, after we present the general structure of SDT in Section 2.1. Section 2.3 explains how SDT views prediction—in particular, SDFs that use predictions to make decisions. The exposition in these sections draws on Manski (2021, 2023).

2.1. Statistical Decision Theory in Abstraction

Wald utilized the now standard decision-theoretic framing of the choice problem of a decision maker (DM) who must choose an action yielding loss that depends on an unknown state of nature. The DM specifies a state space listing the states considered possible. The DM wants to minimize loss, but must choose without knowing the true state.

A fundamental difficulty with loss minimization under uncertainty is apparent even in a simple setting with two feasible actions, say A and B, and two possible choice environments, say s_1 and s_2. Suppose that action A yields smaller loss in environment s_1 and action B yields smaller loss in s_2. If it is not known whether s_1 or s_2 is the actual choice environment, it is not known which action is better. Thus, minimization of loss is logically impossible. At most one can seek a reasonable way to make a choice. A basic issue is how to interpret and justify the word 'reasonable.'

To begin, Section 2.1.1 presents this problem in a setting where the DM does not observe sample data. Section 2.1.2 explains how Wald generalized this setting by supposing that the DM observes sample data that may be informative about the true state. Section 2.1.3 discusses computational approaches.

2.1.1. Decisions Without Sample Data

Consider a DM who faces a predetermined choice set C and believes that the true state of nature s^* lies in state space S. The state space may be finite-dimensional (parametric) or larger (nonparametric). Objective function $L(\cdot, \cdot)$: $C \times S \to R^1$ maps actions and states into loss. The DM ideally would minimize $L(\cdot, s^*)$ over C, but the DM does not know s^*.

It is generally accepted that choice should respect dominance. Action $c \in C$ is weakly dominated if there exists a $d \in C$ such that $L(d, s) \leq L(c, s)$ for all $s \in S$ and $L(d, s) < L(c, s)$ for some $s \in S$. To choose among undominated actions, decision theorists have proposed various ways of using $L(\cdot, \cdot)$ to form functions of actions alone, which can be optimized. In principle, one should only consider undominated actions, but it often is difficult to determine which actions are undominated. Hence, in practice it is common to optimize over the full set of feasible actions. We define decision criteria accordingly. We use max and min notation, without concern for the subtleties that sometimes make it necessary to use sup and inf operations.

Wald (1945) studied choice when the DM places a subjective probability distribution π on the state space, averages state-dependent loss with respect to π, and minimizes subjective average loss $\int L(c, s)d\pi$ over C. The criterion solves

$$(1) \qquad \min_{c \in C} \int L(c, s)d\pi.$$

Wald (1945, 1950) considered choice when the DM does not place a subjective distribution on the state space. In this setting, he studied minimax choice, which selects an action that works uniformly well over all of S in the sense of minimizing the maximum loss attainable across S. The minimax criterion is

$$(2) \qquad \min_{c \in C} \max_{s \in S} L(c, s).$$

Savage (1951), in a book review of Wald (1950), suggested a different formalization of the idea of selecting an action that works uniformly well over all of S. This formalization, which has become known as the minimax-regret (MMR) criterion, solves the problem

$$(3) \quad \min_{c \in C} \max_{s \in S} [L(c, s) - \min_{d \in C} L(d, s)].$$

Here $L(c, s) - \min_{d \in C} L(d, s)$ is the *regret* of action c in state s—i.e., the increment to loss in state s arising from making choice c rather than the optimal choice in that state. The true state being unknown, one evaluates c by its maximum regret over all states and selects an action that minimizes maximum regret. The maximum regret of an action measures its maximum distance from optimality across states.

Criteria (1)—(3) have become the most prominent criteria studied in decision theory, but they are not alone. Hurwicz (1951) suggested minimization of a weighted average of worst and best possible outcomes. Rather than assert a complete subjective distribution on the state space or none, a DM might assert a partial subjective distribution, placing lower and upper probabilities on states. One then might minimize maximum subjective average loss or minimize maximum subjective average regret. These criteria combine elements of averaging across states and concern with uniform performance across states. It appears that this idea was first suggested by Hurwicz (1951). The idea was later taken up by statistical decision theorists. See Berger (1985) and Walley (1990).

2.1.2. Statistical Decision Problems

Statistical decision problems add to the above structure by supposing that the DM observes data generated by some sampling distribution. Knowledge of the sampling distribution is generally incomplete. To express this, one extends state space S to list the feasible sampling distributions, denoted $(Q_s, s \in S)$. Let Ψ_s denote the sample space in state s; Ψ_s is the set of samples that may be drawn under distribution Q_s. The literature typically assumes that the sample space does not vary with s and is known. We assume this and denote the sample space as Ψ. Then a statistical decision function, $c(\cdot): \Psi \to C$, maps the sample data

into a chosen action. Henceforth, $\psi \in \Psi$ is a possible realization of the sample data; that is, a possible training sample.

Wald's concept of a statistical decision function embraces all mappings [data $\rightarrow$ action]. SDF $c(\cdot)$ is a deterministic function after realization of the sample data, but it is a random function ex ante. Hence, the loss achieved by $c(\cdot)$ is a random variable ex ante. Wald's theory evaluates the performance of $c(\cdot)$ in state s by $Q_s\{L[c(\psi), s]\}$, the ex-ante distribution of loss that it yields across realizations ψ of the sampling process.

It remains to ask how DMs might compare the loss distributions yielded by different SDFs. DMs want to minimize loss, so it seems self-evident that they should prefer SDF $d(\cdot)$ to $c(\cdot)$ in state s if $Q_s\{L[d(\psi), s]\}$ is stochastically dominated by $Q_s\{L[c(\psi), s]\}$. It is less obvious how they should compare SDFs whose loss distributions do not stochastically dominate one another.

Wald proposed measurement of the performance of $c(\cdot)$ in state s by its expected loss across samples; that is, $E_s\{L[c(\psi), s]\} \equiv \int L[c(\psi), s] dQ_s$. He used the term *risk* to denote the mean performance of an SDF across samples. An alternative that has drawn only slight attention measures performance by quantile loss (Manski and Tetenov, 2023).

Not knowing the true state, a DM evaluates $c(\cdot)$ by the expected loss vector $(E_s\{L[c(\psi), s]\}, s \in S)$. Using the term *inadmissible* to denote weak dominance when evaluating performance by risk, Wald recommended elimination of inadmissible SDFs from consideration. As in decisions without sample data, there is no clearly best way to choose among admissible SDFs. SDT has mainly studied the same criteria as has decision theory without sample data. Let Γ be a specified set of SDFs, each mapping $\Psi \rightarrow\!\!\!\!\!\rightarrow C$. The statistical versions of criteria (1), (2), and (3) are

$$(4) \qquad \min_{c(\cdot) \in \Gamma} \int E_s\{L[c(\psi), s]\}\, d\pi,$$

$$(5) \qquad \min_{c(\cdot) \in \Gamma} \max_{s \in S} E_s\{L[c(\psi), s]\},$$

$$(6) \qquad \min_{c(\cdot) \in \Gamma} \max_{s \in S} \left(E_s\{L[c(\psi), s]\} - \min_{d \in C} L(d, s) \right).$$

Each of criteria (4) – (6) has been deemed reasonable by some decision theorists, but each has also drawn criticism. To summarize some main points, minimization of subjective average loss (4) (aka Bayes risk) may be appealing if one has a credible basis to place a subjective probability distribution on the state space, but not otherwise. Concern with specification of priors motivated Wald to study the minimax criterion (5). However, minimax was criticized by Savage as 'ultrapessimistic.'

Manski (2004, 2021) put forward a conceptual reason to implement the MMR criterion (6). The conceptual appeal of using maximum regret to measure performance is that it quantifies how lack of knowledge of the true state of nature diminishes the quality of decisions. The term "maximum regret" is a shorthand for the maximum sub-optimality of a decision criterion across the feasible states of nature. An SDF with small maximum regret is uniformly near-optimal across all states. This is a desirable property.

Whichever of criteria (4)—(6) one uses, SDT performs a comprehensive OOS evaluation of an SDF. In each state of nature s, the expected loss $E_s\{L[c(\psi), s]\}$of SDF $c(\cdot)$ measures its performance across all possible training samples. The state-dependent vector $\{E_s\{L[c(\psi), s]\}, s \in S\}$ of expected loss measures performance across all possible populations that may have generated the training sample and all possible populations of decision interest. Direct measurement of performance across all possible populations replaces any need for test samples to be drawn.

Recall that, when arguing for his narrow form of OOS evaluation, Breiman (2001) stated that "the one assumption made in the theory is that the data is drawn i.i.d. from an unknown multivariate distribution." This assumption is unnecessary in SDT. Any sampling distribution can generate the observed data.

2.1.3. Computation

Subject to regularity conditions ensuring that the relevant expectations and extrema exist, problems (4) – (6) offer criteria for decision making with sample data that are broadly applicable in principle. The

primary challenge is computational. Problems (4) − (6) have tractable analytical solutions only in certain cases. Computation commonly requires numerical methods to find approximate solutions.

Expected loss $E_s\{L[c(\psi), s]\}$ typically does not have an explicit form, but it can be well-approximated by Monte Carlo integration. One draws repeated values of ψ from distribution Q_s, computes the average value of $L[c(\psi), s]$ across the values drawn, and uses this to estimate $E_s\{L[c(\psi), s]\}$. Monte Carlo integration can also be used in criterion (4) to approximate the subjective average of expected loss.

The main computational challenges are determination of the extrema across actions in problem (6), across states in problems (5) − (6), and across SDFs in problems (4) − (6). Solution of $\min_{d \in C} L(d, s)$ in (6) is often straightforward but sometimes difficult. Finding extrema over S must cope with the fact that the state space commonly is uncountable. In applications where the quantity to be optimized varies smoothly over S, a simple approach is to compute the extremum over a suitable finite grid of states.

The most difficult computational challenge usually is to optimize over the feasible SDFs. No generally applicable approach is available. Hence, applications of SDT necessarily proceed case-by-case. It may not be tractable to find the best feasible SDF, but one often can evaluate the performance of relatively simple SDFs that researchers use in practice.

2.2. Binary Classification Problems

Binary classification problems are ones in which the choice set C contains two feasible actions, say A and B. An SDF $c(\cdot)$ partitions Ψ into two regions that separate the data yielding choice of each action. These are $\Psi_{c(\cdot)A} \equiv [\psi \in \Psi: c(\psi) = A]$ and $\Psi_{c(\cdot)B} \equiv [\psi \in \Psi: c(\psi) = B]$. Thus, one chooses A if the data lie in set $\Psi_{c(\cdot)A}$ and B if the data lie in $\Psi_{c(\cdot)B}$.

Choice between two actions can be viewed as hypothesis tests, but the Wald perspective on testing differs from that of Neyman and Pearson (1928, 1933). An hypothesis test partitions S into two regions, S_A and S_B, that separate the states in which actions A and B are uniquely optimal. Thus, S_A contains the states $[s \in S: L(A, s) < L(B, s)]$ and S_B contains $[s \in S: L(B, s) < L(A, s)]$. States yielding equal loss may be placed

in S_a or S_B. A test yields a Type I error when the true state lies in S_A but $c(\psi) = B$. It yields a Type II error when the true state lies in S_B, but $c(\psi) = A$.

Neyman-Pearson testing views S_A and S_B asymmetrically, calling the former the null hypothesis and the latter the alternative. A longstanding convention has been to restrict attention to tests in which the probability of a Type I error is no larger than a predetermined value, usually 0.05, for all $s \in S_a$. In contrast, SDT does not restrict attention to tests that yield a predetermined upper bound on the probability of a Type I error.

Wald (1939) proposed evaluation of the performance of an SDF for binary classification by the expected loss that it yields across realizations of the sampling process. The loss distribution in state s is Bernoulli, with mass points max [L(a, s), L(b, s)] and min [L(a, s), L(b, s)]. These coincide if L(a, s) = L(b, s). When $L(a, s) \neq L(b, s)$, let $R_{c(\cdot)s}$ denote the probability that $c(\cdot)$ yields an error, choosing the inferior action over the superior one. That is,

$$(7) \qquad R_{c(\cdot)s} = Q_s[c(\psi) = b] \quad \text{if } L(a, s) < L(b, s),$$
$$= Q_s[c(\psi) = a] \quad \text{if } L(b, s) < L(a, s).$$

These are the probabilities of Type I and Type II errors.

The probabilities that loss equals min [L(a, s), L(b, s)] and max [L(a, s), L(b, s)] are $1 - R_{c(\cdot)s}$ and $R_{c(\cdot)s}$. Hence, expected loss is

$$(8) \qquad E_s\{L[c(\psi), s]\} = R_{c(\cdot)s}\{\max [L(a, s), L(b, s)]\} + [1 - R_{c(\cdot)s}]\{\min [L(a, s), L(b, s)]\}$$
$$= \min [L(a, s), L(b, s)] + R_{c(\cdot)s} \cdot |L(a, s) - L(b, s)|.$$

$R_{c(\cdot)s} \cdot |L(a, s) - L(b, s)|$ is the expected regret of $c(\cdot)$. Thus expected regret, which was defined in abstraction in (6), has a simple form when choice is binary. It is the product of the error probability and the magnitude of the incremental loss when an error occurs.

2.3. Prediction-Based SDFs

We have observed that Wald's concept of an SDF embraces all mappings [data → action]. This very broad domain includes SDFs that make predictions and use the predictions to make decisions. These have the form [data → prediction → action], first making a prediction and then using the prediction to make a decision. There seems to be no accepted term for such SDFs, so we call them *prediction-based* SDFs.

To formalize prediction-based SDFs, we first need to formalize the concept of prediction and embed it in a decision problem. To do this, we bring to bear the venerable idea of probabilistic conditional prediction. This idea postulates a population characterized by a joint probability distribution P(y, x), where (y, x) takes values in some set Y × X. A member is drawn at random from the sub-population with a specified value of x. Then the conditional distribution P(y|x) provides the ideal probabilistic prediction of y conditional on x.

We embed prediction in a decision problem by considering settings in which a value of x is specified and the state space contains a set of feasible conditional distributions P(y|x). Then $P^*(y|x)$ is the unknown true distribution of prediction interest. The DM ideally wants to minimize $L[\cdot, P^*(y|x)]$ over C, but does not know $P^*(y|x)$. One faces a statistical decision problem if one observes data ψ drawn from some sampling process.

In a setting of this type, an SDF is prediction-based if the DM uses the data ψ to form an estimate of $P^*(y|x)$ and acts as if the estimate is accurate. Let $f(\cdot|x)$: Ψ ↛ S be the predictor function used to estimate $P^*(y|x)$. The chosen action with this SDF is $d(\psi) = \text{argmin}_{c \in C} L[c, f(\psi|x)]$.

Rather than estimate the entire conditional distribution $P^*(y|x)$, researchers often use sample data to form a point prediction of outcome y. Let $p(\cdot|x)$: Ψ ↛ Y denote a predictor function that generates a point prediction. The chosen action with an SDF of this type acts as if $P^*(y|x)$ is degenerate, with all its mass at the outcome value $p(\psi|x)$. Thus, the SDF is $d(\psi) = \text{argmin}_{c \in C} L[c, p(\psi|x)]$.

Important special cases occur when the choice set C coincides with the set Y of potential outcomes.

For example, y may be real-valued and the loss function may be MSE when using c to predict y; that is,

(9) $$L[c, P(y|x)] = E[(y - c)^2|x] = \int (y - c)^2 dP(y|x).$$

Or y may be binary and the loss function may be the MCR when using c to predict y; that is,

(10) $$L[c, P(y|x)] = P(y \neq c|x) = \int 1[y \neq c] dP(y|x).$$

In both cases, acting as if P(y|x) is degenerate at p(ψ|x) yields the SDF d(ψ) = p(ψ|x).

From the perspective of SDT, the performance of a prediction-based SDF should be evaluated in the same manner as any SDF. Thus, one should first determine if the SDF is undominated. If so, one may evaluate it by the subjective expected loss it yields, the maximum loss it yields, by its maximum regret, or by some other reasonable criterion.

2.3.1. Regret with Binary Classification and Prediction

In Section 1 we stated the regret of a point predictor p(·) of a binary outcome y in state of nature s, when measuring loss by MSE or MCR. We derive these results here, now conditioning prediction on a specified covariate value x.

When loss is MSE, the risk of p(·|x) in state s is

(11) $$E[y - p(\psi|x)]^2 = V_s(y|x) + V_s[p(\psi|x)] + \{E_s(y|x) - E_s[p(\psi|x)]\}^2.$$

Risk in state s is minimized by setting $p(\psi|x) = E_s(y|x)$ for all data realizations ψ, yielding the minimum risk value $V_s(y|x)$. Hence, the regret of p(·|x) in state s is $V_s[p(\psi)|x] + \{E_s(y|x) - E_s[p(\psi|x)]\}^2$.

When both the outcome y and the point predictor p(·|x) are binary, $V_s[p(\psi|x)] = P_s[p(\psi|x) = 1] - P_s[p(\psi|x) = 1]^2$, $E_s(y|x) = P_s(y = 1|x)$, and $E_s[p(\psi|x)] = P_s[p(\psi|x) = 1]$. Hence, regret is

(12) $P_s[p(\psi|x) = 1] - P_s[p(\psi|x) = 1]^2 + \{P_s(y = 1|x) - P_s[p(\psi|x) = 1]\}^2$

$= P_s[p(\psi|x) = 1] - P_s[p(\psi|x) = 1]^2 + P_s(y = 1|x)^2 + P_s[p(\psi|x) = 1]^2 - 2 \cdot P_s(y = 1|x) \cdot P_s[p(\psi|x) = 1]$

$= P_s[p(\psi|x) = 1][1 - P_s(y = 1|x)] + P_s(y = 1|x)\{P_s(y = 1|x) - P_s[p(\psi|x) = 1]\}$

$= P_s[p(\psi|x) = 1][1 - P_s(y = 1|x)] + P_s(y = 1|x)\{1 - P_s[p(\psi|x) = 1]\} - P_s(y = 1|x)[1 - P_s(y = 1|x)].$

The first term is the probability of a misclassification of the form [y = 0, p(ψ|x) = 1]. The second term is the probability of a misclassification of the form [y = 1, p(ψ|x) = 0]. The third term is the outcome variance $V_s(y|x)$.

When loss is MCR, the risk of p(·|x) in state s is

(13) $P_s[y \neq p(\psi)|x] = P_s[p(\psi) = 1|x][1 - P_s(y = 1|x)] + P_s(y = 1|x)\{1 - P_s[p(\psi) = 1|x]\}.$

Risk in state is minimized by setting p(ψ|x) = 1 for all ψ if $P_s(y = 1|x)] > ½$ and p(ψ|x) = 0 for all ψ if $P_s(y = 1|x)] \leq ½$, yielding the minimum risk value $\min[P_s(y = 1|x), 1 - P_s(y = 1|x)]$. Hence, regret is

(14) $P_s[p(\psi) = 1|x][1 - P_s(y = 1|x)] + P_s(y = 1|x)\{1 - P_s[p(\psi) = 1|x]\} - \min[P_s(y = 1|x), 1 - P_s(y = 1|x)].$

Thus, regret when loss are MSE and MCR differ only in that the former subtracts $P_s(y = 1|x)[1 - P_s(y = 1|x)]$ from the probability of misclassification and the latter subtracts $\min[P_s(y = 1|x), 1 - P_s(y = 1|x)]$.

3. Prediction under Square Loss, with Data on a Subpopulation

We now consider a class of problems in which the training data will be generated from a subpopulation of the population of prediction interest. An important class of applications is to prediction with missing outcome data. Empirical researchers often assume that data will be missing at random, meaning that the

distribution of missing data will be the same as the distribution of observed data. However, this or another assumption fixing the distribution of missing data may not be credible. Prediction of treatment outcomes with observational data presents a particularly difficult problem of prediction with missing data. In this setting, one can observe a treatment outcome only when a person in a study population receives that treatment; that is, when a person is in the treated subpopulation. Without random assignment of persons to treatments, it is generally not credible to assume that the outcome distribution is the same in the untreated subpopulation.

The present discussion restricts attention to prediction under square loss. For notational convenience, we suppress conditioning prediction on covariates x. We showed in Section 2.3 that the regret of predictor $p(\cdot)$ in state s is $V_s[p(\psi)] + \{E_s(y) - E_s[p(\psi)]\}^2$. Hodges and Lehmann (1950) derived the MMR predictor with data from a random sample, when the outcome has known bounded range and all sample data are observed. They assumed no knowledge of the shape of the outcome distribution. Let the outcome range be [0, 1]. Then the MMR predictor is $(\mu_N\sqrt{N} + ½)/(\sqrt{N} + 1)$, where N is sample size and μ_N is the sample mean.

3.1. Prediction with Missing Data

Aiming to extend the analysis of Hodges and Lehmann, Dominitz and Manski (2017) studied prediction of bounded outcomes under square loss when some outcome data are missing. Thus, the training sample is a random sample of observations from the sub-population with observable outcomes. It is challenging to determine the MMR predictor when data are missing. Seeking a tractable approach, we studied “as-if” MMR prediction. We assumed knowledge of the population fraction of observable outcomes, but no knowledge of the distributions of observed and missing outcomes. We used the empirical distribution of the observed data as if it were the population distribution of observable outcomes. We conditioned our MMR analysis on the number K of observed outcomes, viewing it as fixed rather than random.

With no knowledge of the distribution of missing outcomes, the population mean is partially identified when the outcome is bounded. Let y be the outcome, normalized to lie in the [0, 1] interval. Let δ indicate observability of an outcome, $P(\delta = 1)$ and $P(\delta = 0)$ being the fractions of the population whose outcomes are and are not observable. Manski (1989) showed that the identification region for E(y) is the interval $[E(y|\delta = 1)P(\delta = 1), E(y|\delta = 1)P(\delta = 1) + P(\delta = 0)]$.

If this interval were known, the MMR predictor would be its midpoint $E(y|\delta = 1)P(\delta = 1) + \frac{1}{2}P(\delta = 0)$. With sample data on observable outcomes and knowledge of P(δ), one can compute its sample analog $E_K(y|\delta = 1)P(\delta = 1) + \frac{1}{2}P(\delta = 0)$. We showed that the maximum regret of this *midpoint predictor* is $\frac{1}{4}[P(\delta = 1)^2/K + P(\delta = 0)^2]$. We showed that the maximum regret of the midpoint predictor is smaller than that of $E_K(y|\delta = 1)$, a predictor that is commonly used. The latter predictor is well-motivated if data are missing at random but not otherwise.

Whereas Dominitz and Manski (2017) assumed knowledge of P(δ), this knowledge may not be available in practice. One may, however, have available a random sample of size N of the population that enables one to estimate P(δ). A midpoint predictor remains computable when P(δ) is estimated by its sample analog $P_N(\delta)$. Derivation of an analytical expression for maximum regret appears intractable, but numerical computation is feasible. Manski and Tabord-Meehan (2017) documents an algorithm coded in STATA for numerical computation of the maximum regret of this version of the midpoint predictor and other user-specified predictors.

The program is applicable when y is binary or distributed continuously. In the latter case, $P_s(y|\delta = 1)$ and $P_s(y|\delta = 0)$ are approximated by Beta distributions. Subject to these restrictions on the shapes of outcome distributions, the user can specify the state space flexibly. For example, one may assume that nonresponse is no higher than 80% or that the mean outcome for nonresponders is no lower than 0.5. One may bound the difference between the distributions $P_s(y|\delta = 1)$ and $P_s(y|\delta = 0)$.

Whereas Dominitz and Manski (2017) considered the number K of observed outcomes to be fixed, the algorithm considers a sampling process in which one draws at random a fixed number N of population

members and sees the values of the observable outcomes. Hence, the number K of observed outcomes is random. The midpoint predictor is $E_K(y|\delta = 1)P_N(\delta = 1) + \frac{1}{2}P_N(\delta = 0)$.

3.2. Prediction with Missing Data on a Counterfactual Outcome

The analysis of missing data in Dominitz and Manski (2017) may be applied to problems of predicting treatment outcomes in the absence of an ideal randomized experiment. Consider prediction of outcomes following a binary treatment. Suppose that data will be collected on a sample of individuals who receive treatment A and a sample of individuals who receive treatment B. ML researchers refer to this as "A/B testing."

If treatments will be randomly assigned with perfect compliance and if treatment response is individualistic, prediction of outcomes if all members of the population were to receive the same treatment is straightforward. However, A/B testing is rarely ideal. Considering how to proceed, Taddy (2019) wrote (p. 68):

> "More recently, as the field matures and as people recognize that not everything can be explicitly A/B tested, data scientists have discovered the importance of careful causal analysis. One of the most currently active areas of data science is combining ML tools with the sort of counterfactual inference that econometricians have long studied, hence now merging the ML and statistics material with the work of economists."

Taddy and his co-authors made an even stronger statement in Hartford et al. (2017, p. 1414):

> "Counterfactual prediction requires understanding causal relationships between so-called *treatment* and *outcome* variables."

Taddy (2019) was right to observe that structural econometric analysis of treatment selection and outcomes, which has been performed for close to a century, may help to credibly predict treatment outcomes. Yet prediction may be performed without understanding causal relationships. Counterfactual prediction may be relabeled as prediction with missing outcomes and analyzed in the manner of Dominitz

and Manski (2017). Consider prediction of the outcomes that would occur if everyone in a population were to receive treatment A. Sample data on outcomes with treatment A will be observed for those who receive A and will be missing for those who receive B. Thus, the midpoint predictor for the outcome if everyone in the population were to receive A is $E_N[y(A)]P_N(A) + ½P_N(B)$. Analogously, the midpoint predictor for the outcome if everyone were to receive B is $E_N[y(B)]P_N(B) + ½P_N(A)$.

3.3. More Data or Better Data: Choice of Sample Design in Problems with Missing Data

Moving beyond analysis of prediction with a pre-specified sampling process, Dominitz and Manski (2017) addressed the problem of a DM with a finite budget to collect sample data, after which the data will be used to choose a point prediction. Then the DM faces a joint problem of sample design and choice of a predictor. We supposed that two or more sampling processes are available, differing in the cost of data collection and the quality of the data obtained. One may use the budget to draw a large sample of low-quality data or a small sample of high-quality data. We also studied prediction using a pooled sample, combining samples of low-quality and high-quality data.

The analysis performed should be of interest to ML researchers. Much attention has been paid recently to the quality of training samples, including variation in data quality within a training sample, and how data quality impacts ML model predictions. Quality can have different meanings in different applications. As one primary example, Dominitz and Manski considered quality as determined by the prevalence of missing data.

ML researchers routinely bypass missing data problems by producing algorithms that proceed as if the data are missing at random. However, Mitra et al. (2023) raised serious concerns about the impact of so-called “structured missingness” (SM)—that is, not missing at random—on the development of ML models, stating (p. 22):

> “Indeed, it is not an exaggeration to say that issues of missingness are a primary hindrance to efficient learning at scale… However, despite the prevalence of this issue, SM has not yet been systematically

studied and we lack both a theory for SM and the tools need*[sic]* to learn efficiently from data with SM."

Dominitz and Manski (2017) developed tools for prediction with missing data where the DM seeks to choose the sample design that minimizes maximum regret for prediction of a real-valued outcome under square loss. Attention was focused on tractable predictors—in particular, the midpoint predictor defined in Section 3.1 above. The analysis imposed no assumptions that restrict the distribution of unobserved outcomes. Hence, we recognized that the DM must cope with both the statistical imprecision of finite samples and partial identification of the true state of nature.

4. Conditional Prediction with Cross-Covariate Restrictions on the State Space

In Section 2.3, we discussed how SDT studies the use of training sample data to predict an outcome y conditional on a covariate vector x, measuring loss by MSE or MCR. The discussion was abstract, considering data generated by any sampling process, conditioning prediction on any specified covariate value, with any state space and any predictor function. In principle, SDT may be used to perform comprehensive OOS evaluation of any of the huge variety of conditional prediction methods that have been developed from the late 1800s onward by statisticians, econometricians, and ML researchers. SDT is applicable whether or not the distribution P(y|x) of prediction interest is the same as the distribution that generates sample realizations of y conditional on x in the training data (y_i, x_i), i = 1, . . , N.

In practice, SDT has rarely been used to evaluate conditional prediction methods. Prediction has mainly been studied under the assumption that the training data are a random sample drawn from the distribution P(y, x) of prediction interest, or at least that P(y|x) generates the training realizations of y conditional on x. The literature on Bayesian prediction predominately uses the *conditional Bayes* paradigm. This centers attention on the posterior predictive distribution, computed after training data are observed, not on ex ante Bayes risk as in SDT. There has been little minimax or MMR evaluation of conditional prediction methods; an isolated case of MMR evaluation of classical linear regression is Sawa and

Hiromatsu (1973). Research on nonparametric regression has mainly studied asymptotic properties of estimates when the covariate distribution has positive density in a neighborhood of a value of interest and the conditional expectation E(y|x) varies smoothly with x in a local sense, such as differentiability. Donoho et al. (1995) reviews many findings. An isolated instance of MMR evaluation of kernel nonparametric regression is Manski (2023), which bounds the global rather than local variation of E(y|x) with x. See Section 4.2 for discussion.

Whereas statisticians and econometricians have used some sort of statistical theory to study prediction methods, ML researchers have largely performed heuristic K-fold and CTF OOS validation, described in Section 1. An isolated exception is Schmidt-Hieber (2020), who performed some asymptotic analysis of deep neural network predictors when conditional means are compositionally sparse. Despite the absence of theory, ML researchers assert that K-fold and CTF validation exercises show that certain favored methods commonly "work" in practice. Yet the reasons why these methods "work" and the circumstances in which they "work" continue to be controversial. In principle, SDT should be able to shed light on these issues.

4.1. The Fundamental Importance of the Cross-Covariate Structure of the State Space

A focus of controversy has been the asserted capacity of certain prediction methods advocated by ML researchers, initially regression trees and more recently deep neural networks, to successfully predict outcomes when covariate vectors have high dimension relative to sample size. Reacting to the longstanding concern that prediction becomes increasingly difficult as the dimension of the covariate vector increases (aka the curse of dimensionality), Breiman (2001) expressed considerable optimism when he wrote (p. 209): "For decades, the first step in prediction methodology was to avoid the curse. . . . . Recent work has shown that dimensionality can be a blessing." He continued (p. 209):

> "Reducing dimensionality reduces the amount of information available for prediction. The more predictor variables, the more information. There is also information in various combinations of the

predictor variables. Let's try going in the opposite direction: Instead of reducing dimensionality, increase it by adding many functions of the predictor variables."

Logically, Breiman's optimism cannot be completely realistic. The foundation of the curse of dimensionality is that, as the support of the covariate distribution grows, the probability P(x) of drawing any specified value of x into a random training sample of fixed size N necessarily decreases. Hence, the expected number of observations of y associated with any specified value of x necessarily falls. Indeed, it is zero when the covariate distribution is continuous rather than discrete. It follows that, as the support of the covariate distribution grows, prediction of y conditional on the specified value of x with a training sample of size N must increasingly rely on outcome data associated with other covariate values. However, observation of outcomes associated with other covariate values per se conveys no information about P(y|x) at the specified x-value of interest. These data become informative *only* given cross-covariate restrictions on the state space that relate P(y|x) to the conditional outcome distributions P(y|x′), $x' \neq x$.

Statistical theory has long sought to characterize how the statistical imprecision of conditional prediction varies with the maintained cross-covariate restrictions. Analysis is most straightforward in parametric modeling of conditional distributions, where the distributions $\{P(y|x), x \in X\}$, or at least the conditional expectations $\{E(y|x), x \in X\}$ are assumed to all be functions of a finite-dimensional parameter vector. Cross-covariate restrictions are formalized in the specification of the parameter space. Analysis is also relatively straightforward in the semiparametric approach to conditional classification called *maximum score* estimation in econometrics (Manski, 1975, 1985; Manski and Thompson, 1989) and *empirical risk minimization* in the statistical learning theory of Vapnik (1999, 2000). In both the parametric and semiparametric settings, researchers have mainly studied asymptotic questions of consistency and rates of convergence.

Analysis is more subtle, but still intuitive, in classical research on nonparametric regression, where $\{P(y|x), x \in X\}$ or $\{E(y|x), x \in X\}$ are assumed to vary in a smooth manner across suitably nearby values of x. The more recent body of research assuming some concept of *dimensional sparsity* is yet more subtle

in that it does not a priori fix the covariate space over which the conditional distribution or mean may vary. Instead, it constrains the number of elements of the covariate vector across which variation may occur.

The ML literature using deep neural networks as predictor functions replaces dimensional sparsity with the concept of compositional sparsity. Poggio et al. (2017) wrote this (p. 503):

> "The main message is that deep networks have the theoretical guarantee, which shallow networks do not have, that they can avoid the curse of dimensionality for an important class of problems, corresponding to compositional functions, i.e., functions of functions. An especially interesting subset of such compositional functions are hierarchically local compositional functions where all the constituent functions are local in the sense of bounded small dimensionality. The deep networks that can approximate them without the curse of dimensionality are of the deep convolutional type."

Poggio and Fraser (2024) defined compositional sparsity as follows (p. 1): "The property that a compositional function have *[sic]* "few" constituent functions, each depending on only a small subset of inputs."

Thus, if the function governing variation in the outcome y with covariates x happens to be compositionally sparse, then the claim is that a deep neural network will approximate it well. It should not be too surprising that a predictor function whose structure is similar to the unknown, true function of interest will approximate it well. Recognizing this led Shamir (2020), commenting on Schmidt-Hieber (2020), to assert (p. 1912): "Essentially, we have replaced a 'curse of dimensionality' effect with a 'curse of sparsity'."

A logical follow-up question concerns whether it is realistic in practice to assume that unknown functions governing variation in y with x are compositionally sparse. Poggio et al. (2017) asked this question and conjectured a partial answer, writing (p. 517):

> "This line of arguments defines a class of algorithms that is universal and can be optimal for a large set of problems. It does not however explain why problems encountered in practice should match this class of algorithms. Though we and others have argued that the explanation may be in either the physics or the neuroscience of the brain, these arguments…are not (yet) rigorous. Our conjecture is that compositionality is imposed by the wiring of our cortex and is reflected in language. Thus, compositionality of many computations on images may reflect the way we describe and think about them."

Poggio and Fraser (2024) went beyond this conjecture about two applications in which deep neural networks have been widely implemented—computer vision tasks and chat bots—concluding (p. 1):

> "Surprisingly, all functions that are efficiently Turing computable have a compositional sparse representation. Furthermore, deep networks that are also sparse can exploit this general property to avoid the "curse of dimensionality."

As we see it, compositional sparsity is an intriguing concept that warrants further study. In principle, SDT can specify that a state space is composed of compositionally sparse functions and evaluate the predictive performance of deep neural networks that have this structure. However, we expect that this implementation of SDT will be computationally challenging.

We also think it important to question the realism of assuming that unknown conditional means and other features of conditional distributions actually are compositionally sparse. Poggio et al. (2017) and Poggio and Fraser (2024) express optimism about this. We are less sure.

For example, a physical setting in which compositional sparsity does not hold is prediction of the trajectory in space-time of a mass subject to the gravitational attraction of multiple other masses. Basic physics, whether Newtonian or Einsteinian, predicts that the trajectory depends on the joint attraction of all masses in the universe. It is not the case that the trajectory is determined by a "few constituent functions, each depending on only a small subset of inputs."

4.2. Bounded-Variation Restrictions on the State Space

In some applications, it is realistic to bound the variation across covariate values of a conditional mean E(y|x) or other feature of P(y|x). Focusing on the problem of identification rather than decision making with sample data, Manski and Pepper (2000) initiated study of identification with *monotone instrumental variable* assumptions, which assume that a sub-vector of x is ordered, and that E(y|x) varies monotonically across these covariate values. More general *bounded variation* assumptions have been used to study conditional prediction of risk of illness by Manski (2018) and Li, Litvin, and Manski (2023). They have been used to study prediction of treatment response in criminal justice settings by Manski and Pepper (2013,

2018).

In settings where the outcome y is binary, bounded-variation assumptions have the form

(15) $P_s(y = 1|x') + \lambda_-(x, x') \leq P_s(y = 1|x) \leq P_s(y = 1|x') + \lambda_+(x, x'), \quad \text{all } s \in S.$

Here x is a covariate value of prediction interest, x′ is a different value, and $\lambda_-(x, x') \leq \lambda_+(x, x')$ are specified real numbers. Bounded-variation assumptions do not assert that $P_s(y = 1|x)$ varies locally smoothly with x, as in classical nonparametric regression analysis. Nor do they assume the types of cross-covariate invariance embedded in dimensional or compositional sparsity assumptions. They impose a distinct type of cross-covariate restriction on the state space.

As far as we are aware, the only research studying SDT with bounded-variation assumptions are Stoye (2012) and Manski (2023). Both focused on the maximum regret of SDFs in certain problems of choice between two treatments. Stoye (2012) obtained an analytical expression for the MMR decision function when $\lambda_-(x, x') = -\lambda_+(x, x') = \kappa$, where κ is a specified positive real number that does not vary with (x, x′).

Manski (2023) considered numerical computation of maximum regret for general assumptions of form (15), when loss is an extension of the MCR where the magnitude of the loss incurred with a prediction error varies with x and s. The analysis supposed that the DM uses a kernel method to predict $P^*(y = 1|x)$. A kernel estimate is a weighted average of the outcomes across different values of the covariate. The numerical analysis computed maximum regret using alternative weights. This enables one to minimize maximum regret within the class of kernel predictors.

It should be noted that maximum-regret analysis of the performance of kernel predictors differs considerably from classical analysis of kernel methods. The literature has mainly studied asymptotic properties of kernel estimates of mean regressions. Theorems typically assume that x is a real vector whose distribution has positive density in a neighborhood of a value of interest. Theorems generally assume that the conditional expectation E(y|x) varies smoothly with x in a local sense, such as being differentiable. They

typically do not impose bounded-variation assumptions that bound the difference between E(y|x') and E(y|x) at specified covariate values.

5. Conclusion

Whereas statisticians and econometricians have used frequentist or Bayesian statistical theory to propose and evaluate prediction methods, ML researchers have largely performed heuristic K-fold and CTF OOS validation. These types of OOS evaluation cannot yield generalizable lessons. We argue that ML researchers should instead perform comprehensive OOS evaluation using SDT. SDT is remote from the heuristic types of OOS evaluation currently practiced, but it is not remote conceptually. It performs OOS evaluation across all possible (1) training samples, (2) populations that may generate training data, and (3) populations of prediction interest.

SDT is simple in abstraction, but it is often computationally demanding to implement. Some progress has been made in tractable implementation of SDT when prediction accuracy is measured by mean square error or by misclassification rate. SDT also requires the practitioner to specify a decision criterion. Among the criteria that have been studied, we find minimax regret particularly appealing. We call on ML researchers to join with econometricians and statisticians in expanding the domain within which implementation of SDT is tractable.

References

Berger, J. (1985), *Statistical Decision Theory and Bayesian Analysis*, New York: Springer-Verlag.

Breiman, L. (2001), "Statistical Modeling: The Two Cultures," *Statistical Science*, 16, 199–231.

Dominitz, J. and C. Manski (2017), "More Data or Better Data? A Statistical Decision Problem," *Review of Economic Studies*, 84, 1583-1605.

Donoho, D. (2017), "50 Years of Data Science," *Journal of Computational and Graphical Statistics*, 26, 745-766.

Donoho, D., I. Johnstone, G. Kerkyacharian, and D. Picard (1995), "Wavelet Shrinkage: Asymptopia?" *Journal of the Royal Statistics Society, Series B*, 57, 301-369.
Efron, B. (2020), "Estimation, Prediction, and Attribution," *International Statistical Review*, 88, S28-S59.

Ferguson, T. (1967), *Mathematical Statistics: A Decision Theoretic Approach*, San Diego: Academic Press.

Hartford, J., G. Lewis, K. Leyton-Brown, and M. Taddy (2017), "Deep IV: A Flexible Approach for Counterfactual Prediction," In *Proceedings of the 34th International Conference on Machine Learning* 70:1414–23.https://proceedings.mlr.press/v70/hartford17a/hartford17a.pdf.

Hurwicz, L. (1951), "Some Specification Problems and Applications to Econometric Models," *Econometrica*, 19, 343-344.

Kitagawa, T. and A. Tetenov (2018), "Who Should be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, 86, 591-616.

Li, S., V. Litvin, and C. Manski (2023), "Partial Identification of Personalized Treatment Response with Trial-reported Analyses of Binary Subgroups," *Epidemiology*, 34, 319-324.

Manski, C. (1975), "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3, 205-228.

Manski, C. (1985), "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27, 313-333.

Manski, C. (1989), "Anatomy of the Selection Problem," *Journal of Human Resources*, 24, 343-360.

Manski, C. (2004), "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72, 221-246.

Manski, C. (2018), "Credible Ecological Inference for Medical Decisions with Personalized Risk Assessment," *Quantitative Economics*, 9, 541-569.

Manski, C. (2021), "Econometrics for Decision Making: Building Foundations Sketched by Haavelmo and Wald," *Econometrica*, 89, 2827-2853.

Manski, C. (2007a), *Identification for Prediction and Decision*, Cambridge, MA: Harvard University Press.

Manski, C. (2007b), "Minimax-Regret Treatment Choice with Missing Outcome Data," *Journal of Econometrics*, 139, 105-115.

Manski, C. (2019b), "Treatment Choice with Trial Data: Statistical Decision Theory Should Supplant Hypothesis Testing," *The American Statistician*, 73, 296-304.

Manski, C. (2021), "Econometrics for Decision Making: Building Foundations Sketched by Haavelmo and Wald," *Econometrica*, 89, 2827-2853.

Manski, C. (2023), "Probabilistic Prediction for Binary Treatment Choice: with focus on personalized medicine," *Journal of Econometrics*, 234, 647-663.

Manski, C. and J. Pepper (2000), "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68, 997-1010.

Manski, C. and J. Pepper (2013), "Deterrence and the Death Penalty: Partial Identification Analysis Using Repeated Cross Sections," *Journal of Quantitative Criminology*, 29, 123-141.

Manski, C. and J. Pepper (2018), "How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions," *Review of Economics and Statistics*, 100, 232-244.

Manski, C. and M. Tabord-Meehan (2017), "Wald MSE: Evaluating the Maximum MSE of Mean Estimates with Missing Data," *STATA Journal*, 17, 723-735.

Manski, C. and A. Tetenov (2016), "Sufficient Trial Size to Inform Clinical Practice," *Proceedings of the National Academy of Sciences*, 113, 10518-10523.

Manski, C. and A. Tetenov (2019), "Trial Size for Near-Optimal Choice Between Surveillance and Aggressive Treatment: Reconsidering MSLT-II," *The American Statistician*, 73, S1, 305-311.

Manski, C. and A. Tetenov (2023), "Statistical Decision Theory Respecting Stochastic Dominance," *The Japanese Economic Review*, 74, 447-469.

Manski, C. and S. Thompson (1989), "Estimation of Best Predictors of Binary Response," *Journal of Econometrics*, 40, 97-123.

Mitra, R., S. McGough, T. Chakraborti, T. et al. (2023), "Learning from data with structured missingness," *Nat Mach Intell* 5, 13–23. https://doi.org/10.1038/s42256-022-00596-z

Neyman, J., and E. Pearson (1928), "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference," *Biometrika*, 20A, 175-240, 263-294.

Neyman, J., and E. Pearson (1933), "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society of London*, Ser. A, 231, 289-337.

Poggio, T., and M. Fraser (2024), "Compositional Sparsity of Learnable Functions," CBMM Memo No. 145, February 8, 2024, available at https://cbmm.mit.edu/sites/default/files/publications/CBMM-Memo-145.pdf. Forthcoming in *Bulletin of the American Mathematical Society*.

Poggio, T., H. Mhaskar, L. Rosasco, B. Miranda, and Q. Liao (2017), "Why and When Can Deep-but Not Shallow-Networks Avoid the Curse of Dimensionality: A Review," *International Journal of Automation and Computing*, 14, 503-519.

Savage, L. (1951), "The Theory of Statistical Decision," *Journal of the American Statistical Association,* 46, 55-67.

Sawa, T. and T. Hiromatsu (1973), "Minimax Regret Significance Points for a Preliminary Test in Regression Analysis," *Econometrica*, 41, 1093-1101.

Shamir, O. (2020), "Discussion of: 'Nonparametric Regression Using Deep Neural Networks with RELU Activation Function," *The Annals of Statistics*, 48, 1911–1915.

Schmidt-Hieber, J. (2020), "Nonparametric Regression using Deep Neural Networks with ReLU Activation Function," *Annals of Statistics*, 48, 1875-1899.

Stoye, J. (2009), "Minimax Regret Treatment Choice with Finite Samples," *Journal of Econometrics*, 151, 70-81.

Stoye, J. (2012), "Minimax Regret Treatment Choice with Covariates or with Limited Validity of Experiments," *Journal of Econometrics*, 166, 138-156.

Taddy, M. (2019), "The Technological Elements of Artificial Intelligence," in A. Agrawal, J. Gans, and A. Goldfarb (editors), *The Economics of Artificial Intelligence: An Agenda*, Chicago: University of Chicago Press.

Vapnik, V. (1999), "An Overview of Statistical Learning Theory," *IEEE Transactions on Neural Networks*, 10, 988-999.

Vapnik, V. (2000), *The Nature of Statistical Learning Theory*, New York: Springer.

Wald, A. (1939), "Contribution to the Theory of Statistical Estimation and Testing Hypotheses," *Annals of Mathematical Statistics*, 10, 299-326.

Wald, A. (1945), "Statistical Decision Functions Which Minimize the Maximum Risk," *Annals of Mathematics*, 46, 265-280.

Wald A. (1950), *Statistical Decision Functions*, New York: Wiley.

Walley, P. (1991), *Statistical Reasoning with Imprecise Probabilities*, New York: Chapman and Hall.