Journal of Machine Learning Research 20 (2019) 1-81 Submitted 11/18; Revised 8/19; Published 12/19

All Models are Wrong, but *Many* are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously

Aaron Fisher AFISHE27@ALUMNI.JH.EDU
Takeda Pharmaceuticals
Cambridge, MA 02139, USA

Cynthia Rudin CYNTHIA@CS.DUKE.EDU
Departments of Computer Science and Electrical and Computer Engineering
Duke University
Durham, NC 27708, USA

Francesca Dominici FDOMINIC@HSPH.HARVARD.EDU
Department of Biostatistics
Harvard T.H. Chan School of Public Health
Boston, MA 02115, USA

(Authors are listed in order of contribution, with highest contribution listed first.)

Editor: Maya Gupta

Abstract

Variable importance (VI) tools describe how much covariates contribute to a prediction model's accuracy. However, important variables for one well-performing model (for example, a linear model $f(\mathbf{x}) = \mathbf{x}^T\beta$ with a fixed coefficient vector β) may be unimportant for another model. In this paper, we propose model class reliance (MCR) as the range of VI values across *all* well-performing model in a prespecified class. Thus, MCR gives a more comprehensive description of importance by accounting for the fact that many prediction models, possibly of different parametric forms, may fit the data well. In the process of deriving MCR, we show several informative results for permutation-based VI estimates, based on the VI measures used in Random Forests. Specifically, we derive connections between permutation importance estimates for a *single* prediction model, U-statistics, conditional variable importance, conditional causal effects, and linear model coefficients. We then give probabilistic bounds for MCR, using a novel, generalizable technique. We apply MCR to a public data set of Broward County criminal records to study the reliance of recidivism prediction models on sex and race. In this application, MCR can be used to help inform VI for unknown, proprietary models.

Keywords: Rashomon, permutation importance, conditional variable importance, U-statistics, transparency, interpretable models

1. Introduction

Variable importance (VI) tools describe how much a prediction model's accuracy depends on the information in each covariate. For example, in Random Forests, VI is measured by

the decrease in prediction accuracy when a covariate is permuted (Breiman, 2001; Breiman et al., 2001; see also Strobl et al., 2008; Altmann et al., 2010; Zhu et al., 2015; Gregorutti et al., 2015; Datta et al., 2016; Gregorutti et al., 2017). A similar "Perturb" VI measure has been used for neural networks, where noise is added to covariates (Recknagel et al., 1997; Yao et al., 1998; Scardi and Harding, 1999; Gevrey et al., 2003). Such tools can be useful for identifying covariates that must be measured with high precision, for improving the transparency of a "black box" prediction model (see also Rudin, 2019), or for determining what scenarios may cause the model to fail.

However, existing VI measures do not generally account for the fact that many prediction models may fit the data almost equally well. In such cases, the model used by one analyst may rely on entirely different covariate information than the model used by another analyst. This common scenario has been called the "Rashomon" effect of statistics (Breiman et al., 2001; see also Lecué, 2011; Statnikov et al., 2013; Tulabandhula and Rudin, 2014; Nevo and Ritov, 2017; Letham et al., 2016). The term is inspired by the 1950 Kurosawa film of the same name, in which four witnesses offer different descriptions and explanations for the same encounter. Under the Rashomon effect, how should analysts give comprehensive descriptions of the importance of each covariate? How well can one analyst recover the conclusions of another? Will the model that gives the best predictions necessarily give the most accurate interpretation?

To address these concerns, we analyze the *set* of prediction models that provide near-optimal accuracy, which we refer to as a *Rashomon set*. This approach stands in contrast to training to select a *single* prediction model, among a prespecified class of candidate models. Our motivation is that Rashomon sets (defined formally below) summarize the range of effective prediction strategies that an analyst might choose. Additionally, even if the candidate models do not contain the true data generating process, we may hope that some of these models function in similar ways to the data generating process. In particular, we may hope there exist well performing candidate models that place the same importance on a variable of interest as the underlying data generating process does. If so, then studying sets of well-performing models will allow us to deduce information about the data generating process.

Applying this approach to study variable importance, we define *model class reliance* (MCR) as the highest and lowest degree to which any well-performing model within a given class may rely on a variable of interest for prediction accuracy. Roughly speaking, MCR captures the range of explanations, or mechanisms, associated with well-performing models. Because the resulting range summarizes many prediction models simultaneously, rather a single model, we expect this range to be less affected by the choices that an individual analyst makes during the model-fitting process. Instead of reflecting these choices, MCR aims to reflect the nature of the prediction problem itself.

We make several, specific technical contributions in deriving MCR. First, we review a core measure of how much an individual prediction model relies on covariates of interest for its accuracy, which we call *model reliance* (MR). This measure is based on permutation importance measures for Random Forests (Breiman et al., 2001; Breiman, 2001), and can be expanded to describe conditional importance (see Section 8, as well as Strobl et al. 2008). We draw a connection between permutation-based importance estimates (MR) and U-statistics, which facilitates later theoretical results. Additionally, we derive connections

between MR, conditional causal effects, and coefficients for additive models. Expanding on MR, we propose MCR, which generalizes the definition of MR for a *class of models*. We derive finite-sample bounds for MCR, which motivate an intuitive estimator of MCR. Finally, we propose computational procedures for this estimator.

The tools we develop to study Rashomon sets are quite general, and can be used to make finite-sample inferences for arbitrary characteristics of well-performing models. For example, beyond describing variable importance, these tools can describe the range of risk predictions that well-fitting models assign to a particular covariate profile, or the variance of predictions made by well-fitting models. In some cases, these novel techniques may provide finite-sample confidence intervals (CIs) where none have previously existed (see Section 5).

MCR and the Rashomon effect become especially relevant in the context of criminal recidivism prediction. Proprietary recidivism risk models trained from criminal records data are increasingly being used in U.S. courtrooms. One concern is that these models may be relying on information that would otherwise be considered unacceptable (for example, race, sex, or proxies for these variables), in order to estimate recidivism risk. The relevant models are often proprietary, and cannot be studied directly. Still, in cases where the predictions made by these models are publicly available, it may be possible to identify alternative prediction models that are sufficiently similar to the proprietary model of interest.

In this paper, we specifically consider the proprietary model COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), developed by the company Northpointe Inc. (subsequently, in 2017, Northpointe Inc.,Courtview Justice Solutions Inc., and Constellation Justice Systems Inc. joined together under the name Equivant). Our goal is to estimate how much COMPAS relies on either race, sex, or proxies for these variables not measured in our data set. To this end, we apply a broad class of flexible, kernel-based prediction models to predict COMPAS score. In this setting, the MCR interval reflects the highest and lowest degree to which any prediction model in our class can rely on race and sex while still predicting COMPAS score relatively accurately. Equipped with MCR, we can relax the common assumption of being able to correctly specify the unknown model of interest (here, COMPAS) up to a parametric form. Instead, rather than assuming that the COMPAS model itself is contained in our class, we assume that our class contains at least one well-performing alternative model that relies on sensitive covariates to the same degree that COMPAS does. Under this assumption, the MCR interval will contain the VI value for COMPAS. Applying our approach, we find that race, sex, and their potential proxy variables, are likely not the dominant predictive factors in the COMPAS score (see analysis and discussion in Section 10).

The remainder of this paper is organized as follows. In Section 2 we introduce notation, and give a high level summary of our approach, illustrated with visualizations. In Sections 3 and 4 we formally present MR and MCR respectively, and derive theoretical properties of each. We also review related variable importance practices in the literature, such as retraining a model after removing one of the covariates. In Section 5, we discuss general applicability of our approach for determining finite-sample CIs for other problems. In Section 6, we present a general procedure for computing MCR. In Section 7, we give specific implementations of this procedure for (regularized) linear models, and linear models in a reproducing kernel Hilbert space. We also show that, for additive models, MR can be expressed in terms of the model's coefficients. In Section 8 we outline connections between

MR, causal inference, and conditional variable importance. In Section 9, we illustrate MR and MCR with a simulated toy example, to aid intuition. We also present simulation studies for the task of estimating MR for an unknown, underlying conditional expectation function, under misspecification. We analyze a well-known public data set on recidivism in Section 10, described above. All proofs are presented in the appendices.

2. Notation & Technical Summary

The label of "variable importance" measure has been broadly used to describe approaches for either inference (van der Laan, 2006; Díaz et al., 2015; Williamson et al., 2017) or prediction. While these two goals are highly related, we primarily focus on how much prediction models rely on covariates to achieve accuracy. We use terms such as "model reliance" rather than "importance" to clarify this context.

In order to evaluate how much prediction models rely on variables, we now introduce notation for random variables, data, classes of prediction models, and loss functions for evaluating predictions. Let $Z = (Y, X_1, X_2) \in \mathcal{Z}$ be a random variable with outcome $Y \in \mathcal{Y}$ and covariates $X = (X_1, X_2) \in \mathcal{X}$, where the covariate subsets $X_1 \in \mathcal{X}_1$ and $X_2 \in \mathcal{X}_2$ may each be multivariate. We assume that observations of Z are *iid*, that $n \geq 2$, and that solutions to arg min and arg max operations exist whenever optimizing over sets mentioned in this paper (for example, in Theorem 4, below). Our goal is to study how much different prediction models rely on X_1 to predict Y.

We refer to our data set as $\mathbf{Z} = \left[\begin{array}{cc} \mathbf{y} & \mathbf{X} \end{array} \right]$, a matrix composed of a n-length outcome vector $\mathbf{y}$ in the first column, and a $n \times p$ covariate matrix $\mathbf{X} = \left[\begin{array}{cc} \mathbf{X}_1 & \mathbf{X}_2 \end{array} \right]$ in the remaining columns. In general, for a given vector $\mathbf{v}$, let $\mathbf{v}_{[j]}$ denote its j^{th} element(s). For a given matrix $\mathbf{A}$, let $\mathbf{A}'$, $\mathbf{A}_{[i,\cdot]}$, $\mathbf{A}_{[\cdot,j]}$, and $\mathbf{A}_{[i,j]}$ respectively denote the transpose of $\mathbf{A}$, the i^{th} row(s) of $\mathbf{A}$, the j^{th} column(s) of $\mathbf{A}$, and the element(s) in the i^{th} row(s) and j^{th} column(s) of $\mathbf{A}$.

We use the term *model class* to refer to a prespecified subset $\mathcal{F} \subset \{f \mid f : \mathcal{X} \to \mathcal{Y}\}$ of the measurable functions from $\mathcal{X}$ to $\mathcal{Y}$. We refer to member functions $f \in \mathcal{F}$ as *prediction models,* or simply as *models.* Given a model f, we evaluate its performance using a nonnegative *loss function* $L : (\mathcal{F} \times \mathcal{Z}) \to \mathbb{R}_{\geq 0}$. For example, L may be the squared error loss $L_{\text{se}}(f, (y, x_1, x_2)) = (y - f(x_1, x_2))^2$ for regression, or the hinge loss $L_{\text{h}}(f, (y, x_1, x_2)) = (1 - yf(x_1, x_2))_+$ for classification. We use the term *algorithm* to refer to any procedure $\mathcal{A} : \mathcal{Z}^n \to \mathcal{F}$ that takes a data set as input and returns a model $f \in \mathcal{F}$ as output.

2.1. Summary of Rashomon Sets & Model Class Reliance

Many traditional statistical estimates come from descriptions of a *single,* fitted prediction model. In contrast, in this section, we summarize our approach for studying a *set* of near-optimal models. To define this set, we require a prespecified "reference" model, denoted by f_{ref}, to serve as a benchmark for predictive performance. For example, f_{ref} may come from a flowchart used to predict injury severity in a hospital's emergency room, or from another quantitative decision rule that is currently implemented in practice. Given a reference model f_{ref}, we define a *population ϵ-Rashomon set* as the subset of models

with expected loss no more than ϵ above that of f_{ref}. We denote this set as $\mathcal{R}(\epsilon) := \{f \in \mathcal{F} : \mathbb{E}L(f, Z) \leq \mathbb{E}L(f_{\text{ref}}, Z) + \epsilon\}$, where $\mathbb{E}$ denotes expectations with respect to the population distribution. This set can be thought of as representing models that might be arrived at due to differences in data measurement, processing, filtering, model parameterization, covariate selection, or other analysis choices (see Section 4).

Illustrations of Rashomon Sets & Model Class Reliance

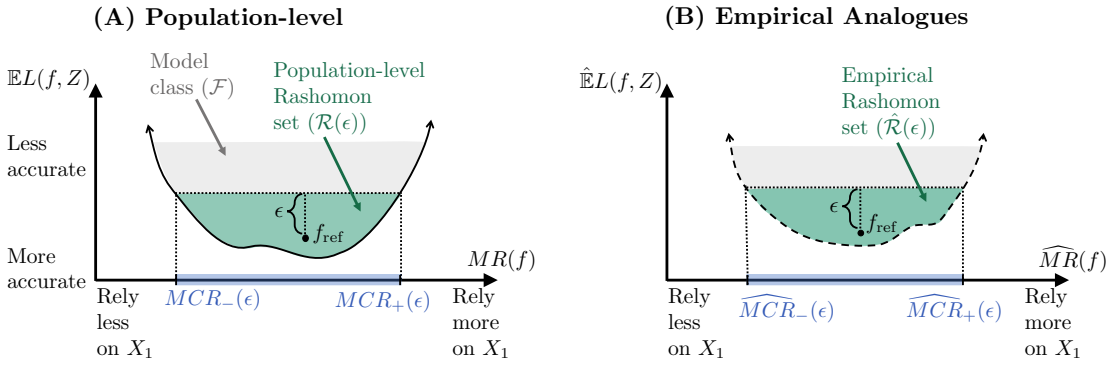

Figure 1: Rashomon sets and model class reliance – Panel (A) illustrates a hypothetical Rashomon set $\mathcal{R}(\epsilon)$, within a model class $\mathcal{F}$. The y-axis shows the expected loss of each model $f \in \mathcal{F}$, and the x-axis shows how much each model f relies on X_1 (defined formally in Section 3). Along the x-axis, the population-level MCR range is highlighted in blue, showing the values of MR corresponding to well-performing models (see Section 4). Panel (B) shows the in-sample analogue of Panel (A). Here, the y-axis denotes the in-sample loss, $\hat{\mathbb{E}}L(f, Z) := \frac{1}{n}\sum_{i=1}^{n} L(f, \mathbf{Z}_{[i,\cdot]})$; the x-axis shows the empirical model reliance of each model $f \in \mathcal{F}$ on X_1 (see Section 3); and the highlighted portion of the x-axis shows empirical MCR (see Section 4).

Figure 1-A illustrates a hypothetical example of a population ϵ-Rashomon set. Here, the y-axis shows the expected loss of each model $f \in \mathcal{F}$, and the x-axis shows how much each model relies on X_1 for its predictive accuracy. More specifically, given a prediction model f, the x-axis shows the percent increase in f's expected loss when noise is added to X_1. We refer to this measure as the *model reliance* (MR) of f on X_1, written informally as

$$MR(f) := \frac{\text{Expected loss of } f \text{ under noise}}{\text{Expected loss of } f \text{ without noise}}. \tag{2.1}$$

The added noise must satisfy certain properties, namely, it must render X_1 completely uninformative of the outcome Y, without altering the marginal distribution of X_1 (for details, see Section 3, as well as Breiman, 2001; Breiman et al., 2001).

Our central goal is to understand how much, or how little, models may rely on covariates of interest (X_1) while still predicting well. In Figure 1-A, this range of possible MR values

is shown by the highlighted interval along the x-axis. We refer to an interval of this type as a *population-level model class reliance* (MCR) range (see Section 4), formally defined as

$$[MCR_-(\epsilon),\ MCR_+(\epsilon)] := \left[\min_{f \in \mathcal{R}(\epsilon)} MR(f),\ \max_{f \in \mathcal{R}(\epsilon)} MR(f)\right]. \tag{2.2}$$

To estimate this range, we use empirical analogues of the population ϵ-Rashomon set, and of MR, based on observed data (Figure 1-B). We define an *empirical ϵ-Rashomon set* as the set of models with *in-sample* loss no more than ϵ above that of f_{ref}, and denote this set by $\hat{\mathcal{R}}(\epsilon)$. Informally, we define the *empirical* MR of a model f on X_1 as

$$\widehat{MR}(f) := \frac{\text{In-sample loss of } f \text{ under noise}}{\text{In-sample loss of } f \text{ without noise}}, \tag{2.3}$$

that is, the extent to which f appears to rely on X_1 in a given sample (see Section 3 for details). Finally, we define the *empirical model class reliance* as the range of empirical MR values corresponding to models with strong in-sample performance (see Section 4), formally written as

$$[\widehat{MCR}_-(\epsilon),\ \widehat{MCR}_+(\epsilon)] := \left[\min_{f \in \hat{\mathcal{R}}(\epsilon)} \widehat{MR}(f),\ \max_{f \in \hat{\mathcal{R}}(\epsilon)} \widehat{MR}(f)\right]. \tag{2.4}$$

In Figure 1-B, the above range is shown by the highlighted portion of the x-axis.

We make several technical contributions in the process of developing MCR.

- **Estimation of MR, and population-level MCR:** Given f, we show desirable properties of $\widehat{MR}(f)$ as an estimator of $MR(f)$, using results for U-statistics (Section 3.1 and Theorem 5). We also derive finite sample bounds for population-level MCR, some of which require a limit on the complexity of $\mathcal{F}$ in the form of a covering number. These bounds demonstrate that, under fairly weak conditions, empirical MCR provides a sensible estimate of population-level MCR (see Section 4 for details).

- **Computation of empirical MCR:** Although empirical MCR is fully determined given a sample, the minimization and maximization in Eq 2.4 require nontrivial computations. To address this, we outline a general optimization procedure for MCR (Section 6). We give detailed implementations of this procedure for cases when the model class $\mathcal{F}$ is a set of (regularized) linear regression models, or a set of regression models in a reproducing kernel Hilbert space (Section 7). The output of our proposed procedure is a closed-form, convex envelope containing $\mathcal{F}$, which can be used to approximate empirical MCR for any performance level ϵ (see Figure 2 for an illustration). Still, for complex model classes where standard empirical loss minimization is an open problem (for example, neural networks), computing empirical MCR remains an open problem as well.

- **Interpretation of MR in terms of model coefficients, and causal effects:** We show that MR for an additive model can be written as a function of the model's coefficients (Proposition 15), and that MR for a binary covariate X_1 can be written as a function of the conditional causal effects of X_1 on Y (Proposition 19).

- **Extensions to conditional importance:** We provide an extension of MR that is analogous to the notion of conditional importance (Strobl et al., 2008). This extension describes how much a model relies on the specific information in X_1 that cannot otherwise be gleaned from X_2 (Section 8.2).
- **Generalizations for Rashomon sets:** Beyond notions of variable importance, we also generalize our finite sample results for MCR to describe arbitrary characterizations of models in a population ϵ-Rashomon set. As we discuss in concurrent work (Coker et al., 2018), this generalization is analogous to the profile likelihood interval, and can, for example, be used to bound the range of risk predictions that well-performing prediction models may assign to a particular set of covariates (Section 5).

We begin in the next section by formally reviewing model reliance.

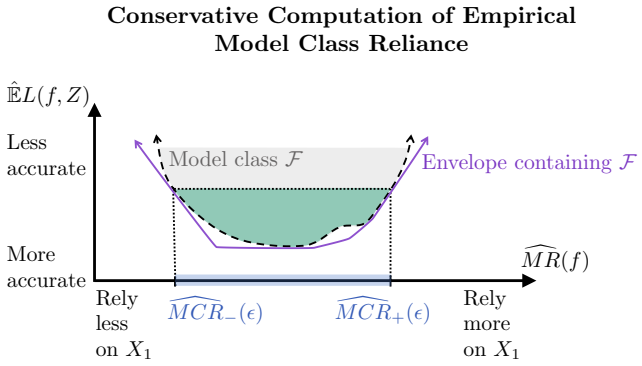

Figure 2: Illustration of output from our empirical MCR computational procedure – Our computation procedure produces a closed-form, convex envelope that contains $\mathcal{F}$ (shown above as the solid, purple line), which bounds empirical MCR for any value of ϵ (see Eq 2.4). The procedure works sequentially, tightening these bounds as much as possible near the ϵ value of interest (Section 6). The results from our data analysis (Figure 8) are presented in the same format as the above purple envelope.

3. Model Reliance

To formally describe how much the expected accuracy of a fixed prediction model f relies on the random variable X_1, we use the notion of a "switched" loss where X_1 is rendered uninformative. Throughout this section, we will treat f as a pre-specified prediction model of interest (as in Hooker, 2007). Let $Z^{(a)} = (Y^{(a)}, X_1^{(a)}, X_2^{(a)})$ and $Z^{(b)} = (Y^{(b)}, X_1^{(b)}, X_2^{(b)})$ be independent random variables, each following the same distribution as $Z = (Y, X_1, X_2)$. We define

$$e_{\text{switch}}(f) := \mathbb{E}L\{f, (Y^{(b)}, X_1^{(a)}, X_2^{(b)})\}$$

as representing the expected loss of model f across pairs of observations $(Z^{(a)}, Z^{(b)})$ in which the values of $X_1^{(a)}$ and $X_1^{(b)}$ have been switched. To see this interpretation of the above equation, note that we have used the variables $(Y^{(b)}, X_2^{(b)})$ from $Z^{(b)}$, but we have used the variable $X_1^{(b)}$ from an independent copy $Z^{(b)}$. This is why we say that $X_1^{(a)}$ and $X_1^{(b)}$ have been switched; the values of $(Y^{(b)}, X_1^{(a)}, X_2^{(b)})$ do not relate to each other as they would if they had been chosen together. An alternative interpretation of $e_{\text{switch}}(f)$ is as the expected loss of f when noise is added to X_1 in such a way that X_1 becomes completely uninformative of Y, but that the marginal distribution of X_1 is unchanged.

As a reference point, we compare $e_{\text{switch}}(f)$ against the standard expected loss when none of the variables are switched, $e_{\text{orig}}(f) := \mathbb{E}L(f, (Y, X_1, X_2))$. From these two quantities, we formally define *model reliance* (MR) as the ratio,

$$MR(f) := \frac{e_{\text{switch}}(f)}{e_{\text{orig}}(f)}, \tag{3.1}$$

as we alluded to in Eq 2.1. Higher values of $MR(f)$ signify greater reliance of f on X_1. For example, an $MR(f)$ value of 2 means that the model relies heavily on X_1, in the sense that its loss doubles when X_1 is scrambled. An $MR(f)$ value of 1 signifies no reliance on X_1, in the sense that the model's loss does not change when X_1 is scrambled. Models with reliance values strictly less than 1 are more difficult to interpret, as they rely less on the variable of interest than a random guess. Interestingly, it is possible to have models with reliance less than one. For instance, a model f' may satisfy $MR(f') < 1$ if it treats X_1 and Y as positively correlated when they are in fact negatively correlated. However, in many cases, the existence of a model $f' \in \mathcal{F}$ satisfying $MR(f') < 1$ implies the existence of another, better performing model $f'' \in \mathcal{F}$ satisfying $MR(f'') = 1$ and $e_{\text{orig}}(f'') \leq e_{\text{orig}}(f')$. That is, although models may exist with MR values less than 1, they will typically be suboptimal (see Appendix A.2).

Model reliance could alternatively be defined as a difference rather than a ratio, that is, as $MR_{\text{difference}}(f) := e_{\text{switch}}(f) - e_{\text{orig}}(f)$. In Appendix A.5, we discuss how many of our results remain similar under either definition.

3.1. Estimating Model Reliance with U-statistics, and Connections to Permutation-based Variable Importance

Given a model f and data set $\mathbf{Z} = \begin{bmatrix} \mathbf{y} & \mathbf{X} \end{bmatrix}$, we estimate $MR(f)$ by separately estimating the numerator and denominator of Eq 3.1. We estimate $e_{\text{orig}}(f)$ with the standard empirical loss,

$$\hat{e}_{\text{orig}}(f) := \frac{1}{n} \sum_{i=1}^{n} L\{f, (\mathbf{y}_{[i]}, \mathbf{X}_{1[i,\cdot]}, \mathbf{X}_{2[i,\cdot]})\}. \tag{3.2}$$

We estimate $e_{\text{switch}}(f)$ by performing a "switch" operation across all observed pairs, as in

$$\hat{e}_{\text{switch}}(f) := \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} L\{f, (\mathbf{y}_{[j]}, \mathbf{X}_{1[i,\cdot]}, \mathbf{X}_{2[j,\cdot]})\}. \tag{3.3}$$

Above, we have aggregated over all possible combinations of the observed values for (Y, X_2) and for X_1, excluding pairings that are actually observed in the original sample. If

the summation over all possible pairs (Eq 3.3) is computationally prohibitive due to sample size, another estimator of $e_{\text{switch}}(f)$ is

$$\hat{e}_{\text{divide}}(f) := \frac{1}{2\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \big[L\{f, (\mathbf{y}_{[i]}, \mathbf{X}_{1[i+\lfloor n/2 \rfloor, \cdot]}, \mathbf{X}_{2[i,\cdot]})\} \tag{3.4}$$

$$+ L\{f, (\mathbf{y}_{[i+\lfloor n/2 \rfloor]}, \mathbf{X}_{1[i,\cdot]}, \mathbf{X}_{2[i+\lfloor n/2 \rfloor,\cdot]})\} \big]. \tag{3.5}$$

Here, rather than summing over all pairs, we divide the sample in half. We then match the first half's values for (Y, X_2) with the second half's values for X_1 (Line 3.4), and vice versa (Line 3.5). All three of the above estimators (Eqs 3.2, 3.3 & 3.5) are unbiased for their respective estimands, as we discuss in more detail shortly.

Finally, we can estimate $MR(f)$ with the plug-in estimator

$$\widehat{MR}(f) := \frac{\hat{e}_{\text{switch}}(f)}{\hat{e}_{\text{orig}}(f)}, \tag{3.6}$$

which we define as the *empirical model reliance* of f on X_1. In this way, we formalize the empirical MR definition in Eq 2.3.

Again, our definition of empirical MR is very similar to the permutation-based variable importance approach of Breiman (2001), where Breiman uses a single random permutation and we consider all possible pairs. To compare these two approaches more precisely, let $\{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{n!}\}$ be a set of n-length vectors, each containing a different permutation of the set $\{1, \ldots, n\}$. The approach of Breiman (2001) is analogous to computing the loss $\sum_{i=1}^{n} L\{f, (\mathbf{y}_{[i]}, \mathbf{X}_{1[\boldsymbol{\pi}_{l[i]},\cdot]}, \mathbf{X}_{2[i,\cdot]})\}$ for a randomly chosen permutation vector $\boldsymbol{\pi}_l \in \{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{n!}\}$. Similarly, our calculation in Eq 3.3 is proportional to the sum of losses over all possible $(n!)$ permutations, excluding the n unique combinations of the rows of $\mathbf{X}_1$ and the rows of $\begin{bmatrix} \mathbf{X}_2 & \mathbf{y} \end{bmatrix}$ that appear in the original sample (see Appendix A.3). Excluding these observations is necessary to preserve the (finite-sample) unbiasedness of $\hat{e}_{\text{switch}}(f)$.

The estimators $\hat{e}_{\text{orig}}(f)$, $\hat{e}_{\text{switch}}(f)$ and $\hat{e}_{\text{divide}}(f)$ all belong to the well-studied class of U-statistics. Thus, under fairly minor conditions, these estimators are unbiased, asymptotically normal, and have finite-sample probabilistic bounds (Hoeffding, 1948, 1963; Serfling, 1980; see also DeLong et al., 1988 for an early use of U-statistics in machine learning, as well as caveats in Demler et al., 2012). To our knowledge, connections between permutation-based importance and U-statistics have not been previously established.

While the above results from U-statistics depend on the model f being fixed a priori, we can also leverage these results to create *uniform* bounds on the MR estimation error for all models in a sufficiently regularized class $\mathcal{F}$. We formally present this bound in Section 4 (Theorem 5), after introducing required conditions on model class complexity. The existence of this uniform bound implies that it is feasible to train a model and to evaluate its importance using the *same data*. This differs from the classical VI approach of Random Forests (Breiman, 2001), which avoids in-sample importance estimation. There, each tree in the ensemble is fit on a random subset of data, and VI for the tree is estimated using the held-out data. The tree-specific VI estimates are then aggregated to obtain a VI estimate for the overall ensemble. Although sample-splitting approaches such as this are helpful in many cases, the uniform bound for MR suggests that they are not strictly necessary, depending on the sample size and the complexity of $\mathcal{F}$.

3.2. Limitations of Existing Variable Importance Methods

Several common approaches for variable selection, or for describing relationships between variables, do not necessarily capture a variable's importance. Null hypothesis testing methods may identify a relationship, but do not describe the relationship's strength. Similarly, checking whether a variable is included by a sparse model-fitting algorithm, such as the Lasso (Hastie et al., 2009), does not describe the extent to which the variable is relied on. Partial dependence plots (Breiman et al., 2001; Hastie et al., 2009) can be difficult to interpret if multiple variables are of interest, or if the prediction model contains interaction effects.

Another common VI procedure is to run a model-fitting algorithm twice, first on all of the data, and then again after removing X_1 from the data set. The losses for the two resulting models are then compared to determine the importance, or "necessity," of X_1 (Gevrey et al., 2003). Because this measure is a function of two prediction models rather than one, it does not measure how much either individual model relies on X_1. We refer to this approach as measuring empirical *Algorithm Reliance* (AR) on X_1, as the model-fitting algorithm is the common attribute between the two models. Related procedures were proposed by Breiman et al. (2001); Breiman (2001), which measure the sufficiency of X_1.

As we discuss in Section 3.1, the permutation-based VI measure from RFs (Breiman, 2001; Breiman et al., 2001) forms the inspiration for our definition of MR. This RF VI measure has been the topic of empirical studies (Archer and Kimes, 2008; Calle and Urrea, 2010; Wang et al., 2016), and several variations of the measure have been proposed (Strobl et al., 2007, 2008; Altmann et al., 2010; Hapfelmeier et al., 2014). Mentch and Hooker (2016) use U-statistics to study predictions of ensemble models fit to subsamples, similar to the bootstrap aggregation used in RFs. Procedures related to "Mean Difference Impurity," another VI measure derived for RFs, have been studied theoretically by Louppe et al. (2013); Kazemitabar et al. (2017). All of this literature focuses on VI measures for RFs, for ensembles, or for individual trees. Our estimator for model reliance differs from the traditional RF VI measure (Breiman, 2001) in that we permute inputs to the overall model, rather than permuting the inputs to each individual ensemble member. Thus, our approach can be used generally, and is not limited to trees or ensemble models.

Outside of the context of RF VI, Zhu et al. (2015) propose an estimand similar to our definition of model reliance, and Gregorutti et al. (2015, 2017) propose an estimand analogous to $e_{\text{switch}}(f) - e_{\text{orig}}(f)$. These recent works focus on the model reliance of f on X_1 specifically when f is equal to the conditional expectation function of Y (that is, $f(x_1, x_2) = \mathbb{E}[Y|X_1 = x_1, X_2 = x_2]$). In contrast, we consider model reliance for arbitrary prediction models f. Datta et al. (2016) study the extent to which a model's predictions are expected to change when a subset of variables is permuted, regardless of whether the permutation affects a loss function L. These VI approaches are specific to a single prediction model, as is MR. In the next section, we consider a more general conception of importance: how much *any* model in a particular set may rely on the variable of interest.

4. Model Class Reliance

Like many statistical procedures, our MR measure (Section 3) produces a description of a *single* predictive model. Given a model with high predictive accuracy, MR describes how much the model's performance hinges on covariates of interest (X_1). However, there will often be many other models that perform similarly well, and that rely on X_1 to different degrees. With this notion in mind, we now study how much *any* well-performing model from a prespecified class $\mathcal{F}$ may rely on covariates of interest.

Recall from Section 2.1 that, in order to define a population ϵ-Rashomon set of near-optimal models, we must choose a "reference" model f_{ref} to serve as a performance benchmark. In order to discuss this choice, we now introduce more explicit notation for the population ϵ-Rashomon set, written as

$$\mathcal{R}(\epsilon, f_{\text{ref}}, \mathcal{F}) := \{f \in \mathcal{F} : e_{\text{orig}}(f) \leq e_{\text{orig}}(f_{\text{ref}}) + \epsilon\}. \tag{4.1}$$

Note that we write $\mathcal{R}(\epsilon, f_{\text{ref}}, \mathcal{F})$ and $\mathcal{R}(\epsilon)$ interchangeably when f_{ref} and $\mathcal{F}$ are clear from context. Similarly, we occasionally write empirical ϵ-Rashomon sets using the more explicit notation $\hat{\mathcal{R}}(\epsilon, f_{\text{ref}}, \mathcal{F}) := \{f \in \mathcal{F} : \hat{e}_{\text{orig}}(f) \leq \hat{e}_{\text{orig}}(f_{\text{ref}}) + \epsilon\}$, but typically abbreviate these sets as $\hat{\mathcal{R}}(\epsilon)$.

While f_{ref} could be selected by minimizing the in-sample loss, the theoretical study of $\mathcal{R}(\epsilon, f_{\text{ref}}, \mathcal{F})$ is simplified under the assumption that f_{ref} is prespecified. For example, f_{ref} may come from a flowchart used to predict injury severity in a hospital's emergency room, or from another quantitative decision rule that is currently implemented in practice. The model f_{ref} can also be selected using sample splitting. In some cases it may be desirable to fix f_{ref} equal to the best-in-class model $f^\star := \arg\min_{f \in \mathcal{F}} e_{\text{orig}}(f)$, but this is generally infeasible because $f^\star$ is unknown. Still, for any $f_{\text{ref}} \in \mathcal{F}$, the Rashomon set $\mathcal{R}(\epsilon, f_{\text{ref}}, \mathcal{F})$ defined using f_{ref} will always be conservative in the sense that it contains the Rashomon set $\mathcal{R}(\epsilon, f^\star, \mathcal{F})$ defined using $f^\star$.

We can now formalize our definitions of population-level MCR and empirical MCR by simply plugging in our definitions for $MR(f)$ and $\widehat{MR}(f)$ (Section 3) into Eqs 2.2 & 2.4 respectively. Studying population-level MCR (Eq 2.2) is the main focus of this paper, as it provides a more comprehensive view of importance than measures from a single model. If $MCR_+(\epsilon)$ is low, then *no* well-performing model in $\mathcal{F}$ places high importance on X_1, and X_1 can be discarded at low cost regardless of future modeling decisions. If $MCR_-(\epsilon)$ is large, then *every* well-performing model in $\mathcal{F}$ must rely substantially on X_1, and X_1 should be given careful attention during the modeling process. Here, $\mathcal{F}$ may itself consist of several parametric model forms (for example, all linear models and all decision tree models with less than 6 single-split nodes). We stress that the range $[MCR_-(\epsilon), MCR_+(\epsilon)]$ does not depend on the fitting algorithm used to select a model $f \in \mathcal{F}$. The range is valid for any algorithm producing models in $\mathcal{F}$, and applies for any $f \in \mathcal{F}$.

In the remainder of this section, we derive finite sample bounds for population-level MCR, from which we argue that empirical MCR provides reasonable estimates of population-level MCR (Section 4.1). In Appendix B.7 we consider an alternate formulation of Rashomon sets and MCR where we replace the relative loss threshold in the definition of $\mathcal{R}(\epsilon)$ with an absolute loss threshold. This alternate formulation can be similar in practice, but still requires the specification of a reference function f_{ref} to ensure that $\mathcal{R}(\epsilon)$ and $\hat{\mathcal{R}}(\epsilon)$ are nonempty.

4.1. Motivating Empirical Estimators of MCR by Deriving Finite-sample Bounds

In this section we derive finite-sample, probabilistic bounds for $MCR_+(\epsilon)$ and $MCR_-(\epsilon)$. Our results imply that, under minimal assumptions, $\widehat{MCR}_+(\epsilon)$ and $\widehat{MCR}_-(\epsilon)$ are respectively within a neighborhood of $MCR_+(\epsilon)$ and $MCR_-(\epsilon)$ with high probability. However, the weakness of our assumptions (which are typical for statistical-learning-theoretic analysis) renders the width of our resulting CIs to be impractically large, and so we use these results only to show conditions under which $\widehat{MCR}_+(\epsilon)$ and $\widehat{MCR}_-(\epsilon)$ form sensible point estimates. In Sections 9.1 & 10, below, we apply a bootstrap procedure to account for sampling variability.

To derive these results we introduce three bounded loss assumptions, each of which can be assessed empirically. Let $b_{\text{orig}}, B_{\text{ind}}, B_{\text{ref}}, B_{\text{switch}} \in \mathbb{R}$ be known constants.

Assumption 1 *(Bounded individual loss) For a given model $f \in \mathcal{F}$, assume that $0 \leq L(f,(y,x_1,x_2)) \leq B_{ind}$ for any $(y,x_1,x_2) \in (\mathcal{Y} \times \mathcal{X}_1 \times \mathcal{X}_2)$.*

Assumption 2 *(Bounded relative loss) For a given model $f \in \mathcal{F}$, assume that $|L(f,(y,x_1,x_2)) - L(f_{ref},(y,x_1,x_2))| \leq B_{ref}$ for any $(y,x_1,x_2) \in \mathcal{Z}$.*

Assumption 3 *(Bounded aggregate loss) For a given model $f \in \mathcal{F}$, assume that $\mathbb{P}\{0 < b_{orig} \leq \hat{e}_{orig}(f)\} = \mathbb{P}\{\hat{e}_{switch}(f) \leq B_{switch}\} = 1$.*

Each assumption is a property of a specific model $f \in \mathcal{F}$. The notation B_{ind} and B_{ref} refer to bounds for any individual observation, and the notation b_{orig} and B_{switch} refer to bounds on the aggregated loss L in a sample. These boundedness assumptions are central to our finite sample guarantees, shown below.

Crucially, loss functions L that are unbounded in general may be used so long as $L(f,(y,x_1,x_2))$ is bounded on a particular domain. For example, the squared-error loss can be used if $\mathcal{Y}$ is contained within a known range, and predictions $f(x_1,x_2)$ are contained within the same range for $(x_1,x_2) \in \mathcal{X} \times \mathcal{X}_2$. We give example methods of determining B_{ind} in Sections 7.3.2 & 7.4.2. For Assumption 3, we can approximate b_{orig} by training a highly flexible model to the data, and setting b_{orig} equal to half (or any positive fraction) of the resulting cross-validated loss. To determine B_{switch} we can simply set $B_{\text{switch}} = B_{\text{ind}}$, although this may be conservative. For example, in the case of binary classification models for non-separable groups (see Section 9.1), no linear classifier can misclassify all observations, particularly after a covariate is permuted. Thus, it must hold that $B_{\text{ind}} > B_{\text{switch}}$. Similarly, if f_{ref} satisfies Assumption 1, then B_{ref} may be conservatively set equal to B_{ind}. If model reliance is redefined as a difference rather than a ratio, then a similar form of the results in this section will apply without Assumption 3 (see Appendix A.5).

Based on these assumptions, we can create a finite-sample upper bound for $MCR_+(\epsilon)$ and lower bound for $MCR_-(\epsilon)$. In other words, we create an "outer" bound that contains the interval $[MCR_-(\epsilon), MCR_+(\epsilon)]$ with high probability.

Theorem 4 *("Outer" MCR Bounds) Given a constant $\epsilon \geq 0$, let $f_{+,\epsilon} \in \arg\max_{\mathcal{R}(\epsilon)} MR(f)$ and $f_{-,\epsilon} \in \arg\min_{\mathcal{R}(\epsilon)} MR(f)$ be prediction models that attain the highest and lowest model*

reliance among models in $\mathcal{R}(\epsilon)$*. If* $f_{+,\epsilon}$ *and* $f_{-,\epsilon}$ *satisfy Assumptions 1, 2 & 3, then*

$$\mathbb{P}\left(MCR_+(\epsilon) > \widehat{MCR}_+(\epsilon_{out}) + \mathcal{Q}_{out}\right) \leq \delta, \text{ and} \tag{4.2}$$

$$\mathbb{P}\left(MCR_-(\epsilon) < \widehat{MCR}_-(\epsilon_{out}) - \mathcal{Q}_{out}\right) \leq \delta, \tag{4.3}$$

where $\epsilon_{out} := \epsilon + 2B_{ref}\sqrt{\frac{\log(3\delta^{-1})}{2n}}$*, and* $\mathcal{Q}_{out} := \frac{B_{switch}}{b_{orig}} - \frac{B_{switch} - B_{ind}\sqrt{\frac{\log(6\delta^{-1})}{n}}}{b_{orig} + B_{ind}\sqrt{\frac{\log(6\delta^{-1})}{2n}}}$.

Eq 4.2 states that, with high probability, $MCR_+(\epsilon)$ is no higher than $\widehat{MCR}_+(\epsilon_{\text{out}})$ added to an error term $\mathcal{Q}_{\text{out}}$. As n increases, ϵ_{out} approaches ϵ and $\mathcal{Q}_{\text{out}}$ approaches zero. One practical implication is that, roughly speaking, if $\widehat{MCR}_+(\epsilon) \approx \widehat{MCR}_+(\epsilon_{\text{out}})$, then the empirical estimator $\widehat{MCR}_+(\epsilon)$ is unlikely to substantially underestimate $MCR_+(\epsilon)$. By similar reasoning, we can conclude from Eq 4.3 that if $\widehat{MCR}_-(\epsilon) \approx \widehat{MCR}_-(\epsilon_{\text{out}})$, then $\widehat{MCR}_-(\epsilon)$ is unlikely to substantially overestimate $MCR_-(\epsilon)$. By setting $\epsilon = 0$, Theorem 4 can also be used to create a finite-sample bound for the reliance of the unique (unknown) best-in-class model on X_1 (see Corollary 22 in Appendix A.4), although describing individual models is not the main focus of this paper.

We provide a visual illustration of Theorem 4 in Figure 3. A brief sketch of the proof is as follows. First, we enlarge the empirical ϵ-Rashomon set by increasing ϵ to ϵ_{out}, such that, by Hoeffding's inequality, $f_{+,\epsilon} \in \hat{\mathcal{R}}(\epsilon_{\text{out}})$ with high probability. When $f_{+,\epsilon} \in \hat{\mathcal{R}}(\epsilon_{\text{out}})$, we know that $\widehat{MR}(f_{+,\epsilon}) \leq \widehat{MCR}_+(\epsilon_{\text{out}})$ by the definition of $\widehat{MCR}_+(\epsilon_{\text{out}})$. Next, the term $\mathcal{Q}_{\text{out}}$ leverages finite-sample results for U-statistics to account for estimation error of $MR(f_{+,\epsilon}) = MCR_+(\epsilon)$ when using the estimator $\widehat{MR}(f_{+,\epsilon})$. Thus, we can relate $\widehat{MR}(f_{+,\epsilon})$ to both $\widehat{MCR}_+(\epsilon_{\text{out}})$ and $MCR_+(\epsilon)$ in order to obtain Eq 4.2. Similar steps can be applied to obtain Eq 4.3.

The bounds in Theorem 4 naturally account for potential overfitting without an explicit limit on model class complexity (such as a covering number, Rademacher complexity, or VC dimension). Instead, these bounds depend on being able to fully optimize MR across sets in the form of $\hat{\mathcal{R}}(\epsilon)$. If we allow our model class $\mathcal{F}$ to become more flexible, then the size of $\hat{\mathcal{R}}(\epsilon)$ will also increase. Because the bounds in Theorem 4 result from optimizing over $\hat{\mathcal{R}}(\epsilon)$, increasing the size of $\hat{\mathcal{R}}(\epsilon)$ results in wider, more conservative bounds. In this way, Eqs 4.2 and 4.3 implicitly capture model class complexity.

So far, Theorem 4 lets us bound the range of MR values corresponding to models that predict well, but it does not tell us whether these bounds are actually attained. Similarly, we can conclude from Theorem 4 that $[MCR_-(\epsilon), MCR_+(\epsilon)]$ is unlikely to exceed the estimated range $[\widehat{MCR}_-(\epsilon), \widehat{MCR}_+(\epsilon)]$ by a substantial margin, but we cannot determine whether this estimated range is unnecessarily wide. For example, consider the models that drive the $\widehat{MCR}_+(\epsilon)$ estimator: the models with strong in-sample accuracy, and high empirical reliance on X_1. These models' in-sample performance could merely be the result of overfitting, in which case they do not tell us direct information about $\mathcal{R}(\epsilon)$. Alternatively, even if all of these models truly do perform well on expectation (that is, even if they are contained in $\mathcal{R}(\epsilon)$), the model with the highest empirical reliance on X_1 may merely be the model for which our empirical MR estimate contains the most error. Either of these scenarios can cause $\widehat{MCR}_+(\epsilon)$ to be unnecessarily high, relative to $MCR_+(\epsilon)$.

Fortunately, both problematic scenarios are solved by requiring a limit on the complexity of $\mathcal{F}$. We propose a complexity measure in the form of a covering number, which allows us control a worst case scenario of either overfitting or MR estimation error. Specifically, we define the set of functions $\mathcal{G}_r$ as an *r-margin-expectation-cover* if for any $f \in \mathcal{F}$ and any distribution D, there exists $g \in \mathcal{G}_r$ such that

$$\mathbb{E}_{Z\sim D} \left| L\left(f, Z\right) - L\left(g, Z\right)\right| \leq r. \tag{4.4}$$

We define the *covering number* $\mathcal{N}(\mathcal{F}, r)$ to be the size of the smallest r-margin-expectation-cover for $\mathcal{F}$. In general, we use $\mathbb{P}_{V\sim D}$ and $\mathbb{E}_{V\sim D}$ to denote probabilities and expectations with respect to a random variable V following the distribution D. We abbreviate these quantities accordingly when V or D are clear from context, for example, as $\mathbb{P}_D$, $\mathbb{P}_V$, or simply $\mathbb{P}$. Unless otherwise stated, all expectations and probabilities are taken with respect to the (unknown) population distribution.

We first show that this complexity measure allows us to control the worst case MR estimation error, that is, the covering number $\mathcal{N}(\mathcal{F}, r)$ provides a uniform bound on the error of $\widehat{MR}(f)$ for all $f \in \mathcal{F}$.

Theorem 5 *(Uniform bound for $\widehat{MR}$) Given $r > 0$, if Assumptions 1 and 3 hold for all $f \in \mathcal{F}$, then*

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}} \left|\widehat{MR}(f) - MR(f)\right| > q(\delta, r, n)\right] \leq \delta,$$

where

$$q(\delta, r, n) := \frac{B_{switch}}{b_{orig}} - \frac{B_{switch} - \left\{B_{ind}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}\left(\mathcal{F}, r\sqrt{2}\right))}{n}} + 2r\sqrt{2}\right\}}{b_{orig} + \left\{B_{ind}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}(\mathcal{F}, r))}{2n}} + 2r\right\}}. \tag{4.5}$$

Theorem 5 states that, with high probability, the largest possible estimation error for $MR(f)$ across all models in $\mathcal{F}$ is bounded by $q(\delta, r, n)$, which can be made arbitrarily small by increasing n and decreasing r. As we noted in Section 3.1, this means that it is possible to train a model and estimate its reliance on variables without using sample-splitting.

The covering number $\mathcal{N}(\mathcal{F}, r)$ can also be used to limit the extent of overfitting (see Appendix B.5.1). As a result, it is possible to set an in-sample performance threshold low enough so that it will only be met by models with strong expected performance (that is, by models truly within $\mathcal{R}(\epsilon)$). To implement this idea of a stricter performance threshold, we contract the empirical ϵ-Rashomon set by subtracting a buffer term from ϵ. This requires that we generalize the definition of an empirical ϵ-Rashomon set to $\hat{\mathcal{R}}(\epsilon, f_{\text{ref}}, \mathcal{F}) := \{f_{\text{ref}}\} \cup \{f \in \mathcal{F} : \hat{e}_{\text{orig}}(f) \leq \hat{e}_{\text{orig}}(f_{\text{ref}}) + \epsilon\}$ for $\epsilon \in \mathbb{R}$, where the explicit inclusion of f_{ref} now ensures that $\hat{\mathcal{R}}(\epsilon, f_{\text{ref}}, \mathcal{F})$ is nonempty, even for $\epsilon < 0$. As before, we typically omit the notation f_{ref} and $\mathcal{F}$, writing $\hat{\mathcal{R}}(\epsilon)$ instead.

We are now prepared to answer the questions of whether the bounds from Theorem 4 are actually attained, and of whether the estimated range $[\widehat{MCR}_-(\epsilon), \widehat{MCR}_+(\epsilon)]$ is unnecessarily wide. Our answer comes in the form of an upper bound on $MCR_-(\epsilon)$, and a lower bound on $MCR_+(\epsilon)$.

Theorem 6 *("Inner" MCR Bounds) Given constants $\epsilon \geq 0$ and $r > 0$, if Assumptions 1, 2 and 3 hold for all $f \in \mathcal{F}$, and then*

$$\mathbb{P}\left(MCR_+(\epsilon) < \widehat{MCR}_+(\epsilon_{in}) - \mathcal{Q}_{in}\right) \leq \delta, \text{ and} \tag{4.6}$$

$$\mathbb{P}\left(MCR_-(\epsilon) > \widehat{MCR}_-(\epsilon_{in}) + \mathcal{Q}_{in}\right) \leq \delta, \tag{4.7}$$

where $\epsilon_{in} := \epsilon - 2B_{ref}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}(\mathcal{F},r))}{2n}} - 2r$, and $\mathcal{Q}_{in} = q\left(\frac{\delta}{2}, r, n\right)$, as defined in Eq 4.5.

Theorem 6 can allow us to infer an "inner" bound that is contained within the interval $[MCR_-(\epsilon), MCR_+(\epsilon)]$ with high probability. In Figure 3, we illustrate the result of Theorem 6, and give a sketch of the proof. This proof follows a similar structure to that of Theorem 4, but incorporates Theorem 5's uniform bound on MR estimation error ($\mathcal{Q}_{\text{in}}$ term), as well as an additional uniform bound on the probability that any model has in-sample loss too far from its expected loss (ϵ_{in} term).

A practical implication of Theorem 6 is that, roughly speaking, if $\widehat{MCR}_+(\epsilon_{\text{in}}) \approx \widehat{MCR}_+(\epsilon)$, then it is unlikely for the empirical estimator $\widehat{MCR}_+(\epsilon)$ to substantially underestimate $MCR_+(\epsilon)$. Taken together with Theorem 4, we can conclude that, if $\widehat{MCR}_+(\epsilon_{\text{in}}) \approx \widehat{MCR}_+(\epsilon_{\text{out}})$, then the estimator $\widehat{MCR}_+(\epsilon)$ is unlikely either to overestimate or to underestimate $MCR_+(\epsilon)$ by very much. In large samples, it may be plausible to expect the condition $\widehat{MCR}_+(\epsilon_{\text{in}}) \approx \widehat{MCR}_+(\epsilon_{\text{out}})$ to hold, since ϵ_{in} and ϵ_{out} both approach ϵ as n increases. In the same way, if $\widehat{MCR}_-(\epsilon_{\text{in}}) \approx \widehat{MCR}_-(\epsilon_{\text{out}})$, we can conclude from Eqs 4.3 & 4.7 that the empirical estimator $\widehat{MCR}_-(\epsilon)$ is unlikely to either overestimate or underestimate $MCR_-(\epsilon)$ by very much. For this reason, we argue that $\widehat{MCR}_-(\epsilon)$ and $\widehat{MCR}_+(\epsilon)$ form sensible estimates of population-level MCR – each is contained within a neighborhood of its respective estimand, with high probability. The secondary x-axis of Figure 3 gives an illustration of this argument.

5. Extensions of Rashomon Sets Beyond Variable Importance

In this section we generalize the Rashomon set approach beyond the study of MR. In Section 5.1, we create finite-sample CIs for other summary characterizations of near-optimal, or best-in-class models. The generalization also helps to illustrate a core aspect of the argument underlying Theorem 4: models with near-optimal performance in the population tend to have relatively good performance in random samples.

In Section 5.2, we review existing literature on near-optimal models.

5.1. Finite-sample Confidence Intervals from Rashomon Sets

Rather than describing how much a model relies on X_1, here we assume the analyst is interested in an arbitrary characteristic of a model. We denote this characteristic of interest as $\phi : \mathcal{F} \to \mathbb{R}$. For example, if $f_{\boldsymbol{\beta}}$ is the linear model $f_{\boldsymbol{\beta}}(x) = x'\boldsymbol{\beta}$, then ϕ may be defined as the norm of the associated coefficient vector (that is, $\phi(f_{\boldsymbol{\beta}}) = \|\boldsymbol{\beta}\|_2^2$) or the prediction $f_{\boldsymbol{\beta}}$ would assign given a specific covariate profile x_{new} (that is, $\phi(f_{\boldsymbol{\beta}}) = f_{\boldsymbol{\beta}}(x_{\text{new}})$).

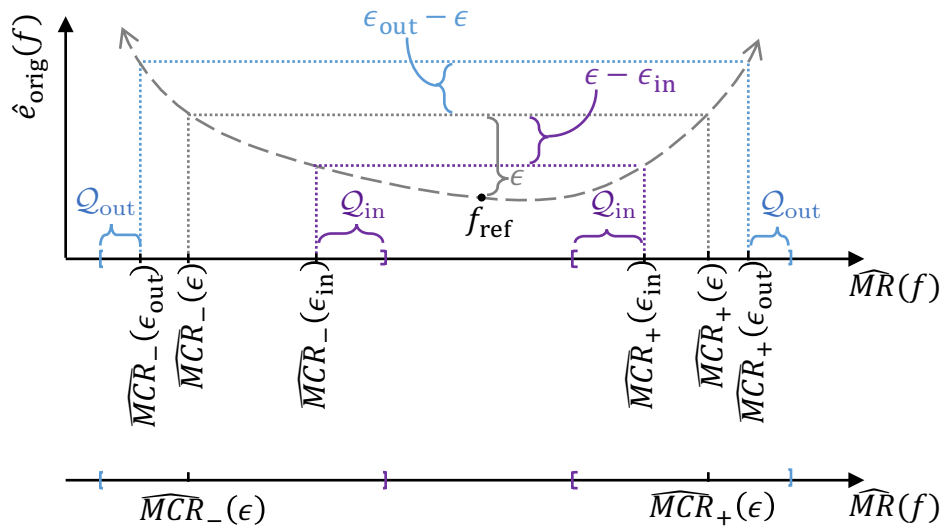

Figure 3: Illustration of terms in Theorems 4 and 6 – Above we show the relation between empirical MR (x-axis) and empirical loss (y-axis) for models f in a hypothetical model class $\mathcal{F}$. We mark f_{ref} by the black point. For each possible model reliance value $r \geq 0$, the curved, dashed line shows the lowest possible empirical loss for a function in $f \in \mathcal{F}$ satisfying $\widehat{MR}(f) = r$. The set $\hat{\mathcal{R}}(\epsilon)$ contains all models in $\mathcal{F}$ within the dotted gray lines. To create the bounds from Theorem 4, we expand the empirical ϵ-Rashomon set by increasing ϵ to ϵ_{out}, such that $f_{+,\epsilon}$ (or $f_{-,\epsilon}$) is contained in $\hat{\mathcal{R}}(\epsilon_{\text{out}})$ with high probability. We then add (or subtract) $\mathcal{Q}_{\text{out}}$ to account for estimation error of $\widehat{MR}(f_{+,\epsilon})$ (or $\widehat{MR}(f_{-,\epsilon})$). These steps are illustrated above in blue, with the final bounds shown by the blue bracket symbols along the x-axis. To create the bounds for $MCR_+(\epsilon)$ (and $MCR_-(\epsilon)$) in Theorem 6, we constrict the empirical ϵ-Rashomon set by decreasing ϵ to ϵ_{in}, such that all models with high expected loss are simultaneously excluded from $\hat{\mathcal{R}}(\epsilon_{\text{in}})$ with high probability. We then subtract (or add) $\mathcal{Q}_{\text{in}}$ to simultaneously account for MR estimation error for models in $\hat{\mathcal{R}}(\epsilon_{\text{in}})$. These steps are illustrated above in purple, with the final bounds shown by the purple bracket symbols along the x-axis. For emphasis, below this figure we show a copy of the x-axis with selected annotations, from which it is clear that $\widehat{MCR}_-(\epsilon)$ and $\widehat{MCR}_+(\epsilon)$ are always within the bounds produced by Theorems 4 and 6. With high probability, $\widehat{MCR}_-(\epsilon)$ and $\widehat{MCR}_+(\epsilon)$ are within a neighborhood of $MCR_-(\epsilon)$ and $MCR_+(\epsilon)$ respectively.

Given a descriptor ϕ, we now show a general result that allows creation of finite-sample CIs for the best performing models $\mathcal{R}(\epsilon)$. The resulting CIs are themselves based on empirical Rashomon sets.

Proposition 7 *(Finite sample CIs from Rashomon sets) Let* $\epsilon' := \epsilon + 2B_{ref}\sqrt{\frac{\log(2\delta^{-1})}{2n}}$*, let* $\hat{\phi}_-(\epsilon') := \min_{f \in \hat{\mathcal{R}}(\epsilon')} \phi(f)$ *and let* $\hat{\phi}_+(\epsilon') := \max_{f \in \hat{\mathcal{R}}(\epsilon')} \phi(f)$*.*

If Assumption 2 holds for all $f \in \mathcal{R}(\epsilon)$*, then*

$$\mathbb{P}\left[\{\phi(f) : f \in \mathcal{R}(\epsilon)\} \subseteq \left[\hat{\phi}_-(\epsilon'), \hat{\phi}_+(\epsilon')\right]\right] \geq 1 - \delta.$$

Proposition 7 generates a finite-sample CI for the range of values $\phi(f)$ corresponding to well-performing models, $\{\phi(f) : f \in \mathcal{R}(\epsilon)\}$. This CI, denoted by $\left[\hat{\phi}_-(\epsilon'), \hat{\phi}_+(\epsilon')\right]$, can itself be interpreted as the range of values $\phi(f)$ corresponding to models f with empirical loss not substantially above that of f_{ref}. Thus, the interval has both a rigorous coverage rate and a coherent in-sample interpretation. The proof of Proposition 7 uses Hoeffding's inequality to show that models in $\mathcal{F}$ are contained in $\hat{\mathcal{R}}(\epsilon')$ with high probability, that is, that models with good expected performance tend to perform well in random samples.

An immediate corollary of Proposition 7 is that we can generate finite-sample CIs for all best-in-class models $f^\star \in \arg\min_{f \in \mathcal{F}} \mathbb{E}L(f, Z)$ by setting $\epsilon = 0$. This corollary can be further strengthened if a single model $f^\star$ is assumed to uniquely minimize $\mathbb{E}L(f, Z)$ over $f \in \mathcal{F}$ (see Appendix B.6).

Note that Proposition 7 implicitly assumes that $\phi(f)$ can be determined exactly for any model $f \in \mathcal{F}$, in order for the interval $\left[\hat{\phi}_-(\epsilon'), \hat{\phi}_+(\epsilon')\right]$ to be precisely determined. This assumption does not hold, for example, if $\phi(f) = MR(f)$, or if $\phi(f) = \text{Var}\{f(X_1, X_2)\}$, as these quantities depend on both f and the (unknown) population distribution. In such cases, an additional correction factor must be incorporated to account for estimation error of $\phi(f)$, such as the $\mathcal{Q}_{\text{out}}$ term in Theorem 4.

In concurrent work, Coker et al. (2018) show that profile likelihood intervals take the same form as the interval $[\hat{\phi}_-(\epsilon'), \hat{\phi}_+(\epsilon')]$ in Proposition 7. This means that a profile likelihood interval can also be expressed by minimizing and maximizing over an empirical Rashomon set. More specifically, consider the case where the loss function L is the negative of the *known* log likelihood function, and where f_{ref} is the maximum likelihood estimate of the "true model," which in this case is $f^\star$. If additional minor assumptions are met (see Appendix A.6 for details), then the $(1-\delta)$-level profile likelihood interval for $\phi(f^\star)$ is equal to $[\hat{\phi}_-(\frac{\chi_{1,1-\delta}}{2n}), \hat{\phi}_+(\frac{\chi_{1,1-\delta}}{2n})]$, where $\hat{\phi}_-$ and $\hat{\phi}_+$ are defined as in Proposition 7, and $\chi_{1,1-\delta}$ is the $1-\delta$ percentile of a chi-square distribution with 1 degree of freedom.

Relative to a profile likelihood approach, the advantage of Proposition 7 is that it does not require asymptotics, it does not require that the likelihood be known up to a parametric form, and it can be extended to study the *set* of near-optimal prediction models $\mathcal{R}(\epsilon)$, rather than a single, potentially misspecified prediction model $f^\star$. This is especially useful when different near-optimal models accurately describe different aspects of the underlying data generating process, but none capture it completely. The disadvantage of Proposition 7 is that the required performance threshold of $\epsilon' = \epsilon + 2B_{\text{ref}}\sqrt{\frac{\log(2\delta^{-1})}{2n}}$ decreases more slowly than the performance threshold of $\frac{\chi_{1,1-\delta}}{2n}$ required in a profile likelihood interval. Because

our results from Section 4.1 carry a similar disadvantage, we use these results primarily to motivate point estimates describing the Rashomon set $\mathcal{R}(\epsilon)$.

Still, it is worth emphasizing the generality of Proposition 7. Through this result, Rashomon sets allow us to reframe a wide set of finite-sample inference problems as in-sample optimization problems. The implied CIs are not necessarily in closed form, but the approach still opens an exciting pathway for deriving non-asymptotic results. For example, they imply that existing methods for profile likelihood intervals might be able to be reapplied to achieve finite-sample results. For highly complex model classes where profile likelihoods are difficult to compute, such as neural networks or random forests, approximate inference is sometimes achieved via approximate optimization procedures (for example, Markov chain Monte Carlo for Bayesian additive regression trees, in Chipman et al., 2010). Proposition 7 shows that similar approximate optimization methods *could be repurposed to establish approximate, finite-sample inferences for the same model classes.*

5.2. Related Literature on the Rashomon Effect

Breiman et al. (2001) introduced the "Rashomon effect" of statistics as a problem of ambiguity: if many models fit the data well, it is unclear which model we should try to interpret. Breiman suggests that the ensembling many well-performing models together can resolve this ambiguity, as the new ensemble model may perform better than any of its individual members. However, this approach may only push the problem from the member level to the ensemble level, as there may also be many different ensemble models that fit the data well.

The Rashomon effect has also been considered in several subject areas outside of VI, including those in non-statistical academic disciplines (Heider, 1988; Roth and Mehta, 2002). Tulabandhula and Rudin (2014) optimize a decision rule to perform well under the predicted range of outcomes from any well-performing model. Statnikov et al. (2013) propose an algorithm to discover multiple Markov boundaries, that is, minimal sets of covariates such that conditioning on any one set induces independence between the outcome and the remaining covariates. Nevo and Ritov (2017) report interpretations corresponding to a set of well-fitting, sparse linear models. Meinshausen and Bühlmann (2010) estimate structural aspects of an underlying model (such as the variables included in that model) based on how stable those aspects are across a set of well-fitting models. This set of well-fitting models is identified by repeating an estimation procedure in a series of perturbed samples, using varying levels of regularization (see also Azen et al., 2001). Letham et al. (2016) search for a pair of well-fitting dynamical systems models that give maximally different predictions.

6. Calculating Empirical Estimates of Model Class Reliance

In this section, we propose a binary search procedure to bound the values of $\widehat{MCR}_-(\epsilon)$ and $\widehat{MCR}_+(\epsilon)$ (see Eq 2.4), which respectively serve as estimates of $MCR_-(\epsilon)$ and $MCR_+(\epsilon)$ (see Section 4.1). Each step of this search consists of minimizing a linear combination of $\hat{e}_{\text{orig}}(f)$ and $\hat{e}_{\text{switch}}(f)$ across $f \in \mathcal{F}$. Our approach is related to the fractional programming approach of Dinkelbach (1967), but accounts for the fact that the problem is constrained by the value of the denominator, $\hat{e}_{\text{orig}}(f)$. We additionally show that, for many model classes,

Model class and loss function ($\mathcal{F}$ & L)	**Computing $\widehat{MCR}_-$**	**Computing $\widehat{MCR}_+$**
(L2 Regularized) Linear models, with the squared error loss	Highly tractable (QP1QC, see Sections 7.2 & 7.3)	Highly tractable (QP1QC, see Sections 7.2 & 7.3)
Linear models in a reproducing kernel Hilbert space, with the squared error loss	Moderately tractable (QP1QC, see Section 7.4.1)	Moderately tractable (QP1QC, see Section 7.4.1)
Cases where irrelevant covariates do not improve predictions	Moderately tractable (Convex optimization problems, see Proposition 11)	Potentially intractable
Cases where minimizing the empirical loss is a convex optimization problem	Potentially intractable (DC programs, see Section 6.3)	Potentially intractable (DC programs, see Section 6.3)

Table 1: Tractability of empirical MCR computation for different model classes – For each case, we describe the tractability of computing $\widehat{MCR}_-$ and $\widehat{MCR}_+$ using our proposed approaches. Computing empirical MCR can be reduced to a sequence of optimization problems, the form of which are noted in parentheses within the above table.

computing $\widehat{MCR}_-(\epsilon)$ only requires that we minimize convex combinations of $\hat{e}_{\text{orig}}(f)$ and $\hat{e}_{\text{switch}}(f)$, which is no more difficult than minimizing the average loss over an expanded and reweighted sample (See Eq 6.2 & Proposition 11).

Computing $\widehat{MCR}_+(\epsilon)$ however will require that we are able to minimize arbitrary linear combinations of $\hat{e}_{\text{orig}}(f)$ and $\hat{e}_{\text{switch}}(f)$. In Section 6.3, we outline how this can be done for convex model classes – classes for which the loss function is convex in the model parameter. Later, in Section 7, we give more specific computational procedures for when $\mathcal{F}$ is the class of linear models, regularized linear models, or linear models in a reproducing kernel Hilbert space (RKHS). We summarize the tractability of computing empirical MCR for different model classes in Table 1.

To simplify notation associated with the reference model f_{ref}, we present our computational results in terms of bounds on empirical MR subject to performance thresholds on the *absolute* scale. More specifically, we present bound functions b_- and b_+ satisfying $b_-(\epsilon_{\text{abs}}) \leq \widehat{MR}(f) \leq b_+(\epsilon_{\text{abs}})$ simultaneously for all $\{f, \epsilon_{\text{abs}} : \hat{e}_{\text{orig}}(f) \leq \epsilon_{\text{abs}}, f \in \mathcal{F}, \epsilon_{\text{abs}} > 0\}$ (Figures 2 & 8 show examples of these bounds). The binary search procedures we propose can be used to tighten these boundaries at a particular value ϵ_{abs} of interest.

We briefly note that as an alternative to the global optimization procedures we discuss below, heuristic optimization procedures such as simulated annealing can also prove useful in bounding empirical MCR. By definition, the empirical MR for any model in $\hat{\mathcal{R}}(\epsilon)$ forms a lower bound for $\widehat{MCR}_+(\epsilon)$, and an upper bound for $\widehat{MCR}_-(\epsilon)$. Heuristic maximization and minimization of empirical MR can be used to tighten these boundaries.

Throughout this section, we assume that $0 < \min_{f \in \mathcal{F}} \hat{e}_{\text{orig}}(f)$, to ensure that MR is finite.

6.1. Binary Search for Empirical MR Lower Bound

Before describing our binary search procedure, we introduce additional notation used in this section. Given a constant $\gamma \in \mathbb{R}$ and prediction model $f \in \mathcal{F}$, we define the linear combination $\hat{h}_{-,\gamma}$, and its minimizers (for example, $\hat{g}_{-,\gamma,\mathcal{F}}$), as

$$\hat{h}_{-,\gamma}(f) := \gamma \hat{e}_{\text{orig}}(f) + \hat{e}_{\text{switch}}(f), \quad \text{and} \quad \hat{g}_{-,\gamma,\mathcal{F}} \in \arg\min_{f \in \mathcal{F}} \hat{h}_{-,\gamma}(f).$$

We do not require that $\hat{h}_{-,\gamma}$ is uniquely minimized, and we frequently use the abbreviated notation $\hat{g}_{-,\gamma}$ when $\mathcal{F}$ is clear from context.

Our goal in this section is to derive a lower bound on $\widehat{MR}$ for subsets of $\mathcal{F}$ in the form of $\{f \in \mathcal{F} : \hat{e}_{\text{orig}}(f) \leq \epsilon_{\text{abs}}\}$. We achieve this by minimizing a series of linear objective functions in the form of $\hat{h}_{-,\gamma}$, using a similar method to that of Dinkelbach (1967). Often, minimizing the linear combination $\hat{h}_{-,\gamma}(f)$ is more tractable than minimizing the MR ratio directly.

Almost all of the results shown in this section, and those in Section 6.2, also hold if we replace $\hat{e}_{\text{switch}}$ with $\hat{e}_{\text{divide}}$ throughout (see Eq 3.5), including in the definition of $\widehat{MR}$ and $\hat{h}_{-,\gamma}(f)$. The exception is Proposition 11, below, which we may still expect to approximately hold if we replace $\hat{e}_{\text{switch}}$ with $\hat{e}_{\text{divide}}$.

Given an observed sample, we define the following condition for a pair of values $\{\gamma, \epsilon_{\text{abs}}\} \in \mathbb{R} \times \mathbb{R}_{>0}$, and argmin function $\hat{g}_{-,\gamma}$:

Condition 8 (Criteria to continue search for $\widehat{MR}$ lower bound) $\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma}) \geq 0$ *and* $\hat{e}_{orig}(\hat{g}_{-,\gamma}) \leq \epsilon_{abs}$.

We are now equipped to determine conditions under which we can tractably create a lower bound for empirical MR.

Lemma 9 *(Lower bound for $\widehat{MR}$) If $\gamma \in \mathbb{R}$ satisfies $\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma}) \geq 0$, then*

$$\frac{\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma})}{\epsilon_{abs}} - \gamma \leq \widehat{MR}(f) \tag{6.1}$$

for all $f \in \mathcal{F}$ satisfying $\hat{e}_{orig}(f) \leq \epsilon_{abs}$. It also follows that

$$-\gamma \leq \widehat{MR}(f) \qquad \text{for all } f \in \mathcal{F}.$$

Additionally, if $f = \hat{g}_{-,\gamma}$ and at least one of the inequalities in Condition 8 holds with equality, then Eq 6.1 holds with equality.

Lemma 9 reduces the challenge of lower-bounding $\widehat{MR}(f)$ to the task of minimizing the linear combination $\hat{h}_{-,\gamma}(f)$. The result of Lemma 9 is not only a single boundary for a particular value of ϵ_{abs}, but a boundary *function* that holds all values of $\epsilon_{\text{abs}} > 0$, with lower values of ϵ_{abs} leading to more restrictive lower bounds on $\widehat{MR}(f)$.

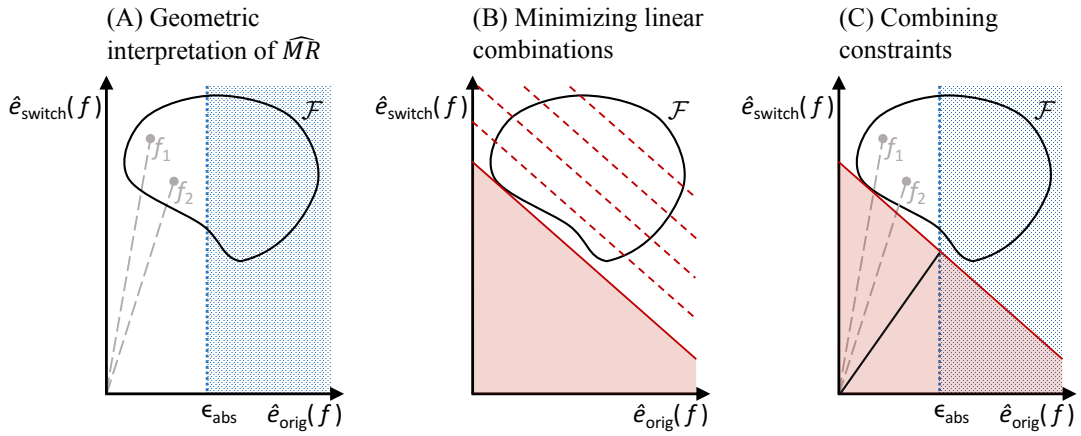

Figure 4: Above, we illustrate the geometric intuition for Lemma 9. In Panel (A), we show an example of a hypothetical model class $\mathcal{F}$, marked by the enclosed region. For each model $f \in \mathcal{F}$, the x-axis shows $\hat{e}_{\text{orig}}(f)$ and the y-axis shows $\hat{e}_{\text{switch}}(f)$. Here, we can see that the condition $\min_{f \in \mathcal{F}} \hat{e}_{\text{orig}}(f) > 0$ holds. The blue dotted region marks models with higher empirical loss. We mark two example models within $\mathcal{F}$ as f_1 and f_2. The slopes of the lines connecting the origin to f_1 and f_2 are equal to $\widehat{MR}(f_1)$ and $\widehat{MR}(f_2)$ respectively. Our goal is to lower-bound the slope corresponding to $\widehat{MR}$ for any model f satisfying $\hat{e}_{\text{orig}}(f) \leq \epsilon_{\text{abs}}$. In Panel (B), we consider the linear combination $\hat{h}_{-,\gamma}(f) = \gamma \hat{e}_{\text{orig}}(f) + \hat{e}_{\text{switch}}(f)$ for $\gamma = 1$. Above, contour lines of $\hat{h}_{-,\gamma}$ are shown in red. The solid red line indicates the smallest possible value of $\hat{h}_{-,\gamma}$ across $f \in \mathcal{F}$. Specifically, its y-intercept equals $\min_{f \in \mathcal{F}} \hat{h}_{-,\gamma}(f)$. If we can determine this minimum, we can determine a linear border constraint on $\mathcal{F}$, that is, we will know that no points corresponding to models $f \in \mathcal{F}$ may lie in the shaded region above. Additionally, if $\min_{f \in \mathcal{F}} \hat{h}_{-,\gamma}(f) \geq 0$ (see Lemma 9), then we know that the origin is either excluded by this linear constraint, or is on the boundary. In Panel (C), we combine the two constraints from Panels (A) & (B) to see that models $f \in \mathcal{F}$ satisfying $\hat{e}_{\text{orig}}(f) \leq \epsilon_{\text{abs}}$ must correspond to points in the white, unshaded region above. Thus, as long as the unshaded region does not contain the origin, any line connecting the origin to the a model f satisfying $\hat{e}_{\text{orig}}(f) \leq \epsilon_{\text{abs}}$ (for example, here, f_1,f_2) must have a slope at least as high as that of the solid black line above. It can be shown algebraically that the black line has slope equal to the left-hand side of Eq 6.1. Thus the left-hand side of Eq 6.1 is a lower bound for $\widehat{MR}(f)$ for all $\{f \in \mathcal{F} : \hat{e}_{\text{orig}}(f) \leq \epsilon_{\text{abs}}\}$.

In addition to the formal proof for Lemma 9, we provide a heuristic illustration of the result in Figure 4, to aid intuition.

It remains to determine which value of γ should be used in Eq 6.1. The following lemma implies that this value can be determined by a binary search, given a particular value of interest for ϵ_{abs}.

Lemma 10 *(Monotonicity for $\widehat{MR}$ lower bound binary search) The following monotonicity results hold:*

1. *$\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma})$ is monotonically increasing in γ.*
2. *$\hat{e}_{orig}(\hat{g}_{-,\gamma})$ is monotonically decreasing in γ.*
3. *Given ϵ_{abs}, the lower bound from Eq 6.1, $\left\{\frac{\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma})}{\epsilon_{abs}} - \gamma\right\}$, is monotonically decreasing in γ in the range where $\hat{e}_{orig}(\hat{g}_{-,\gamma}) \leq \epsilon_{abs}$, and increasing otherwise.*

Given a particular performance level of interest, ϵ_{abs}, Point 3 of Lemma 10 tells us that the value of γ resulting in the tightest lower bound from Eq 6.1 occurs when γ is as low as possible while still satisfying Condition 8. Points 1 and 2 show that if γ_0 satisfies Condition 8, and one of the equalities in Condition 8 holds with equality, then Condition 8 holds for all $\gamma \geq \gamma_0$. Together, these results imply that we can use a binary search to determine the value of γ to be used in Lemma 9, reducing this value until Condition 8 is no longer met. In addition to the formal proof for Lemma 10, we provide an illustration of the result in Figure 5 to aid intuition.

Next we present simple conditions under which the binary search for values of γ can be restricted to the nonnegative real line. This result substantially extends the computational tractability of our approach, as minimizing $\hat{h}_{-,\gamma}$ for $\gamma \geq 0$ is equivalent to minimizing a reweighted empirical loss over an expanded sample of size n^2:

$$\hat{h}_{-,\gamma}(f) = \gamma \hat{e}_{\text{orig}}(f) + \hat{e}_{\text{switch}}(f) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_\gamma(i,j) L\{f, (\mathbf{y}_{[i]}, \mathbf{X}_{1[j,\cdot]}, \mathbf{X}_{2[i,\cdot]})\}, \tag{6.2}$$

where $w_\gamma(i,j) = \frac{\gamma 1(i=j)}{n} + \frac{1(i\neq j)}{n(n-1)} \geq 0$.

Proposition 11 *(Nonnegative weights for $\widehat{MR}$ lower bound binary search) Assume that L and $\mathcal{F}$ satisfy the following conditions.*

1. *(Predictions are sufficient for computing the loss) The loss $L\{f, (Y, X_1, X_2)\}$ depends on the covariates (X_1, X_2) only via the prediction function f, that is, $L\{f, (y, x_1^{(a)}, x_2^{(a)})\} = L\{f, (y, x_1^{(b)}, x_2^{(b)})\}$ whenever $f(x_1^{(a)}, x_2^{(a)}) = f(x_1^{(b)}, x_2^{(b)})$.*
2. *(Irrelevant information does not improve predictions) For any distribution D satisfying $X_1 \perp_D (X_2, Y)$, there exists a function f_D satisfying*

$$\mathbb{E}_D L\{f_D, (Y, X_1, X_2)\} = \min_{f\in\mathcal{F}} \mathbb{E}_D L\{f, (Y, X_1, X_2)\},$$

and

$$f_D(x_1^{(a)}, x_2) = f_D(x_1^{(b)}, x_2) \text{ for any } x_1^{(a)}, x_1^{(b)} \in \mathcal{X}_1 \text{ and } x_2 \in \mathcal{X}_2. \tag{6.3}$$

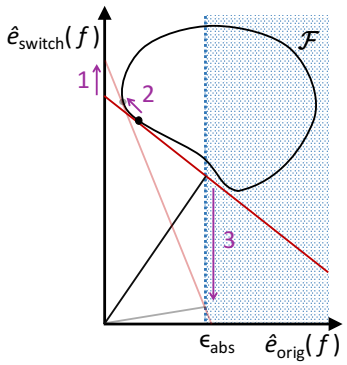

Figure 5: Monotonicity for binary search. Above we show a version of Figure 4-C for two alternative values of γ. This figure is meant to add intuition for the monotonicity results in Lemma 10, in addition to the formal proof. Increasing γ is equivalent to *decreasing* the slope of the red line in Figure 4-C. We define two values $\gamma_1 < \gamma_2$, where γ_1 corresponds to the solid red line, above, and γ_2 corresponds to the semi-transparent red line. The y-intercept values of these lines are equal to $\hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_1})$ and $\hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_2})$ respectively (see Figure 4-C caption). The solid and semi-transparent black dots mark $\hat{g}_{-,\gamma_1}$ and $\hat{g}_{-,\gamma_2}$ respectively. Plugging γ_1 and γ_2 into Eq 6.1 yields two lower bounds for $\widehat{MR}$, marked by the slopes of the solid and semi-transparent black lines respectively (see Figure 4-C caption). We see that (1) $\hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_1}) \leq \hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_2})$, that (2) $\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_1}) \geq \hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_2})$, and that (3) the left-hand side of Eq 6.1 is decreasing in γ when $\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma}) \leq \epsilon_{\text{abs}}$. These three conclusions are marked by arrows in the above figure, with numbering matching the enumerated list in Lemma 10.

Let $\gamma = 0$. Under the above assumptions, it follows that either (i) there exists a function $\hat{g}_{-,0}$ minimizing $\hat{h}_{-,0}$ that does not satisfy Condition 8, or (ii) $\hat{e}_{orig}(\hat{g}_{-,0}) \leq \epsilon_{abs}$ and $\widehat{MR}(g_{-,0}) \leq 1$ for any function $\hat{g}_{-,0}$ minimizing $\hat{h}_{-,0}$.

The implication of Proposition 11 is that, when the conditions of Proposition 11 are met, the search region for γ can be limited to the nonnegative real line, and minimizing $\hat{h}_{-,\gamma}$ will be no harder than minimizing a reweighted empirical loss over an expanded sample (Eq 6.2). To see this, recall that for a fixed value of ϵ_{abs} we can tighten the boundary in Lemma 9 by conducting a binary search for the smallest value of γ that satisfies Condition 8. If setting γ equal to 0 does not satisfy Condition 8, and the search for γ can be restricted to the nonnegative real line, where minimizing $\hat{h}_{-,0}$ is more tractable (see Eq 6.2). Alternatively, if $\hat{e}_{\text{orig}}(g_{-,0}) \leq \epsilon_{\text{abs}}$ and $\widehat{MR}(g_{-,0}) \leq 1$, then we have identified a well-performing model $g_{-,0}$ with empirical MR no greater than 1. For $\epsilon_{\text{abs}} = \hat{e}_{\text{orig}}(f_{\text{ref}}) + \epsilon$, this implies that $\widehat{MCR}_{-}(\epsilon) \leq 1$, which is a sufficiently precise conclusion for most interpretational purposes (see Appendix A.2).

Because of the fixed pairing structure used in $\hat{e}_{\text{divide}}$, Proposition 11 will not necessarily hold if we replace $\hat{e}_{\text{switch}}$ with $\hat{e}_{\text{divide}}$ throughout (see Appendix C.3). However, since $\hat{e}_{\text{divide}}$ approximates $\hat{e}_{\text{switch}}$, we can expect Proposition 11 to hold approximately. The bound from Eq 6.1 still remains valid if we replace $\hat{e}_{\text{switch}}$ with $\hat{e}_{\text{divide}}$ and limit γ to the nonnegative reals, although in some cases it may not be as tight.

6.2. Binary Search for Empirical MR Upper Bound

We now briefly present a binary search procedure to upper bound $\widehat{MR}$, which mirrors the procedure from Section 6.1. Given a constant $\gamma \in \mathbb{R}$ and prediction model $f \in \mathcal{F}$, we define the linear combination $\hat{h}_{+,\gamma}$, and its minimizers (for example, $\hat{g}_{+,\gamma,\mathcal{F}}$), as

$$\hat{h}_{+,\gamma}(f) := \hat{e}_{\text{orig}}(f) + \gamma \hat{e}_{\text{switch}}(f), \quad \text{and} \quad \hat{g}_{+,\gamma,\mathcal{F}} \in \arg\min_{f \in \mathcal{F}} \hat{h}_{+,\gamma}(f).$$

As in Section 6.1, $\hat{h}_{+,\gamma}$ need not be uniquely minimized, and we generally abbreviate $\hat{g}_{+,\gamma,\mathcal{F}}$ as $\hat{g}_{+,\gamma}$ when $\mathcal{F}$ is clear from context.

Given an observed sample, we define the following condition for a pair of values $\{\gamma, \epsilon_{\text{abs}}\} \in \mathbb{R}_{\leq 0} \times \mathbb{R}_{>0}$, and argmin function $\hat{g}_{+,\gamma}$:

Condition 12 (Criteria to continue search for $\widehat{MR}$ upper bound) $\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma}) \geq 0$ *and* $\hat{e}_{orig}(\hat{g}_{+,\gamma}) \leq \epsilon_{abs}$.

We can now develop a procedure to upper bound $\widehat{MR}$, as shown in the next lemma.

Lemma 13 *(Upper bound for $\widehat{MR}$) If $\gamma \in \mathbb{R}$ satisfies $\gamma \leq 0$ and $\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma}) \geq 0$, then*

$$\widehat{MR}(f) \leq \left\{ \frac{\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma})}{\epsilon_{abs}} - 1 \right\} \gamma^{-1} \tag{6.4}$$

for all $f \in \mathcal{F}$ satisfying $\hat{e}_{orig}(f) \leq \epsilon_{abs}$. It also follows that

$$\widehat{MR}(f) \leq |\gamma^{-1}| \qquad \text{for all } f \in \mathcal{F}. \tag{6.5}$$

Additionally, if $f = \hat{g}_{+,\gamma}$ and at least one of the inequalities in Condition 12 holds with equality, then Eq 6.4 holds with equality.

As in Section 6.1, it remains to determine the value of γ to use in Lemma 13, given a value of interest for $\epsilon_{\text{abs}} \geq \min_{f \in \mathcal{F}} \hat{e}_{\text{orig}}(f)$. The next lemma tells us that the boundary from Lemma 13 is tightest when γ is as low as possible while still satisfying Condition 12.

Lemma 14 *(Monotonicity for $\widehat{MR}$ upper bound binary search) The following monotonicity results hold:*

1. *$\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma})$ is monotonically increasing in γ.*
2. *$\hat{e}_{orig}(\hat{g}_{+,\gamma})$ is monotonically decreasing in γ for $\gamma \leq 0$, and Condition 12 holds for $\gamma = 0$ and $\epsilon_{abs} \geq \min_{f \in \mathcal{F}} \hat{e}_{orig}(f)$.*

3. *Given ϵ_{abs}, the upper boundary $\left\{\frac{\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma})}{\epsilon_{abs}} - 1\right\}\gamma^{-1}$ is monotonically increasing in γ in the range where $\hat{e}_{orig}(\hat{g}_{+,\gamma}) \leq \epsilon_{abs}$ and $\gamma < 0$, and decreasing in the range where $\hat{e}_{orig}(\hat{g}_{+,\gamma}) > \epsilon_{abs}$ and $\gamma < 0$.*

Together, the results from Lemma 14 imply that we can use a binary search across $\gamma \in \mathbb{R}$ to tighten the boundary on $\widehat{MR}$ from Lemma 13.

6.3. Convex Models

In this section we show that empirical MCR can be conservatively computed when the loss function is convex in the model parameters – that is, when the models $f_\theta \in \mathcal{F}$ are indexed by a d-dimensional parameter $\theta \in \Theta \subseteq \mathbb{R}^d$, and when the loss function $L(f_\theta, (y, x_1, x_2))$ is convex in θ for all $(x_1, x_2, y) \in \mathcal{X}_1 \times (\mathcal{X}_2, \mathcal{Y})$.

Fortunately, neither Lemma 9 nor Lemma 13 require an exact minimum for $\hat{h}_{-,\gamma}$ or $\hat{h}_{+,\gamma}$. For Lemma 9, any lower bound on $\hat{h}_{-,\gamma}$ is sufficient to determine a lower bound on $MR(f)$. Likewise, for Lemma 13, any lower bound on $\hat{h}_{+,\gamma}$ is sufficient to determine an upper bound on $MR(f)$.

To find these lower bounds, we note that for "convex" model classes (defined above) the optimization problems in Sections 6.1 & 6.2 can be written either as convex optimization problems, or as difference convex function (DC) programs. A DC program is one that can be written as

$$\min_{\{\theta : c_{\text{DC}}(\theta) \leq k, \theta \in \Theta\}} g_{\text{DC}}(\theta) - h_{\text{DC}}(\theta),$$

where c_{DC} is a constraint function, $k \in \mathbb{R}^1$, and g_{DC}, h_{DC}, and c_{DC} are convex. Although precise solutions to DC problems are not always tractable, lower bounds can be attained by branch-and-bound (B&B) methods (Horst and Thoai, 1999). A simple B&B approach is to partition Θ into a set of simplexes. Within the j^{th} simplex, a lower bound on $g_{\text{DC}}(\theta) - h_{\text{DC}}(\theta)$ can be determined by replacing h_{DC} with the hyperplane function h_j satisfying $h_j(v) = h_{\text{DC}}(v)$ at each vertex v of the j^{th} simplex. Within this partition, $g_{\text{DC}}(\theta) - h_{\text{DC}}(\theta)$ is lower bounded by $l_j := \min_\theta g_{\text{DC}}(\theta) - h_j(\theta)$, which can be computed as the solution to a convex optimization problem. Any partition for which l_j is found to be too high is disregarded. Once a bound l_j is computed for each partition, the partition with the lowest value l_j is selected to be subdivided further, and additional lower bounds are recomputed for each new, resulting partition. This procedure continues until a sufficiently tight lower bound is attained (for more detailed procedures, see Horst and Thoai, 1999).

This approach allows us to conservatively approximate bounds on $\widehat{MR}(f)$ in the form of Eq 6.1 & 6.4 by replacing $\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma})$ and $\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma})$ with lower bounds from the B&B procedure. Although it will always yield valid bounds, the procedure may converge slowly when the dimension of Θ is large, giving highly conservative results. For some special cases of model classes however, even high dimensional DC problems simplify greatly. We discuss these cases in the next section.

7. MR & MCR for Linear Models, Additive Models, and Regression Models in a Reproducing Kernel Hilbert Space

For linear or additive models, many simplifications can be made to our approaches for MR and MCR. To simplify the interpretation of MR, we show below that population-level MR for a linear model can be expressed in terms of the model's coefficients (Section 7.1). To simplify computation, we show that the cost of computing empirical MR for a linear model grows only linearly in n (Section 7.1), even though the number of terms in the definition of empirical MR grows quadratically (see Eqs 3.3 & 3.6).

Moving on from MR, we show how empirical MCR can be computed for the class of linear models (Section 7.2), for regularized linear models (Section 7.3), and for regression models in a reproducing kernel Hilbert space (RKHS, Section 7.4). To do this, we build on the approach in Section 6 by giving approaches for minimizing arbitrary combinations of $\hat{e}_{\text{switch}}(f)$ and $\hat{e}_{\text{orig}}(f)$ across $f \in \mathcal{F}$. Even when the associated objective functions are non-convex, we can tractably obtain global minima for these model classes. We also discuss procedures to determine an upper bound B_{ind} on the loss for any observation when using these model classes (see Assumption 1).

Throughout this section, we assume that $\mathcal{X} \subset \mathbb{R}^p$ for $p \in \mathbb{Z}^+$, that $\mathcal{Y} \subset \mathbb{R}^1$, and that L is the squared error loss function $L(f, (y, x_1, x_2) = (y - f(x_1, x_2))^2$. As in Section 6, we also assume that $0 < \min_{f \in \mathcal{F}} \hat{e}_{\text{orig}}(f)$, to ensure that empirical MR is finite.

7.1. Interpreting and Computing MR for Linear or Additive Models

We begin by considering MR for linear models evaluated with the squared error loss. For this setting, we can show both an interpretable definition of MR, as well as a computationally efficient formula for $\hat{e}_{\text{switch}}(f)$.

Proposition 15 *(Interpreting MR, and computing empirical MR for linear models) For any prediction model f, let $e_{orig}(f)$, $e_{switch}(f)$, $\hat{e}_{\text{orig}}(f)$, and $\hat{e}_{switch}(f)$ be defined based on the squared error loss $L(f, (y, x_1, x_2)) := (y - f(x_1, x_2))^2$ for $y \in \mathbb{R}$, $x_1 \in \mathbb{R}^{p_1}$, and $x_2 \in \mathbb{R}^{p_2}$, where p_1 and p_2 are positive integers. Let $\beta = (\beta_1, \beta_2)$ and f_β satisfy $\beta_1 \in \mathbb{R}^{p_1}$, $\beta_2 \in \mathbb{R}^{p_2}$, and $f_\beta(x) = x'\beta = x_1'\beta_1 + x_2'\beta_2$. Then*

$$MR(f_\beta) = 1 + \frac{2}{e_{orig}(f_\beta)} \left\{ Cov(Y, X_1)\beta_1 - \beta_2' Cov(X_2, X_1)\beta_1 \right\}, \tag{7.1}$$

and, for finite samples,

$$\hat{e}_{switch}(f_\beta) = \frac{1}{n} \left\{ \mathbf{y}'\mathbf{y} - 2 \begin{bmatrix} \mathbf{X}_1'\mathbf{W}\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{bmatrix}' \beta + \beta' \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{W}\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{W}\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix} \beta \right\}, \tag{7.2}$$

where $\mathbf{W} := \frac{1}{n-1}(\mathbf{1}_n\mathbf{1}_n' - \mathbf{I}_n)$, $\mathbf{1}_n$ is the n-length vector of ones, and $\mathbf{I}_n$ is the $n \times n$ identity matrix.

Eq 7.1 shows that model reliance for linear models can be interpreted in terms of the population covariances, the model coefficients, and the model's accuracy. Gregorutti et al. (2017) show an equivalent formulation of Eq 7.1 under the stronger assumptions that f_β is

equal to the conditional expectation function of Y (that is, $f_\beta(x) = \mathbb{E}(Y|X = x)$), and the covariates X_1 and X_2 are centered.

Eq 7.2 shows that, although the number of terms in the definition of $\hat{e}_{\text{switch}}$ grows quadratically in n (see Eq 3.3), the computational complexity of $\hat{e}_{\text{switch}}(f_\beta)$ for a linear model f_β grows *only linearly* in n. Specifically, the terms $\mathbf{X}_1'\mathbf{W}\mathbf{y}$ and $\mathbf{X}_1'\mathbf{W}\mathbf{X}_2$ in Eq 7.2 can be computed as $\frac{1}{n-1}\{(\mathbf{X}_1'\mathbf{1}_n)(\mathbf{1}_n'\mathbf{y}) - (\mathbf{X}_1'\mathbf{y})\}$ and $\frac{1}{n-1}\{(\mathbf{X}_1'\mathbf{1}_n)(\mathbf{1}_n'\mathbf{X}_2) - (\mathbf{X}_1'\mathbf{X}_2)\}$ respectively, where the computational complexity of each term in parentheses grows linearly in n.

As in Gregorutti et al. (2017), both results in Proposition 15 readily generalize to additive models of the form $f_{g_1,g_2}(X_1, X_2) := g_1(X_1) + g_2(X_2)$, since permuting X_1 is equivalent to permuting $g_1(X_1)$.

7.2. Computing Empirical MCR for Linear Models

Building on the computational result from the previous section, we now consider empirical MCR computation for linear model classes of the form

$$\mathcal{F}_{\text{lm}} := \left\{ f_\beta : f_\beta(x) = x'\beta, \quad \beta \in \mathbb{R}^p \right\}.$$

In order to implement the computational procedure from Sections 6.1 and 6.2, we must be able to minimize arbitrary linear combinations of $\hat{e}_{\text{orig}}(f_\beta)$ and $\hat{e}_{\text{switch}}(f_\beta)$. Fortunately, for linear models, this minimization reduces to a quadratic program, as we show in the next remark.

Remark 16 *(Tractability of empirical MCR for linear model classes) For any $f_\beta \in \mathcal{F}_{lm}$ and any fixed coefficients $\xi_{orig}, \xi_{switch} \in \mathbb{R}$, the linear combination*

$$\xi_{orig}\hat{e}_{orig}(f_\beta) + \xi_{switch}\hat{e}_{switch}(f_\beta) \tag{7.3}$$

is proportional in β to the quadratic function $-2\mathbf{q}'\beta + \beta'\mathbf{Q}\beta$, where

$$\mathbf{Q} := \xi_{orig}\mathbf{X}'\mathbf{X} + \xi_{switch}\begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{W}\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{W}\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix}, \qquad \mathbf{q} := \left(\xi_{orig}\mathbf{y}'\mathbf{X} + \xi_{switch}\begin{bmatrix} \mathbf{X}_1'\mathbf{W}\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{bmatrix}' \right)',$$

and $\mathbf{W} := \frac{1}{n-1}(\mathbf{1}_n\mathbf{1}_n' - \mathbf{I}_n)$. Thus, minimizing $\xi_{orig}\hat{e}_{orig}(f_\beta) + \xi_{switch}\hat{e}_{switch}(f_\beta)$ is equivalent to an unconstrained (possibly non-convex) quadratic program.

Because our empirical MCR computation procedure from Sections 6.1 and 6.2 consists of minimizing a sequence of objective functions in the form of Eq 7.3, Remark 16 shows us that this procedure is tractable for the class of unconstrained linear models.

7.3. Regularized Linear Models

Next, we continue to build on the results from Section 7.2 to calculate boundaries on $\widehat{MR}$ for *regularized* linear models. We consider model classes formed by quadratically constrained subsets of $\mathcal{F}_{\text{lm}}$, defined as

$$\mathcal{F}_{\text{lm},r_{\text{lm}}} := \left\{ f_\beta : f_\beta(x) = x'\beta, \quad \beta \in \mathbb{R}^p, \quad \beta'\mathbf{M}_{\text{lm}}\beta \le r_{\text{lm}} \right\}, \tag{7.4}$$

where $\mathbf{M}_{\text{lm}}$ and r_{lm} are pre-specified. Again, this class describes linear models with a quadratic constraint on the coefficient vector.

7.3.1. Calculating MCR

As in Section 7.2, calculating bounds on $\widehat{MR}$ via Lemmas 9 & 13 requires that are able to minimizing linear combinations $\xi_{\text{orig}}\hat{e}_{\text{orig}}(f_\beta) + \xi_{\text{switch}}\hat{e}_{\text{switch}}(f_\beta)$ across $f_\beta \in \mathcal{F}_{\text{lm},r_{\text{lm}}}$ for arbitrary $\xi_{\text{orig}}, \xi_{\text{switch}} \in \mathbb{R}$. Applying Remark 16, we can again equivalently minimize $-2\mathbf{q}'\beta + \beta'\mathbf{Q}\beta$ subject to the constraint in Eq 7.4:

$$\begin{aligned} \text{minimize} \quad & -2\mathbf{q}'\beta + \beta'\mathbf{Q}\beta \\ \text{subject to} \quad & \beta'\mathbf{M}_{\text{lm}}\beta \leq r_{\text{lm}}. \end{aligned} \tag{7.5}$$

The resulting optimization problem is a (possibly non-convex) quadratic program with one quadratic constraint (QP1QC). This problem is well-studied, and is related to the trust region problem (Boyd and Vandenberghe, 2004; Pólik and Terlaky, 2007; Park and Boyd, 2017). Thus, the bounds on MCR presented in Sections 6.1 and 6.2 again become computationally tractable for the class of quadratically constrained linear models.

7.3.2. Upper Bounding the Loss

One benefit of constraining the coefficient vector ($\beta'\mathbf{M}_{\text{lm}}\beta \leq r_{\text{lm}}$) is that it facilitates determining an upper bound B_{ind} on the loss function $L(f_\beta, (y, x)) = (y - x'\beta)^2$, which automatically satisfies Assumption 1 for all $f \in \mathcal{F}_{\text{lm},r_{\text{lm}}}$. The following lemma gives sufficient conditions to determine B_{ind}.

Lemma 17 *(Loss upper bound for linear models) If $\mathbf{M}_{lm}$ is positive definite, Y is bounded within a known range, and there exists a known constant $r_{\mathcal{X}}$ such that $x'\mathbf{M}_{lm}^{-1}x \leq r_{\mathcal{X}}$ for all $x \in (\mathcal{X}_1 \times \mathcal{X}_2)$, then Assumption 1 holds for the model class $\mathcal{F}_{lm,r_{lm}}$, the squared error loss function, and the constant*

$$B_{ind} = \max\left[\left\{\min_{y\in\mathcal{Y}}(y) - \sqrt{r_{\mathcal{X}} r_{lm}}\right\}^2, \left\{\max_{y\in\mathcal{Y}}(y) + \sqrt{r_{\mathcal{X}} r_{lm}}\right\}^2\right].$$

In practice, the constant $r_{\mathcal{X}}$ can be approximated by the empirical distribution of X and Y. The motivation behind the restriction $x'\mathbf{M}_{\text{lm}}^{-1}x \leq r_{\mathcal{X}}$ in Lemma 17 is to create complementary constraints on X and β. For example, if $\mathbf{M}_{\text{lm}}$ is diagonal, then the smallest elements of $\mathbf{M}_{\text{lm}}$ correspond to directions along which β is least restricted by $\beta'\mathbf{M}_{\text{lm}}\beta \leq r_{\text{lm}}$ (Eq 7.5), as well as the directions along which x is most restricted by $x'\mathbf{M}_{\text{lm}}^{-1}x \leq r_{\mathcal{X}}$ (Lemma 17).

7.4. Regression Models in a Reproducing Kernel Hilbert Space (RKHS)

We now expand our scope of model classes by considering regression models in a reproducing kernel Hilbert space (RKHS), which allow for nonlinear and nonadditive features of the covariates. We show that, as in Section 7.3, minimizing a linear combination of $\hat{e}_{\text{orig}}(f)$ and $\hat{e}_{\text{switch}}(f)$ across models f in this class can be expressed as a QP1QC, which allows us to implement the binary search procedure of Sections 6.1 & 6.2.

First we introduce notation required to describe regression in a RKHS. Let $\mathbf{D}$ be a $(R\times p)$ matrix representing a pre-specified dictionary of R reference points, such that each row of $\mathbf{D}$ is contained in $\mathcal{X} = \mathbb{R}^p$. Let k be a pre-specified positive definite kernel function, and let μ

be a prespecified estimate of $\mathbb{E}Y$. Let $\mathbf{K_D}$ be the $R \times R$ matrix with $\mathbf{K_D}_{[i,j]} = k(\mathbf{D}_{[i,\cdot]}, \mathbf{D}_{[j,\cdot]})$. We consider prediction models of the following form, where the distance to each reference point is used as a regression feature:

$$\mathcal{F}_{\mathbf{D},r_k} = \left\{ f_\alpha : f_\alpha(x) = \mu + \sum_{i=1}^{R} k\left(x, \mathbf{D}_{[i,\cdot]}\right) \alpha_{[i]}, \quad \|f_\alpha\|_k \leq r_k, \quad \alpha \in \mathbb{R}^R \right\}. \tag{7.6}$$

Above, the norm $\|f_\alpha\|_k$ is defined as

$$\|f_\alpha\|_k := \sum_{i=1}^{R} \sum_{j=1}^{R} \alpha_{[i]} \alpha_{[j]} k(\mathbf{D}_{[i,\cdot]}, \mathbf{D}_{[j,\cdot]}) = \alpha' \mathbf{K_D} \alpha. \tag{7.7}$$

In the next two sections, we show that bounds on empirical MCR can again be tractably computed for this class, and that the loss for models in this class can be feasibly upper bounded.

7.4.1. Calculating MCR

Again, calculating bounds on $\widehat{MR}$ from Lemmas 9 & 13 requires us to be able to minimize arbitrary linear combinations of $\hat{e}_{\text{orig}}(f_\alpha)$ and $\hat{e}_{\text{switch}}(f_\alpha)$.

Given a size-n sample of test observations $\mathbf{Z} = \begin{bmatrix} \mathbf{y} & \mathbf{X} \end{bmatrix}$, let $\mathbf{K}_{\text{orig}}$ be the $n \times R$ matrix with elements $\mathbf{K}_{\text{orig}[i,j]} = k\left(\mathbf{X}_{[i,\cdot]}, \mathbf{D}_{[j,\cdot]}\right)$. Let $\mathbf{Z}_{\text{switch}} = \begin{bmatrix} \mathbf{y}_{\text{switch}} & \mathbf{X}_{\text{switch}} \end{bmatrix}$ be the $(n(n-1)) \times (1+p)$ matrix with rows that contain the set $\{(\mathbf{y}_{[i]}, \mathbf{X}_{1[j,\cdot]}, \mathbf{X}_{2[i,\cdot]}) : i, j \in \{1, \ldots, n\} \text{ and } i \neq j\}$. Finally, let $\mathbf{K}_{\text{switch}}$ be the $n(n-1) \times R$ matrix with $\mathbf{K}_{\text{switch}[i,j]} = k\left(\mathbf{X}_{\text{switch}[i,\cdot]}, \mathbf{D}_{[j,\cdot]}\right)$.

For any two constants $\xi_{\text{orig}}, \xi_{\text{switch}} \in \mathbb{R}$, we can show that minimizing the linear combination $\xi_{\text{orig}} \hat{e}_{\text{orig}}(f_\alpha) + \xi_{\text{switch}} \hat{e}_{\text{switch}}(f_\alpha)$ over $\mathcal{F}_{\mathbf{D},r_k}$ is equivalent to the minimization problem

$$\text{minimize} \quad \frac{\xi_{\text{orig}}}{n} \|\mathbf{y} - \mu - \mathbf{K}_{\text{orig}} \alpha\|_2^2 + \frac{\xi_{\text{switch}}}{n(n-1)} \|\mathbf{y}_{\text{switch}} - \mu - \mathbf{K}_{\text{switch}} \alpha\|_2^2 \tag{7.8}$$

$$\text{subject to} \quad \alpha' \mathbf{K_D} \alpha < r_k. \tag{7.9}$$

Like Problem 7.5, Problem 7.8-7.9 is a QP1QC. To show Eqs 7.8-7.9, we first write $\hat{e}_{\text{orig}}(f_\alpha)$ as

$$\hat{e}_{\text{orig}}(f_\alpha) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \mathbf{y}_{[i]} - f_\alpha(\mathbf{X}_{[i,\cdot]}) \right\}^2 \tag{7.10}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left\{ \mathbf{y}_{[i]} - \mu - \sum_{j=1}^{R} k(\mathbf{X}_{[i,\cdot]}, \mathbf{D}_{[j,\cdot]}) \alpha_{[j]} \right\}^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left\{ \mathbf{y}_{[i]} - \mu - \mathbf{K}'_{\text{orig}[i,\cdot]} \alpha \right\}^2$$

$$= \frac{1}{n} \|\mathbf{y} - \mu - \mathbf{K}_{\text{orig}} \alpha\|_2^2. \tag{7.11}$$

Following similar steps, we can obtain

$$\hat{e}_{\text{switch}}(f_\alpha) = \frac{1}{n(n-1)} \|\mathbf{y}_{\text{switch}} - \mu - \mathbf{K}_{\text{switch}}\alpha\|_2^2.$$

Thus, for any two constants $\xi_{\text{orig}}, \xi_{\text{switch}} \in \mathbb{R}$, we can see that $\xi_{\text{orig}}\hat{e}_{\text{orig}}(f_\alpha) + \xi_{\text{switch}}\hat{e}_{\text{switch}}(f_\alpha)$ is quadratic in α. This means that we can tractably compute bounds on empirical MCR for this class as well.

7.4.2. Upper Bounding the Loss

Using similar steps as in Section 7.3.2, the following lemma gives sufficient conditions to determine B_{ind} for the case of regression in a RKHS.

Lemma 18 *(Loss upper bound for regression in a RKHS) Assume that Y is bounded within a known range, and there exists a known constant $r_{\mathbf{D}}$ such that $v(x)'\mathbf{K}_{\mathbf{D}}^{-1}v(x) \leq r_{\mathbf{D}}$ for all $x \in (\mathcal{X}_1 \times \mathcal{X}_2)$, where $v : \mathbb{R}^p \to \mathbb{R}^R$ is the function satisfying $v(x)_{[i]} = k(x, \mathbf{D}_{[i,\cdot]})$. Under these conditions, Assumption 1 holds for the model class $\mathcal{F}_{\mathbf{D},r_k}$, the squared error loss function, and the constant*

$$B_{ind} = \max\left[\left\{\min_{y\in\mathcal{Y}}(y) - (\mu + \sqrt{r_{\mathbf{D}}r_k})\right\}^2, \left\{\max_{y\in\mathcal{Y}}(y) + (\mu + \sqrt{r_{\mathbf{D}}r_k})\right\}^2\right].$$

Thus, for regression models in a RKHS, we can satisfy Assumption 1 for all models in the class.

8. Connections Between MR and Causality

Our MR approach can be fundamentally described as studying how a model's behavior changes under an intervention on the underlying data. We aim to study the causal effect of this intervention on the *model's* performance. This goal mirror's the conventional causal inference goal of studying how an intervention on variables will change outcomes generated by a process in *nature*.

This section explores this connection to causal inference further. Section 8.1 shows that when the prediction model in question is the conditional expectation function from nature itself, MR reduces to commonly studied quantities in the causal literature. Section 8.2 proposes an alternative to MR that focuses on interventions, or data perturbations, that are likely to occur in the underlying data generating process.

8.1. Model Reliance and Causal Effects

In this section, we show a connection between population-level model reliance and the conditional average causal effect. For consistency with the causal inference literature, we temporarily rename the random variables (Y, X_1, X_2) as (Y, T, C), with realizations (y, t, c). Here, $T := X_1$ represents a binary treatment indicator, $C := X_2$ represents a set of baseline covariates ("C" is for "covariates"), and Y represents an outcome of interest. Under this notation, $e_{\text{orig}}(f)$ represents the expected loss of a prediction function f, and $e_{\text{switch}}(f)$ denotes the expected loss in a pair of observations in which the treatment has been switched.

Let $f_0(t,c) := \mathbb{E}(Y|C=c, T=t)$ be the (unknown) conditional expectation function for Y, where we place no restrictions on the functional form of f_0.

Let Y_1 and Y_0 be potential outcomes under treatment and control respectively, such that $Y = Y_0(1-T) + Y_1 T$. The treatment effect for an individual is defined as $Y_1 - Y_0$, and the average treatment effect is defined as $\mathbb{E}(Y_1 - Y_0)$. Let $\text{CATE}(c) := \mathbb{E}(Y_1 - Y_0|C=c)$ be the (unknown) conditional average treatment effect of T for all patients with $C=c$. Causal inference methods typically assume $(Y_1, Y_0) \perp T|C$ (conditional ignorability), and $0 < \mathbb{P}(T=1|C=c) < 1$ for all values of c (positivity), in order for f_0 and CATE to be well defined and identifiable.

The next proposition quantifies the relation between the conditional average treatment effect function (CATE) and the model reliance of f_0 on X_1.

Proposition 19 *(Causal interpretations of MR) For any prediction model f, let $e_{orig}(f)$ and $e_{switch}(f)$ be defined based on the squared error loss $L(f,(y,t,c)) := (y - f(t,c))^2$.*

If $(Y_1, Y_0) \perp T|C$ (conditional ignorability) and $0 < \mathbb{P}(T=1|C=c) < 1$ for all values of c (positivity), then $MR(f_0)$ is equal to

$$1 + \frac{Var(T)}{\mathbb{E}_{T,C} Var(Y|T,C)} \sum_{t\in\{0,1\}} \left\{ \mathbb{E}(Y_1 - Y_0|T=t)^2 + Var\left(CATE(C)|T=t\right) \right\}, \tag{8.1}$$

where $Var(T)$ is the marginal variance of the treatment assignment.

We see above that model reliance decomposes into several terms that are each individually important in causal inference: the treatment prevalence (via $\text{Var}(T)$); the variability in Y that is not explained by C or T; the magnitude of the average treatment effect, conditional on T; and the variance of the conditional average treatment effect across subgroups. For example, if all patients are treated, then scrambling the treatment in a random pair of observations has no effect on the loss. In this case we see that $\text{Var}(T) = 0$ and $MR(f_0) = 1$, indicating no reliance. When $\text{Var}(T) > 0$, a higher average treatment effect magnitude ($\mathbb{E}(Y_1 - Y_0|T=t)^2$) corresponds to f_0 requiring T more heavily to predict Y, all else equal. Similarly, if there is a high degree of treatment effect heterogeneity across subgroups (that is, when $\text{Var}\left(\text{CATE}(C)|T=t\right)$ is large), the model f_0 will again use T more heavily when predicting Y. For example, a treatment may be important for predicting Y even if the average treatment effect is zero, so long as the treatment helps some subgroups more than others.

8.2. Conditional Importance: Adjusting for Dependence Between X_1 and X_2

One common scenario where multiple models achieve low loss is when the sets of predictors X_1 and X_2 are highly correlated, or contain redundant information. Models may predict well either through reliance on X_1, or through reliance on X_2, and so MCR will correctly identify a wide range of potential reliances on X_1. However, we may specifically be interested how much models rely on the information in X_1 that cannot alternatively be gleaned from X_2.

For example, age and accumulated wealth may be correlated, and both may be predictive of future promotion. We may wish to know the how much a model for predicting promotion relies on information that is uniquely available from wealth measurements.

To formalize this notion, we define an alternative to e_{switch} where noise is added to X_1 in a way that accounts for the dependence between X_1 and X_2. Given a fixed prediction model f, we ask: how well would the model f perform if the values of X_1 were scrambled *across observations with the same value for* X_2. Specifically, let $Z^{(a)} = (Y^{(a)}, X_1^{(a)}, X_2^{(a)})$ and $Z^{(b)} = (Y^{(b)}, X_1^{(b)}, X_2^{(b)})$ denote a pair of independent random vectors following the same distribution as $Z = (Y, X_1, X_2)$, as in Section 3, and let

$$e_{\text{cond}}(f) := \mathbb{E}_{X_2}\mathbb{E}_{(Y^{(b)},X_1^{(a)},X_2^{(b)})}\left[L\{f,(Y^{(b)},X_1^{(a)},X_2^{(b)})\}|X_2^{(a)} = X_2^{(b)} = X_2\right]. \tag{8.2}$$

In words, $e_{\text{cond}}(f)$ is the expected loss of a given model f across pairs of observations $(Z^{(a)}, Z^{(b)})$ in which the values of $X_1^{(a)}$ and $X_1^{(b)}$ have been switched, given that these pairs match on X_2. This quantity can also be interpreted as the expected loss of f if noise were added to X_1 in such a way that X_1 was no longer informative of Y, given X_2, but that the joint distribution of the covariates (X_1, X_2) was maintained.

We then define conditional model reliance, or "core" model reliance (CMR) for a fixed function f as

$$CMR(f) = \frac{e_{\text{cond}}(f)}{e_{\text{orig}}(f)}.$$

That is, CMR is the factor by which the model's performance degrades when the information unique to X_1 is removed. If $X_1 \perp X_2$, then X_1 contains no redundant information, and CMR and MR are equivalent. Otherwise, all else equal, CMR will decrease as X_2 becomes more predictive of X_1. Analogous to MCR, we define conditional MCR (CMCR) in the same way as in Eq 2.2, but with MR replaced with CMR. In comparison with MCR, CMCR will generally result in a range that is closer to 1 (null reliance).

An advantage of CMR is that it restricts the "noise-corrupted" inputs to be within the domain $\mathcal{X}$, rather than the expanded domain $\mathcal{X}_1 \times \mathcal{X}_2$ considered by MR. This means that CMR will not be influenced by impossible combinations of x_1 and x_2, while MR may be influenced by them. Hooker (2007) discuss a similar issue, arguing that evaluations of a prediction model's behavior in different circumstances should be weighted by, for example, how likely those circumstances are to occur.

A challenge facing the CMR approach is that matched pairs such as those in Eq 8.2 may occur rarely, making it difficult to estimate CMR nonparametrically. We explore this estimation issue next.

8.2.1. Estimation of CMR by Weighting, Matching, or Imputation

If the covariate space is discrete and low dimensional, nonparametric methods based on weighting or matching can be effective means of estimating CMR. Specifically, we can weight each pair of sample points i, j according to how likely the covariate combination $(\mathbf{X}_{1[i,\cdot]}, \mathbf{X}_{2[j,\cdot]})$ is to occur, as in

$$\hat{e}_{\text{weight}}(f) := \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} w(\mathbf{X}_{1[i,\cdot]}, \mathbf{X}_{2[j,\cdot]}) \times L\{f,(\mathbf{y}_{[j]}, \mathbf{X}_{1[i,\cdot]}, \mathbf{X}_{2[j,\cdot]})\},$$

where $w(x_1, x_2) := \frac{\mathbb{P}(X_1=x_1|X_2=x_2)}{\mathbb{P}(X_1=x_1)}$ is an importance weight (see also Hooker, 2007). Here, pairs of observations corresponding to unlikely or impossible combinations of covariates are

down-weighted or discarded, respectively. If the probabilities $\mathbb{P}(X_1 = x_1|X_2 = x_2)$ and $\mathbb{P}(X_1 = x_1)$ are known, then $\hat{e}_{\text{weight}}(f)$ is unbiased for $e_{\text{cond}}(f)$ (see Appendix A.7).

Alternatively, if X_2 is discrete and low dimensional, we can restrict estimates of $e_{\text{cond}}(f)$ to only consider pairs of sample observations in which X_2 is constant, or "matched," as in

$$\hat{e}_{\text{match}}(f) := \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \frac{1(\mathbf{X}_{2[j,\cdot]} = \mathbf{X}_{2[i,\cdot]})}{\mathbb{P}(X_2 = \mathbf{X}_{2[i,\cdot]})} \times L\{f, (\mathbf{y}_{[j]}, \mathbf{X}_{1[i,\cdot]}, \mathbf{X}_{2[j,\cdot]})\}. \tag{8.3}$$

This approach allows estimation of CMR without knowledge of the conditional distribution $\mathbb{P}(X_1 = x_1|X_2 = x_2)$. If the inverse probability weight $\mathbb{P}(X_2 = \mathbf{X}_{2[i,\cdot]})^{-1}$ is known, then $\hat{e}_{\text{match}}(f)$ is unbiased for $e_{\text{cond}}(f)$ (see Appendix A.7). The weight $\mathbb{P}(X_2 = \mathbf{X}_{2[i,\cdot]})^{-1}$ accounts for the fact that, for any given value x_2, the proportion of observations of X_2 taking the value x_2 will generally not be the same as the proportion of matched pairs $(X_2^{(a)}, X_2^{(b)})$ taking value the x_2, and so simply summing over all matched pairs would lead to bias. In practice, the proportion $\mathbb{P}(X_2 = \mathbf{X}_{2[i,\cdot]})$ can be approximated as $\frac{1}{n-1} \sum_{j' \neq i} 1(\mathbf{X}_{2[i,\cdot]} = \mathbf{X}_{2[j',\cdot]})$, with minor adjustments to Eq 8.3 to avoid dividing by zero. The resulting estimate is analogous to exact matching procedures commonly used in causal inference, which are known to work best when the covariates are discrete and low dimensional, in order for exact matches to be common (Stuart, 2010).

However, when the covariate space is continuous or high dimensional, we typically cannot estimate CMR nonparametrically. For such cases, we propose to estimate CMR under an assumption of homogeneous residuals. Specifically, we define μ_1 to be the conditional expectation function $\mu_1(x_2) = \mathbb{E}(X_1|X_2 = x_2)$, and assume that the random residual $X - \mu_1(X_2)$ is independent of X_2. Under this assumption, it can be shown that

$$e_{\text{cond}}(f) = \mathbb{E}L\left[f, (Y^{(b)}, \left\{X_1^{(a)} - \mu_1(X_2^{(a)})\right\} + \mu_1(X_2^{(b)}), X_2^{(b)})\right].$$

That is, $e_{\text{cond}}(f)$ is equal to the expected loss of f across random pairs of observations $(Z^{(a)}, Z^{(b)})$ in which the value *of the residual terms* (in curly braces) have been switched. Because of the independence assumption, no matching or weighting is required. If μ_1 is known, then we can again produce an unbiased estimate using the U-statistic

$$\hat{e}_{\text{impute}}(f) := \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} L\left[f, (\mathbf{y}_{[j]}, \left\{\mathbf{X}_{1[i,\cdot]} - \mu_1(\mathbf{X}_{2[i,\cdot]})\right\} + \mu_1(\mathbf{X}_{2[j,\cdot]}), \mathbf{X}_{2[j,\cdot]})\right].$$

This estimator aggregates over all pairs in our sample, switching the values of the residual terms (in curly braces) within each pair. In practice, when μ_1 is not known, an estimate of μ_1 can be achieved via regression or related machine learning techniques, and plugged in to the above equation. In this way, the assumption that $X - \mu_1(X_2) \perp X_2$ allows us to estimate CMR without explicitly modeling the joint distribution of X_1 and X_2.

In the existing literature, Strobl et al. (2008) introduce a similar procedure for estimating conditional variable importance. However, a formal comparison to Strobl et al. is complicated by the fact that the authors do not define a specific estimand, and that their approach is limited to tree-based regression models. Other existing approaches conditional importance approaches include methods for redefining X_1 and X_2 to induce approximate

independence, before computing an importance measure analogous to MR. This can be done by reducing the total number of covariates used, and hence reducing how well any one variable can be predicted by the others (as in Gregorutti et al., 2017). Alternatively, variables in X_2 that are predictive of X_1 can be regrouped directly into X_1 (as in Tolo̧si and Lengauer, 2011; see also the discussion from Kirk, Lewin and Stumpf, in Meinshausen and Bühlmann 2010).

In summary, CMR allows us to see how much a model relies on the information uniquely available in X_1. While CMR is more difficult to estimate than MR, several tractable approaches exist when X_2 is discrete, or when a homogenous residual assumption can be applied. One may also consider extending CMR by conditioning only on a subset of X_2. For example, we may consider conditioning only on elements of X_2 that are believed to causally effect X_1, by changing the outer expectation in Eq 8.2. For simplicity, we focus on the base case of estimating MR in this paper. Similar results could potentially be carried over for CMR as well.

9. Simulations

In this section, we first present a toy example to illustrate the concepts of MR, MCR, and AR. We then present a Monte Carlo simulation studying the effectiveness of bootstrap CIs for MCR.

9.1. Illustrative Toy Example with Simulated Data

To illustrate the concepts of MR, MCR, and AR (see Section 3.2), we consider a toy example where $X = (X_1, X_2) \in \mathbb{R}^2$, and $Y \in \{-1, 1\}$ is a binary group label. Our primary goal in this section is to build intuition for the differences between these three importance measures, and so we demonstrate them here only in a single sample. We focus on the empirical versions of our importance metrics ($\widehat{MR}$, $\widehat{MCR}_-$ and $\widehat{MCR}_+$), and compare them against AR, which is typically interpreted as an in-sample measure (Breiman, 2001), or as an intermediate step to estimate an alternate importance measure in terms of variable rankings (Gevrey et al., 2003; Olden et al., 2004).

We simulate $X|Y = -1$ from an independent, bivariate normal distribution with means $\mathbb{E}(X_1|Y = -1) = \mathbb{E}(X_2|Y = -1) = 0$ and variances $\text{Var}(X_1|Y = -1) = \text{Var}(X_2|Y = -1) = \frac{1}{9}$. We simulate $X|Y = 1$ by drawing from the same bivariate normal distribution, and then adding the value of a random vector $(C_1, C_2) := (\cos(U), \sin(U))$, where U is a random variable uniformly distributed on the interval $[-\pi, \pi]$. Thus, (C_1, C_2) is uniformly distributed across the unit circle.

Given a prediction model $f : \mathcal{X} \to \mathbb{R}$, we use the sign of $f(X_1, X_2)$ as our prediction of Y. For our loss function, we use the hinge loss $L(f, (y, x_1, x_2)) = (1 - yf(x_1, x_2))_+$, where $(a)_+ = a$ if $a \geq 0$ and $(a)_+ = 0$ otherwise. The hinge loss function is commonly used as a convex approximation to the zero-one loss $L(f, (y, x_1, x_2)) = 1[y \neq \text{sign}\{f(x_1, x_2)\}]$.

We simulate two samples of size 300 from the data generating process described above, one to be used for training, and one to be used for testing. Then, for the class of models

used to predict Y, we consider the set of degree-3 polynomial classifiers

$$\mathcal{F}_{d3} = \big\{ f_{\boldsymbol{\theta}} : f_{\boldsymbol{\theta}}(x_1, x_2) = \boldsymbol{\theta}_{[1]} + \boldsymbol{\theta}_{[2]} x_1 + \boldsymbol{\theta}_{[3]} x_2 \\ + \boldsymbol{\theta}_{[4]} x_1^2 + \boldsymbol{\theta}_{[5]} x_2^2 + \boldsymbol{\theta}_{[6]} x_1 x_2 \\ + \boldsymbol{\theta}_{[7]} x_1^3 + + \boldsymbol{\theta}_{[8]} x_2^3 + \boldsymbol{\theta}_{[9]} x_1^2 x_2 + \boldsymbol{\theta}_{[10]} x_1 x_2^2; \;\; \|\boldsymbol{\theta}_{[-1]}\|_2^2 \leq r_{d3} \big\},$$

where $\boldsymbol{\theta}_{[-1]}$ denotes all elements of $\boldsymbol{\theta}$ except $\boldsymbol{\theta}_{[1]}$, and where we set r_{d3} to the value that minimizes the 10-fold cross-validated loss in the training data. Let $\mathcal{A}_{d3}$ be the algorithm that minimizes the hinge loss over the (convex) feasible region $\{f_\theta : \|\boldsymbol{\theta}_{[-1]}\|_2^2 \leq r_{d3}\}$. We apply $\mathcal{A}_{d3}$ to the training data to determine a reference model f_{ref}. Also using the training data, we set ϵ equal to 0.10 multiplied by the cross-validated loss of $\mathcal{A}_{d3}$, such that $\mathcal{R}(\epsilon, f_{\text{ref}}, \mathcal{F}_{d3})$ contains all models in $\mathcal{F}_{d3}$ that exceed the loss of f_{ref} by no more than approximately 10% (see Eq 4.1). We then calculate empirical AR, MR, and MCR using the test observations.

We begin by considering the AR of $\mathcal{A}_{d3}$ on X_1. Calculating AR requires us to fit two separate models, first using all of the variables to fit a model on the training data, and then again using only X_2. In this case, the first model is equivalent to f_{ref}. We denote the second model as $\hat{f}_2$. To compute AR, we evaluate f_{ref} and $\hat{f}_2$ in the test observations. We illustrate this AR computation in Figure 6-A, marking the classification boundaries for f_{ref} and $\hat{f}_2$ by the black dotted line and the blue dashed lines respectively, and marking the test observations by labelled points ("x" for $Y = 1$, and "o" for $Y = -1$). Comparing the loss associated with these two models gives one form of AR–an estimate of the necessity of X_1 for the algorithm $\mathcal{A}_{d3}$. Alternatively, to estimate the *sufficiency* of X_1, we can compare the reference model f_{ref} against the model resulting from retraining algorithm $\mathcal{A}_{d3}$ only using X_1. We refer to this third model as $\hat{f}_1$, and mark its classification boundary by the solid blue lines in Figure 6-A.

Each of the classifiers in Figure 6-A can also be evaluated for its reliance on X_1, as shown in Figure 6-C. Here, we use $\hat{e}_{\text{divide}}$ in our calculation of $\widehat{MR}$ (see Eq 3.5). Unsurprisingly, the classifier fit without using X_1 (blue dashed line) has a model reliance of $\widehat{MR}(\hat{f}_2) = 1$. The reference model f_{ref} (dotted black line) has a model reliance of $\widehat{MR}(f_{\text{ref}}) = 3.47$. Each $\widehat{MR}$ value has an interpretation contained to a single model. That is, $\widehat{MR}$ compares a *single model's* behavior under different data distributions, rather than the AR approach of comparing *different models'* behavior on marginal distributions from a single joint distribution.

We illustrate MCR in Figure 6-B. In contrast to AR, MCR is only ever a function of well-performing prediction models. Here, we consider the empirical ϵ-Rashomon set $\hat{\mathcal{R}}(\epsilon, f_{\text{ref}}, \mathcal{F}_{d3})$, the subset of models in $\mathcal{F}_{d3}$ with test loss no more than ϵ above that of f_{ref}. We show the classification boundary associated with 15 well-performing models contained in $\hat{\mathcal{R}}(\epsilon, f_{\text{ref}}, \mathcal{F}_{d3})$ by the gray solid lines. We also show two of the models in $\hat{\mathcal{R}}(\epsilon, f_{\text{ref}}, \mathcal{F}_{d3})$ that approximately maximize and minimize empirical reliance on X_1 among models in $\hat{\mathcal{R}}(\epsilon, f_{\text{ref}}, \mathcal{F}_{d3})$. We denote these models as $\hat{f}_{+,\epsilon}$ and $\hat{f}_{-,\epsilon}$, and mark them by the solid green and dashed green lines respectively. For every model shown in Figure 6-B, we also mark its model reliance in Figure 6-C. We can then see from Figure 6-C that $\widehat{MR}$ for each model in $\hat{\mathcal{R}}(\epsilon, f_{\text{ref}}, \mathcal{F}_{d3})$ is contained between $\widehat{MR}(\hat{f}_{-,\epsilon})$ and $\widehat{MR}(\hat{f}_{+,\epsilon})$, up to a small approximation error.

In summary, unlike AR, MCR is only a function of models that fit the data well.

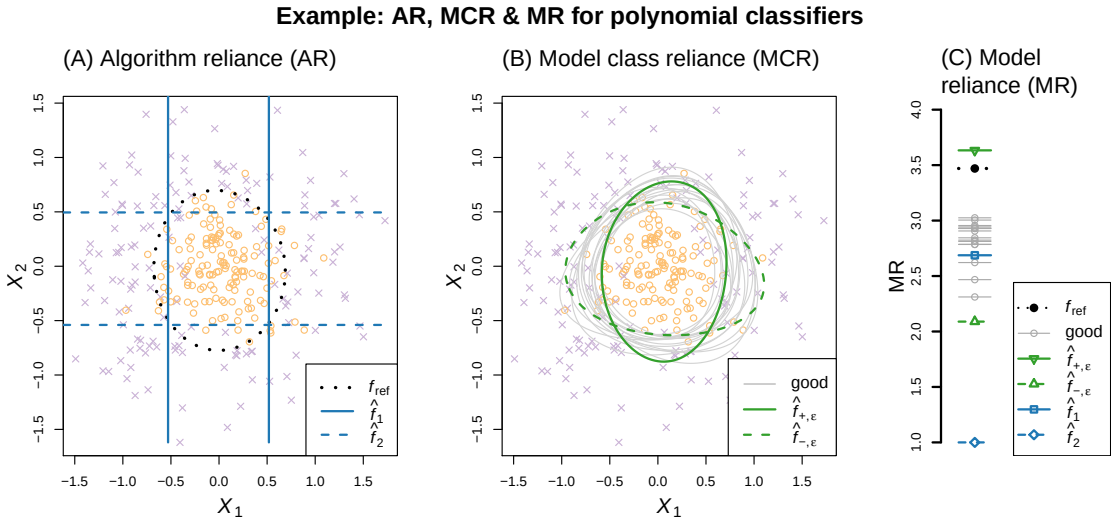

Figure 6: Example of AR, MCR & MR for polynomial classifiers. Panels (A) & (B) show the same 300 draws from a simulated data set, with the classification of each data point marked by "x" for $Y = 1$, and "o" for $Y = -1$. In Panel (A), for AR, we show single-feature models formed by dropping a covariate. Because these models take only a single input, we represent their classification boundaries as straight lines. In Panel (B), for MCR, we show the classification boundaries for several (two-feature) models with low in-sample loss. Of these models, the model with minimal dependence on X_1 is shown by the dashed green oval, and the model with maximal dependence on X_1 is shown by the solid green oval. Panel (C) shows the empirical model reliance on X_1 for each of the models in Panels (A) & (B). We see in Panel (C) that, as expected, no well-performing model relies (empirically) on X_1 more than $\hat{f}_{+,\epsilon}$ does, or relies (empirically) on X_1 less than $\hat{f}_{-,\epsilon}$ does. That is, no well-performing model has an empirical MR value greater than $\widehat{MCR}_+(\epsilon)$, or less than $\widehat{MCR}_-(\epsilon)$.

9.2. Simulations of Bootstrap Confidence Intervals

In this section we study the performance of MCR under model class misspecification. Our goal will be to estimate how much the conditional expectation function $f_0(x) = \mathbb{E}(Y|X = x)$ relies on subsets of covariates. Given a reference model f_{ref} and model class $\mathcal{F}$, our ability to describe $MR(f_0)$ will hinge on two conditions:

Condition 20 (Nearly correct model class) *The class $\mathcal{F}$ contains a well-performing model $\tilde{f} \in \mathcal{R}(\epsilon, f_{ref}, \mathcal{F})$ satisfying $MR(\tilde{f}) = MR(f_0)$ (see Eq 4.1).*

Condition 21 (Bootstrap coverage) *Bootstrap CIs for empirical MCR give appropriate coverage of population-level MCR.*

Condition 20 ensures that the interval $[MCR_-(\epsilon), MCR_+(\epsilon)]$ contains $MR(f_0)$, and Condition 21 ensures that this interval can be estimated in finite samples. Condition 20 can also be interpreted as saying that the model reliance value of $MR(f_c)$ is "well supported" by the class $\mathcal{F}$, even if $\mathcal{F}$ does not contain f_0. Our primary goal is to assess whether CIs derived from MCR can give appropriate coverage of $MR(f_0)$, which depends on both conditions. As a secondary goal, we also would like to be able to assess Conditions 20 & 21 individually.

Verifying the above conditions requires that we are able to calculate population-level MCR. To this end, we draw samples with replacement from a finite population of 20,000 observations, in which MCR can also be calculated directly. To derive a CI based on MCR, we divide each simulated sample $\mathcal{Z}_s$ into a training subset and analysis subset. We use the training subset to fit a reference model $f_{\text{ref},s}$, which is required for our definition of population-level MCR. We calculate a bootstrap CI by drawing 500 bootstrap samples from the analysis subset, and computing $\widehat{MCR}_-(\epsilon)$ and $\widehat{MCR}_+(\epsilon)$ in each bootstrap sample by optimizing over $\hat{\mathcal{R}}(\epsilon, f_{\text{ref},s}, \mathcal{F})$. We then take the 2.5% percentile of $\widehat{MCR}_-(\epsilon)$ values across bootstrap samples, and the 97.5% percentile of $\widehat{MCR}_+(\epsilon)$ values across bootstrap samples, as the lower and upper endpoints of our CI, respectively. We repeat this procedure for both X_1 and X_2.

We generate data according to a model with increasing amounts of nonlinearity. For $\gamma \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$, we simulate continuous outcomes as $Y = f_0(X) + E$, where f_0 is the function $f_0(\mathbf{x}) = \sum_{j=1}^{p} j\mathbf{x}_{[j]} - \gamma\mathbf{x}_{[j]}^2$; the covariate dimension p is equal to 2, with X_1 and X_2 defined as the first and second elements of X; the covariates X are drawn from a multivariate normal distribution with $\mathbb{E}(X_1) = \mathbb{E}(X_2) = 0$, $\text{Var}(X_1) = \text{Var}(X_2) = 0$, and $\text{Cov}(X_1, X_2) = 1/4$; and E is a normally distributed noise variable with mean zero and variance equal to $\sigma_E^2 := \text{Var}(f_0(X))$. We consider sample sizes of $n = 400$ and 800, of which $n_{tr} = 200$ or 300 observations are assigned to the training subset respectively.

To implement our approach, we use the model class $\mathcal{F}_{\text{lm}} = \{f_{\boldsymbol{\beta}} : f_{\boldsymbol{\beta}}(\mathbf{x}) = \boldsymbol{\beta}_{[1]} + \sum_{j=1}^{2} \mathbf{x}_{[j]}\boldsymbol{\beta}_{[j+1]}, \boldsymbol{\beta} \in \mathbb{R}^3\}$. We set the performance threshold ϵ equal to $0.1 \times \sigma_E^2$. We refer to this MCR implementation with $\mathcal{F}_{\text{lm}}$ as "MCR-Linear."

As a comparator method, we consider a simpler bootstrap approach, which we refer to as "Standard-Linear." Here, we take 500 bootstrap samples from the simulated data $\mathcal{Z}_s$. In each bootstrap sample, indexed by b, we set aside n_{tr} training points to train a model $f_b \in \mathcal{F}_{\text{lm}}$, and calculate $\widehat{MR}(f_b)$ from the remaining data points. We then create a 95% bootstrap percentile CI for $MR(f_0)$ by taking the 2.5% and 97.5% percentiles of $\widehat{MR}(f_b)$ across $b = 1, \ldots, 500$.

9.2.1. Results

Overall, we find that MCR provides more robust and conservative intervals for the reliance of f_0 on X_1 and X_2, relative to standard bootstrap approaches. We also find that higher sample size generally exacerbates coverage errors due to misspecification, as methods become more certain of biased results.

MCR-Linear gave proper coverage for up to moderate levels of misspecification ($\gamma = 0.3$), where Standard-Linear began to break down (Figure 7). For larger levels of misspecification ($\gamma \geq 0.4$), both MCR-Linear and Standard-Linear failed to give appropriate coverage.

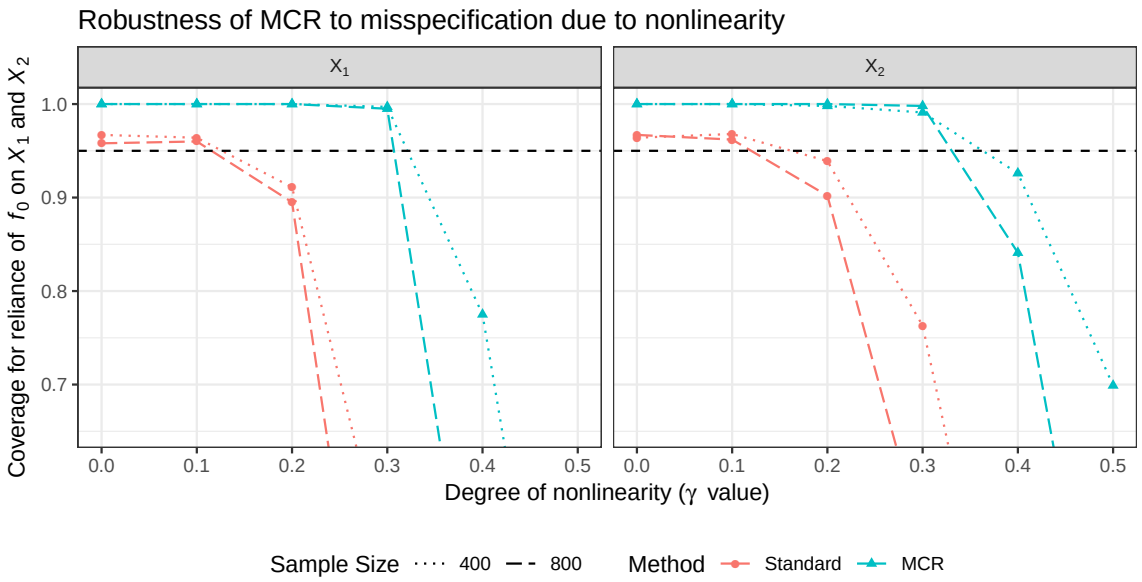

Figure 7: MR Coverage - The y-axis shows coverage rate for the reliance of f_0 on either X_1 (left column) or X_2 (right column), where X_2 is simulated to be more influential that X_1. The x-axis shows increasing levels of misspecification (γ). All methods aim to have at least 95% coverage for each scenario (dashed horizontal line).

The increased robustness of MCR comes at the cost of wider CIs. Intervals for MCR-Linear were typically larger than intervals for Standard-Linear by a factor of approximately 2-4. This is partly due to the fact that CIs for MCR are meant to cover the range of values $[MCR_{-}(\epsilon), MCR_{+}(\epsilon)]$ (defined using $f_{\text{ref},s}$), rather than to cover a single point.

When investigating Conditions 20 & 21 individually, we find that the coverage errors for MCR-Linear were largely attributable to violations of Condition 20. Condition 21 appears to hold conservatively for all scenarios studied–within each scenario, at least 95.9% of bootstrap CIs contained population-level MCR.

These simulation results highlight an aspect of MCR that is both a strength and a weakness: MCR is generic. MCR does not assume a particular means by which misspecification may occur, and is less powerful than sensitivity analyses which make that assumption correctly. Nonetheless, MCR still appears to add robustness. For sufficiently strong signals, an informative interval may still be returned. In our applied data analysis, below, we see that this is indeed the case.

10. Data Analysis: Reliance of Criminal Recidivism Prediction Models on Race and Sex

Evidence suggests that bias exists among judges and prosecutors in the criminal justice system (Spohn, 2000; Blair et al., 2004; Paternoster and Brame, 2008). In an aim to counter this bias, machine learning models trained to predict recidivism are increasingly

being used to inform judges' decisions on pretrial release, sentencing, and parole (Monahan and Skeem, 2016; Picard-Fritsche et al., 2017). Ideally, prediction models can avoid human bias and provide judges with empirically tested tools. But prediction models can also mirror the biases of the society that generates their training data, and perpetuate the same bias at scale. In the case of recidivism, if arrest rates across demographic groups are not representative of underlying crime rate (Beckett et al., 2006; Ramchand et al., 2006; U.S. Department of Justice - Civil Rights Devision, 2016), then bias can be created in both (1) the outcome variable, future crime, which is measured imperfectly via arrests or convictions, and (2) the covariates, which include the number of prior convictions on a defendant's record (Corbett-Davies et al., 2016; Lum and Isaac, 2016). Further, when a prediction model's behavior and mechanisms are an opaque black box, the model can evade scrutiny, and fail to offer recourse or explanations to individuals rated as "high risk."

We focus here on the issue of transparency, which takes an important role in the recent debate about the proprietary recidivism prediction tool COMPAS (Larson et al., 2016; Corbett-Davies et al., 2016). While COMPAS is known to not rely explicitly on race, there is concern that it may rely implicitly on race via proxies–variables statistically dependent with race (see further discussion in Section 11).

Our goal is to identify bounds for how much COMPAS relies on different covariate subsets, either implicitly or explicitly, under certain assumptions (defined below). We analyze a public data set of defendants from Broward County, Florida, in which COMPAS scores have been recorded (Larson et al., 2016). Within this data set, we only included defendants measured as African-American or Caucasian (3,373 in total) due to sparseness in the remaining categories. The outcome of interest (Y) is the COMPAS violent recidivism score. Of the available covariates, we consider three variables which we refer to as "admissible": an individual's age, their number of priors, and an indicator of whether the current charge is a felony. We also consider two variables which we refer to as "inadmissible": an individual's race and sex. Our labels of "admissible" and "inadmissible" are not intended to be legally precise–indeed, the boundary between these types of labels is not always clear (see Section 10.2). We compute empirical MCR and AR for each variable group, as well as bootstrap CIs for MCR (see Section 9.2).

To compute empirical MCR and AR, we consider a flexible class of linear models in a RKHS to predict the COMPAS score (described in more detail below). Given this class, the MCR range (See Eq 2.2) captures the highest and lowest degree to which any model in the class may rely on each covariate subset. We assume that our class contains at least one model that relies on "inadmissible variables" to the same extent that COMPAS relies either on "inadmissible variables" or on proxies that are unmeasured in our sample (analogous to Condition 20). We make the same assumption for "admissible variables." These assumptions can be interpreted as saying that the reliance values of COMPAS are relatively "well supported" by our chosen model class, and allows us to identify bounds on the MR values for COMPAS. We also consider the more conventional, but less robust approach of AR (Section 3.2), that is, how much would the accuracy suffer for a model-fitting algorithm trained on COMPAS score if a variable subset was removed?

These computations require that we predefine our loss function, model class, and performance threshold. We define MR, MCR, and AR in terms of the squared error loss $L(f, (y, x_1, x_2)) = \{y - f(x_1, x_2)\}^2$. We define our model class $\mathcal{F}_{\mathbf{D}, r_k}$ in the form of Eq 7.6,

where we determine $\mathbf{D}$, μ, k, and r_k based on a subset $\mathcal{S}$ of 500 training observations. We set $\mathbf{D}$ equal to the matrix of covariates from $\mathcal{S}$; we set μ equal to the mean of Y in $\mathcal{S}$; we set k equal to the radial basis function $k_{\sigma_s}(\mathbf{x}, \tilde{\mathbf{x}}) = \exp\left(-\frac{\|\mathbf{x}-\tilde{\mathbf{x}}\|^2}{2\sigma_s}\right)$, where we choose σ_s to minimize the cross-validated loss of a Nadaraya-Watson kernel regression (Hastie et al., 2009) fit to $\mathcal{S}$; and we select the parameters r_k by cross-validation on $\mathcal{S}$. We set ϵ equal to 0.1 times the cross-validated loss on $\mathcal{S}$. Also using $\mathcal{S}$, we train a reference model $f_{\text{ref}} \in \mathcal{F}_{\mathbf{D},r_k}$. Using the held-out 2,873 observations, we then estimate $MR(f_{\text{ref}})$ and MCR for $\mathcal{F}_{\mathbf{D},r_k}$. To calculate AR, we train models from $\mathcal{F}_{\mathbf{D},r_k}$ using $\mathcal{S}$, and evaluate their performance in the held-out observations.

10.1. Results

Our results imply that race and sex play somewhere between a null role and a modest role in determining COMPAS score, but that they are less important than "admissible" factors (Figure 8). As a benchmark for comparison, the empirical MR of f_{ref} is equal to 1.09 for "inadmissible variables," and 2.78 for "admissible variables." The AR is equal to 0.94 and 1.87 for "inadmissible" and "admissible" variables respectively, roughly in agreement with MR. The MCR range for "inadmissible variables" is equal to [1.00,1.56], indicating that *for any model in* $\mathcal{F}_{\mathbf{D},r_k}$ with empirical loss no more than ϵ above that of f_{ref}, the model's loss can increase by no more than 56% if race and sex are permuted. Such a statement cannot be made solely based on AR or MR methods, as these methods do not upper bound the reliance values of well-performing models. The bootstrap 95% CI for MCR on "inadmissible variables" is [1.00, 1.73]. Thus, *under our assumptions, if COMPAS relied on sex, race, or their unmeasured proxies by a factor greater than 1.73, then intervals as low as what we observe would occur with probability* < 0.05.

For "admissible variables" the MCR range is equal to [1.77,3.61], with a 95% bootstrap CI of [1.62, 3.96]. *Under our assumptions, this implies if COMPAS relied on age, number of priors, felony indication, or their unmeasured proxies by a factor lower than 1.77, then intervals as high as what we observe would occur with probability* < 0.05. This result is consistent with Rudin et al. (2019), who find age to be highly predictive of COMPAS score.

It is worth noting that the upper limit of 3.61 maximizes empirical MR on "admissible variables" not only among well-performing models, but globally across all models in the class (see Figure 8, and Eq 6.5). In other words, it is not possible to find models in $\mathcal{F}_{\mathbf{D},r_k}$ that perform arbitrarily poorly on perturbed data, but still perform well on unperturbed data, and so the ratio of $\hat{e}_{\text{switch}}(f)$ to $\hat{e}_{\text{orig}}(f)$ has a finite upper bound. Because the regularization constraints of $\mathcal{F}_{\mathbf{D},r_k}$ preclude MR values higher than 3.61, the MR of COMPAS on "admissible variables" may be underestimated by empirical MCR. Note also that both MCR intervals are left-truncated at 1, as it is often sufficiently precise to conclude that there exists a well-performing model with no reliance on the variables of interest (that is, MR equal to 1; see Appendix A.2).

10.2. Discussion & Limitations

Asking whether a proprietary model relies on sex and race, after adjusting for other covariates, is related to the fairness metric known as conditional statistical parity (CSP). A decision rule satisfies CSP if its decisions are independent of a sensitive variable, conditional

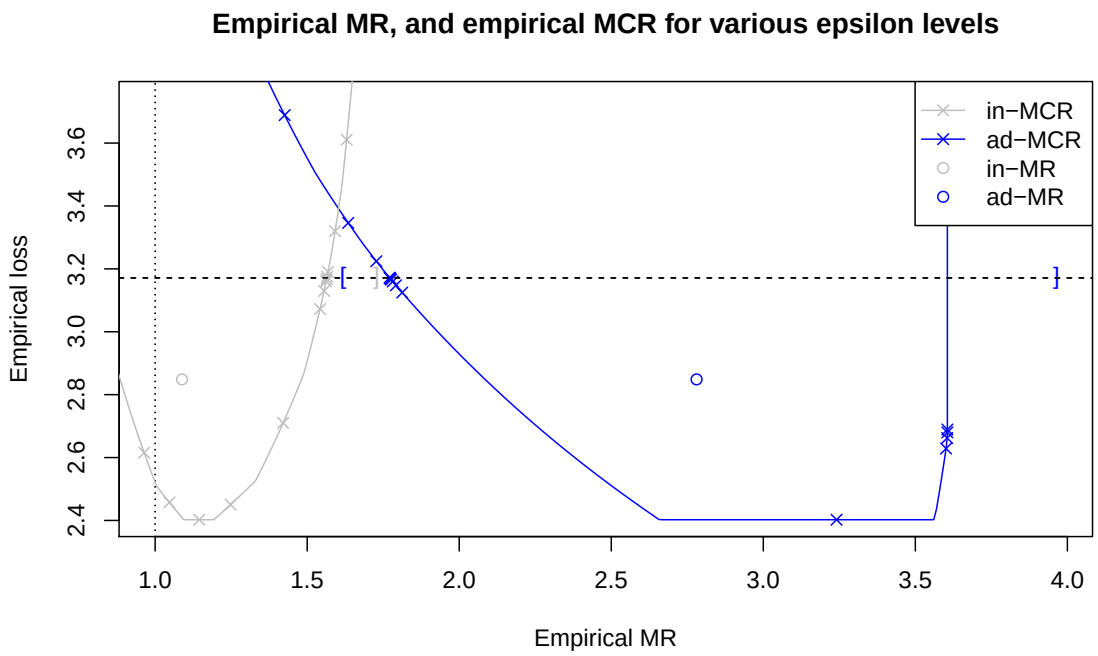

Figure 8: Empirical MR and MCR for Broward County criminal records data set - For any prediction model f, the y-axis shows empirical loss ($\hat{e}_{\text{stnd}}(f)$) and the x-axis shows empirical reliance ($\widehat{MR}(f)$) on each covariate subset. Null reliance (MR equal to 1.0) is marked by the vertical dotted line. Reliances on different covariate subsets are marked by color ("admissible" = blue; "inadmissible" = gray). For example, model reliance values for f_{ref} are shown by the two circular points, one for "admissible" variables and one for "inadmissible" variables. MCR for different values of ϵ can be represented as boundaries on this coordinate space. To this end, for each covariate subset, we compute conservative boundary functions (shown as solid lines, or "bowls") guaranteed to contain *all models in the class* (see Section 6). Specifically, all models in $f \in \mathcal{F}_{\mathbf{D},r_k}$ are guaranteed to have an empirical loss ($\hat{e}_{\text{stnd}}(f)$) and empirical MR value ($\widehat{MR}(f)$) for "inadmissible variables" corresponding to a point within the gray bowl. Likewise, all models in $\mathcal{F}_{\mathbf{D},r_k}$ are guaranteed to have an empirical loss and empirical MR value for "admissible variables" corresponding to a point within the blue bowl. Points shown as "$\times$" represent additional models in $\mathcal{F}_{\mathbf{D},r_k}$ discovered during our computational procedure, and thus show where the "bowl" boundary is tight. The goal of our computation procedure (see Section 6) is to tighten the boundary as much as possible near the ϵ value of interest, shown by the dashed horizontal line above. This dashed line has a y-intercept equal to the loss of the reference model plus the ϵ value of interest. Bootstrap CIs for $MCR_{-}(\epsilon)$ and $MCR_{+}(\epsilon)$ are marked by brackets.

on a set of "legitimate" covariates C (Corbett-Davies et al., 2017; see also Kamiran et al., 2013). Roughly speaking, CSP reflects the idea that groups of people with similar covariates C are treated similarly (Dwork et al., 2012), regardless of the sensitive variable (for example, race or sex). However, the criteria becomes superficial if too many variables are included in C, and care should be taken to avoid including proxies for the sensitive variables. Several other fairness metrics have also been proposed, which often form competing objectives (Kleinberg et al., 2017; Chouldechova, 2017; Nabi and Shpitser, 2018; Corbett-Davies et al., 2017). Here, if COMPAS was not influenced by race, sex, or variables related to race or sex (conditional on a set of "legitimate" variables), it would satisfy CSP.

Unfortunately, it is often difficult to distinguish between "legitimate" (or "admissible") variables and "illegitimate" variables. Some variables function both as part of a reasonable predictor for risk, and, separately, as a proxy for race. Because of disproportional arrest rates, particularly for misdemeanors and drug-related offenses (U.S. Department of Justice - Civil Rights Devision, 2016; Lum and Isaac, 2016), prior misdemeanor convictions may act as such a proxy (Corbett-Davies et al., 2016; Lum and Isaac, 2016).

Proxy variables for race (defined as being statistically dependent with race) that are unmeasured in our sample are also not the only reason that race could be predictive of COMPAS score. Other inputs to the COMPAS algorithm might be associated with race *only conditionally* on variables we categorize as "admissible." However, our result from Section 10.1 that race has limited predictive utility for COMPAS score suggests that such conditional relationships are also limited.

11. Conclusion

In this article, we propose MCR as the upper and lower limit on how important a set of variables can be to any well-performing model in a class. In this way, MCR provides a more comprehensive and robust measure of importance than traditional importance measures for a single model. We derive bounds on MCR, which motivate our choice of point estimates. We also derive connections between permutation importance, U-statistics, conditional variable importance, and conditional causal effects. We apply MCR in a data set of criminal recidivism, in order to help inform the characteristics of the proprietary model COMPAS.

Several exciting areas remain open for future research. One research direction closely related to our current work is the development of exact or approximate MCR computation procedures for other model classes and loss functions. We have shown that, for model classes where minimizing the empirical loss is a convex optimization problem, MCR can be conservatively computed via a series of convex optimization problems. Further, we have shown that computing $\widehat{MCR}_-$ is often no more challenging that minimizing the empirical loss over a reweighted sample. General computation procedures for MCR are still an open research area.

Another direction is to consider MCR for variable selection. If MCR_+ is small for a variable, then no well-performing predictive model can heavily depend on that variable, indicating that it can be eliminated.

Our theoretical analysis of Rashomon sets depends on $\mathcal{F}$ and f_{ref} being prespecified. Above, we have actualized this by splitting our sample into subsets of size n_1 and n_2, using the first subset to determine $\mathcal{F}$ and f_{ref}, and conditioning on $\mathcal{F}$ and f_{ref} when estimating

MCR in the second subset. As a result, the boundedness constants in our assumptions (B_{ind}, $B_{\text{ref}}, B_{\text{switch}}$, and b_{orig}) depend on $\mathcal{F}$, and hence on n_1. However, because our results are non-asymptotic, we have not explored how Rashomon sets behave when n_1 and n_2 grow at different rates. An exciting future extension of this work is to study sequences of triples $\{\epsilon_{n_1}, f_{\text{ref},n_1}, \mathcal{F}_{n_1}\}$ that change as n_1 increases, and the corresponding Rashomon sets $\mathcal{R}(\epsilon_{n_1}, f_{\text{ref},n_1}, \mathcal{F}_{n_1})$, as this may more thoroughly capture how model classes are determined by analysts.

While we develop Rashomon sets with the goal of studying MR, Rashomon sets can also be useful for finite sample inferences about a wide variety of other attributes of best-in-class models (for example, Section 5). Characterizations of a Rashomon set itself may also be of interest. For example, in ongoing work, we are studying the size of a Rashomon set, and its connection to generalization of models and model classes (Semenova and Rudin, 2019). We are additionally developing methods for visualizing Rashomon sets (Dong and Rudin, 2019).

Acknowledgments

Support for this work was provided by the National Institutes of Health (grants P01CA134294, R01GM111339, R01ES024332, R35CA197449, R01ES026217, P50MD010428, DP2MD012722, R01MD012769, & R01ES028033), by the Environmental Protection Agency (grants 83615601 & 83587201-0), and by the Health Effects Institute (grant 4953-RFA14-3/16-4).

Appendix A. Miscellaneous Supplemental Sections

All labels for items in the following appendices begin with a letter (for example, Section A.2), while references to items in the main text contain only numbers (for example, Proposition 19).

A.1. Code

R code for our example in Section 9.1 and analysis in Section 10 is available at `https://github.com/aaronjfisher/mcr-supplement`.

A.2. Model Reliance Less than 1

While it is counterintuitive, it is possible for the expected loss of a prediction model to *decrease* when the information in X_1 is removed. Roughly speaking, a "pathological" model f_{silly} may use the information in X_1 to "intentionally" misclassify Y, such that $e_{\text{switch}}(f_{\text{silly}}) < e_{\text{orig}}(f_{\text{silly}})$ and $MR(f_{\text{silly}}) < 1$. The model f_{silly} may even be included in a population ϵ-Rashomon set (see Section 4) if it is still possible to predict Y sufficiently well from the information in X_2.

However, in these cases there will often exist another model that outperforms f_{silly}, and that has MR equal to 1 (i.e., no reliance on X_1). To see this, consider the case where $\mathcal{F} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \mathbb{R}^d\}$ is indexed by a parameter $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}_{\text{silly}}$ and $\boldsymbol{\theta}^\star$ be parameter values

such that $f_{\boldsymbol{\theta}_{\text{silly}}}$ is equivalent to f_{silly}, and $f_{\boldsymbol{\theta}^\star}$ is the best-in-class model. If $f_{\boldsymbol{\theta}^\star}$ satisfies $MR(f_{\boldsymbol{\theta}^\star}) > 1$ and if the model reliance function MR is continuous in $\boldsymbol{\theta}$, then there exists a parameter value $\boldsymbol{\theta}_1$ between $\boldsymbol{\theta}_{\text{silly}}$ and $\boldsymbol{\theta}^\star$ such that $MR(f_{\boldsymbol{\theta}_1}) = 1$. Further, if the loss function L is convex in $\boldsymbol{\theta}$, then $e_{\text{orig}}(f_{\boldsymbol{\theta}^\star}) \leq e_{\text{orig}}(f_{\boldsymbol{\theta}_1}) \leq e_{\text{orig}}(f_{\text{silly}})$, and any population ϵ-Rashomon set containing f_{silly} will also contain $f_{\boldsymbol{\theta}_1}$.

A.3. Relating $\hat{e}_{\text{switch}}(f)$ to All Possible Permutations of the Sample

Following the notation in Section 3, let $\{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{n!}\}$ be a set of n-length vectors, each containing a different permutation of the set $\{1, \ldots, n\}$. We show in this section that $\hat{e}_{\text{switch}}(f)$ is equal to the product of

$$\sum_{l=1}^{n!} \sum_{i=1}^{n} L\{f, (\mathbf{y}_{[i]}, \mathbf{X}_{1[\boldsymbol{\pi}_{l[i]},\cdot]}, \mathbf{X}_{2[i,\cdot]})\} 1(\boldsymbol{\pi}_{l[i]} \neq i), \tag{A.1}$$

and a proportionality constant that is only a function of n.

First, consider the sum

$$\sum_{l=1}^{n!} \sum_{i=1}^{n} L\{f, (\mathbf{y}_{[i]}, \mathbf{X}_{1[\boldsymbol{\pi}_{l[i]},\cdot]}, \mathbf{X}_{2[i,\cdot]})\}, \tag{A.2}$$

which omits the indicator function found in Eq A.1.

The summation in Eq A.2 contains $n(n!)$ terms, each of which is a two-way combination of the form $L\{f, (\mathbf{y}_{[i]}, \mathbf{X}_{1[j,\cdot]}, \mathbf{X}_{2[i,\cdot]})\}$ for $i, j \in \{1, \ldots n\}$. There are only n^2 unique combinations of this form, and each must occur in at least $(n-1)!$ of the $n(n!)$ terms in Eq A.2. To see this, consider selecting two integer values $\tilde{i}, \tilde{j} \in \{1, \ldots n\}$, and enumerating all occurrences of the term $L\{f, (\mathbf{y}_{[\tilde{i}]}, \mathbf{X}_{1[\tilde{j},\cdot]}, \mathbf{X}_{2[\tilde{i},\cdot]})\}$ within the sum in Eq A.2. Of the permutation vectors $\{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{n!}\}$, we know that $(n-1)!$ of them place $\tilde{i}$ in the $\tilde{j}^{th}$ position, i.e., that satisfy $\boldsymbol{\pi}_{l[\tilde{i}]} = \tilde{j}$. For each such permutation $\boldsymbol{\pi}_l$, the inner summation in Eq A.2 over all possible values of i must include the term $L\{f, (\mathbf{y}_{[\tilde{i}]}, \mathbf{X}_{1[\boldsymbol{\pi}_{l[\tilde{i}]},\cdot]}, \mathbf{X}_{2[\tilde{i},\cdot]})\} = L\{f, (\mathbf{y}_{[\tilde{i}]}, \mathbf{X}_{1[\tilde{j},\cdot]}, \mathbf{X}_{2[\tilde{i},\cdot]})\}$. Thus, Eq A.2 contains at least $(n-1)!$ occurrences of the term $L\{f, (\mathbf{y}_{[\tilde{i}]}, \mathbf{X}_{1[\tilde{j},\cdot]}, \mathbf{X}_{2[\tilde{i},\cdot]})\}$.

So far, we have shown that each unique combination occurs at least $(n-1)!$ times, but it also follows that each unique combination must occur precisely $(n-1)!$ times. This is because each of the n^2 unique combinations must occur at least $(n-1)!$ times, which accounts for $n^2((n-1)!) = n(n!)$ terms in total. As noted above, Eq has A.2 has only $n(n)!$ terms, so there can be no additional terms. We can then simplify Eq A.2 as

$$\sum_{l=1}^{n!} \sum_{i=1}^{n} L\{f, (\mathbf{y}_{[i]}, \mathbf{X}_{1[\boldsymbol{\pi}_{l[i]},\cdot]}, \mathbf{X}_{2[i,\cdot]})\} = (n-1)! \sum_{i=1}^{n} \sum_{j=1}^{n} L\{f, (\mathbf{y}_{[i]}, \mathbf{X}_{1[j,\cdot]}, \mathbf{X}_{2[i,\cdot]})\}.$$

By the same logic, we can simplify Eq A.1 as

$$\begin{aligned}
&\sum_{l=1}^{n!}\sum_{i=1}^{n} L\{f,(\mathbf{y}_{[i]},\mathbf{X}_{1[\boldsymbol{\pi}_{l[i]},\cdot]},\mathbf{X}_{2[i,\cdot]})\}1(\boldsymbol{\pi}_{l[i]}\neq i)\\
=&(n-1)!\left\{\sum_{i=1}^{n}\sum_{j=1}^{n} L\{f,(\mathbf{y}_{[i]},\mathbf{X}_{1[j,\cdot]},\mathbf{X}_{2[i,\cdot]})\}1(j\neq i)\right\}\\
=&(n-1)!\sum_{i=1}^{n}\sum_{j\neq i} L\{f,(\mathbf{y}_{[i]},\mathbf{X}_{1[j,\cdot]},\mathbf{X}_{2[i,\cdot]})\}, \qquad \text{(A.3)}
\end{aligned}$$

and Line A.3 is proportional to $\hat{e}_{\text{switch}}(f)$ up to a function of n.

A.4. Bound for MR of the Best-in-class Prediction Model

Although describing individual models is not the primary focus of this work, a corollary of Theorem 4 is that we can create a probabilistic bound for the reliance of the (unknown) best-in-class model $f^\star$ on X_1.

Corollary 22 *(Bound on Best-in-class MR) Let $f^\star \in \arg\min_{f\in\mathcal{F}} e_{orig}(f)$ be a prediction model that attains the lowest possible expected loss, and let $f_{+,\epsilon}$ and $f_{-,\epsilon}$ be defined as in Theorem 4. If $f_{+,\epsilon}$ and $f_{-,\epsilon}$ satisfy Assumptions 1, 2 and 3, then*

$$\mathbb{P}\left(MR(f^\star)\in\left[\widehat{MCR}_-\left(\epsilon_{best}\right)-\mathcal{Q}_{best},\quad \widehat{MCR}_+\left(\epsilon_{best}\right)+\mathcal{Q}_{best}\right]\right)\geq 1-\delta,$$

where $\epsilon_{best} := 2B_{ref}\sqrt{\frac{\log(6\delta^{-1})}{2n}}$ *, and* $\mathcal{Q}_{best} := \frac{B_{switch}}{b_{orig}} - \frac{B_{switch}-B_{ind}\sqrt{\frac{\log(12\delta^{-1})}{n}}}{b_{orig}+B_{ind}\sqrt{\frac{\log(12\delta^{-1})}{2n}}}$.

The above result does not require that $f^\star$ be unique. If several models achieve the minimum possible expected loss, the above boundaries apply simultaneously for each of them. In the special case when the true conditional expectation function $\mathbb{E}(Y|X_1,X_2)$ is equal to $f^\star$, then we have a boundary for the reliance of the function $\mathbb{E}(Y|X_1,X_2)$ on X_1. This reliance bound can also be translated into a causal statement using Proposition 19.

A.5. Ratios versus Differences in MR Definition

We choose our ratio-based definition of model reliance, $MR(f) = \frac{e_{\text{switch}}(f)}{e_{\text{orig}}(f)}$, so that the measure can be comparable across problems, regardless of the scale of Y. However, several existing works define VI measures in terms of differences (Strobl et al., 2008; Datta et al., 2016; Gregorutti et al., 2017), analogous to

$$MR_{\text{difference}}(f) := e_{\text{switch}}(f) - e_{\text{orig}}(f). \qquad \text{(A.4)}$$

While this difference measure is less readily interpretable, it has several computational advantages. The mean, variance, and asymptotic distribution of the estimator $\widehat{MR}_{\text{difference}}(f) := \hat{e}_{\text{switch}}(f) - \hat{e}_{\text{orig}}(f)$ can be easily determined using results for U-statistics, without the use of the delta method (Dorfman, 1938; Lehmann and Casella, 2006; see also Ver Hoef, 2012).

Estimates in the form of $\widehat{MR}_{\text{difference}}(f)$ will also be more stable when $\min_{f \in \mathcal{F}} e_{\text{orig}}(f)$ is small, relative to estimates for the ratio-based definition of MR. To improve interpretability, we may also normalize $MR_{\text{difference}}(f)$ by dividing by the variance of Y, which can be easily estimated without the use of models, as in Williamson et al. (2017).

Under the difference-based definition for MR (Eq A.4), the results from Theorem 4, Theorem 6, and Corollary 22 will still hold under the following modified definitions of $\mathcal{Q}_{\text{out}}$, $\mathcal{Q}_{\text{in}}$, and $\mathcal{Q}_{\text{best}}$:

$$\mathcal{Q}_{\text{out,difference}} := \left(1 + \frac{1}{\sqrt{2}}\right) B_{\text{ind}} \sqrt{\frac{\log(6\delta^{-1})}{n}},$$

$$\mathcal{Q}_{\text{in,difference}} := B_{\text{ind}} \left\{ \sqrt{\frac{\log(8\delta^{-1}\mathcal{N}\left(\mathcal{F}, r\sqrt{2}\right))}{n}} + \sqrt{\frac{\log(8\delta^{-1}\mathcal{N}\left(\mathcal{F}, r\right))}{2n}} \right\} + 2r(\sqrt{2}+1), \text{ and}$$

$$\mathcal{Q}_{\text{best,difference}} := \left(1 + \frac{1}{\sqrt{2}}\right) B_{\text{ind}} \sqrt{\frac{\log(12\delta^{-1})}{n}}.$$

Respectively replacing $\mathcal{Q}_{\text{out}}$, $\mathcal{Q}_{\text{in}}$, $\mathcal{Q}_{\text{best}}$, MR, and $\widehat{MR}$ with $\mathcal{Q}_{\text{out,difference}}$, $\mathcal{Q}_{\text{in,difference}}$, $\mathcal{Q}_{\text{best,difference}}$, $MR_{\text{difference}}$ and $\widehat{MR}_{\text{difference}}$ entails only minor changes to the corresponding proofs (see Appendices B.3, B.5, and B.4). The results will also hold without Assumption 3, as is suggested by the fact that b_{orig} and B_{switch} do not appear in $\mathcal{Q}_{\text{out,difference}}$, $\mathcal{Q}_{\text{in,difference}}$, or $\mathcal{Q}_{\text{best,difference}}$.

We also prove an analogous version of Theorem 5, on uniform bounds for $\widehat{MR}_{\text{difference}}$, in Appendix B.5.1.

A.6. Rashomon Sets and Profile Likelihood Intervals

We note in Section 5.1 that, under certain conditions, the CIs returned from Proposition 7 take the same form as profile likelihood CIs (Coker et al., 2018). For completeness, we briefly review this connection. We assume here that models $f_\theta \in \mathcal{F}$ are indexed by a finite dimensional parameter vector $\theta \in \Theta$, where $\theta = (\gamma, \psi)$ contains a 1-dimensional parameter of interest $\gamma \in \mathbb{R}^1$, and a nuisance parameter $\psi \in \Psi$. We further assume and that $e_{\text{orig}}(f_\theta)$ is minimized by a unique parameter value $\theta^\star = (\gamma^\star, \psi^\star) \in \Theta$, and that our goal is to learn about $\gamma^\star$.

If $s_\theta := \int_{\mathcal{Z}} \exp\{-L(f_\theta, z)\} dz$ is finite for all $\theta \in \Theta$, we can convert L into the likelihood function $\mathcal{L} : (\mathcal{Z} \times \Theta) \to \mathbb{R}^1$ satisfying $\mathcal{L}(z; \theta) = \exp\{-L(f_\theta, z)\}/s_\theta$. As an abbreviation, let $\mathcal{L}(\mathbf{Z}; \theta)$ denote $\prod_{i=1}^n \mathcal{L}(\mathbf{Z}_{[i,\cdot]}; \theta)$. Additionally, let $\hat{\theta} := \arg\min_{\theta \in \Theta} \hat{e}_{\text{orig}}(f_\theta)$ be the empirical loss minimizer, and hence the maximum likelihood estimator of $\theta^\star$. If $\mathcal{L}$ is indeed the correct likelihood function, then $\theta^\star = (\gamma^\star, \psi^\star)$ corresponds to the true parameter vector. Further, if $\phi(f_\theta) = \phi(f_{(\gamma,\psi)}) = \gamma$ returns the parameter element of interest (γ), then the $(1-\delta)$-level

profile likelihood interval for $\phi(f_{\theta^\star}) = \gamma^\star$ is

$$\begin{aligned}
\text{PLI}(\delta) :=& \left\{ \gamma \ : \ \log \mathcal{L}(\mathbf{Z}; \hat{\theta}) - \log \mathcal{L}(\mathbf{Z}; \hat{\theta}_\gamma) \leq \frac{\chi_{1,1-\delta}}{2}, \ \text{ where } \ \hat{\theta}_\gamma = \underset{\{\theta \in \Theta \, : \, \phi(f_\theta) = \gamma\}}{\arg\max} \mathcal{L}(\mathbf{Z}; \theta) \right\} \\
=& \left\{ \gamma \ : \ \exists \hat{\theta}_\gamma \ \text{ satisfying } \ \phi\left(f_{\hat{\theta}_\gamma}\right) = \gamma \ \text{ and } \ \log \mathcal{L}(\mathbf{Z}; \hat{\theta}) - \log \mathcal{L}(\mathbf{Z}; \hat{\theta}_\gamma) \leq \frac{\chi_{1,1-\delta}}{2} \right\} \\
=& \left\{ \gamma \ : \ \exists \hat{\theta}_\gamma \ \text{ satisfying } \ \phi\left(f_{\hat{\theta}_\gamma}\right) = \gamma \ \text{ and } \ \hat{e}_{\text{orig}}\left(f_{\hat{\theta}_\gamma}\right) \leq \hat{e}_{\text{orig}}\left(f_{\hat{\theta}}\right) + \frac{\chi_{1,1-\delta}}{2n} \right\} \\
=& \left\{ \gamma \ : \ \exists f_{\hat{\theta}_\gamma} \ \text{ satisfying } \ \phi\left(f_{\hat{\theta}_\gamma}\right) = \gamma \ \text{ and } \ f_{\hat{\theta}_\gamma} \in \hat{\mathcal{R}}\left(\frac{\chi_{1,1-\delta}}{2n}, f_{\hat{\theta}}, \mathcal{F}\right) \right\} \qquad \text{(A.5)}
\end{aligned}$$

where $\chi_{1,1-\delta}$ is the $1-\delta$ percentile of a chi-square distribution with 1 degree of freedom. If $\text{PLI}(\alpha)$ is indeed a contiguous interval, then maximizing and minimizing $\phi(f_\theta)$ across models f_θ in the empirical Rashomon set in Eq A.5 yields the same interval.

A.7. Unbiased Estimates of CMR

We claim in Section 8.2 that both

$$\hat{e}_{\text{match}}(f) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \frac{1(\mathbf{X}_{2[j,\cdot]} = \mathbf{X}_{2[i,\cdot]})}{\mathbb{P}(X_2 = \mathbf{X}_{2[i,\cdot]})} \times L\{f, (\mathbf{y}_{[j]}, \mathbf{X}_{1[i,\cdot]}, \mathbf{X}_{2[j,\cdot]})\}.$$

and

$$\hat{e}_{\text{weight}}(f) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \frac{\mathbb{P}(X_1 = \mathbf{X}_{1[i,\cdot]} | X_2 = \mathbf{X}_{2[j,\cdot]})}{\mathbb{P}(X_1 = \mathbf{X}_{1[i,\cdot]})} \times L\{f, (\mathbf{y}_{[j]}, \mathbf{X}_{1[i,\cdot]}, \mathbf{X}_{2[j,\cdot]})\},$$

are unbiased for

$$e_{\text{cond}}(f) = \mathbb{E}_{X_2} \mathbb{E}\left[L\{f, (Y^{(b)}, X_1^{(a)}, X_2^{(b)})\} | X_2^{(a)} = X_2^{(b)}, X_2 \right].$$

To show that $\hat{e}_{\text{match}}(f)$ is unbiased, we first note that each summation term in $\hat{e}_{\text{match}}(f)$ has the same expectation. Following the notation in Section 3, let $Z^{(a)} = (Y^{(a)}, X_1^{(a)}, X_2^{(a)})$ and $Z^{(b)} = (Y^{(b)}, X_1^{(b)}, X_2^{(b)})$ be independent random variables following the same distribution as $Z = (Y, X_1, X_2)$. The expectation of $\hat{e}_{\text{match}}(f)$ is

$$\begin{aligned}
\mathbb{E}\hat{e}_{\text{match}}(f) =& \mathbb{E}\left[\frac{1(X_2^{(a)} = X_2^{(b)})}{p_{x_2}(X_2^{(a)})} \times L\{f, (Y^{(b)}, X_1^{(a)}, X_2^{(b)})\} \right] \\
=& \mathbb{E}_{X_2^{(a)}} \mathbb{E}\left[\frac{1(X_2^{(a)} = X_2^{(b)})}{p_{x_2}(X_2^{(a)})} \times L\{f, (Y^{(b)}, X_1^{(a)}, X_2^{(b)})\} | X_2^{(a)} \right] \\
=& \mathbb{E}_{X_2^{(a)}} \left\{ p_{x_2}(X_2^{(a)}) \mathbb{E}\left[\frac{1}{p_{x_2}(X_2^{(a)})} \times L\{f, (Y^{(b)}, X_1^{(a)}, X_2^{(b)})\} | X_2^{(a)} = X_2^{(b)}, X_2^{(a)} \right] + 0 \right\} \\
=& \mathbb{E}_{X_2^{(a)}} \mathbb{E}\left[L\{f, (Y^{(b)}, X_1^{(a)}, X_2^{(b)})\} | X_2^{(a)} = X_2^{(b)}, X_2^{(a)} \right] \\
=& e_{\text{cond}}(f).
\end{aligned}$$

To show that $\hat{e}_{\text{weight}}(f)$ is unbiased, we similarly note that each summation term in $\hat{e}_{\text{weight}}(f)$ has the same expectation. Without loss of generality, we show the result for discrete variables (Y, X_1, X_2). Let $\mathcal{Y}_{x_2}$ be the domain of Y conditional on the event that $X_2 = x_2$. The expectation of $\hat{e}_{\text{weight}}(f)$ is

$$\begin{aligned}
\mathbb{E}\hat{e}_{\text{weight}}(f) &= \sum_{x_2^{(b)} \in \mathcal{X}_2} \sum_{y^{(b)} \in \mathcal{Y}_{x_2^{(b)}}} \sum_{x_1^{(a)} \in \mathcal{X}_1} \left[L\{f, (y^{(b)}, x_1^{(a)}, x_2^{(b)})\} \left\{ \frac{\mathbb{P}(X_1 = x_1^{(a)} | X_2 = x_2^{(b)})}{\mathbb{P}(X_1 = x_1^{(a)})} \right\} \right. \\
&\qquad \left. \times \mathbb{P}(X_1 = x_1^{(a)}) \mathbb{P}(Y = y^{(b)}, X_2 = x_2^{(b)}) \right] \\
&= \sum_{x_2^{(b)} \in \mathcal{X}_2} \mathbb{P}(X_2 = x_2^{(b)}) \sum_{y^{(b)} \in \mathcal{Y}_{x_2^{(b)}}} \sum_{x_1^{(a)} \in \mathcal{X}_1} \Big[L\{f, (y^{(b)}, x_1^{(a)}, x_2^{(b)})\} \\
&\qquad \times \mathbb{P}(X_1 = x_1^{(a)} | X_2 = x_2^{(b)}) \mathbb{P}(Y = y^{(b)} | X_2 = x_2^{(b)}) \Big] \\
&= \mathbb{E}_{X_2^{(b)}} \mathbb{E} \left[\int L\{f, (Y^{(b)}, X_1^{(a)}, X_2^{(b)})\} | X_2^{(a)} = X_2^{(b)}, X_2^{(b)}) \right] \\
&= e_{\text{cond}}(f).
\end{aligned}$$

Appendix B. Proofs for Statistical Results

We present proofs for our statistical results in this section, and conclude by presenting proofs for our computational results in Appendix C.

B.1. Lemma Relating Empirical and Population Rashomon Sets

Throughout the remaining proofs, it will be useful to express the definition of population ϵ-Rashomon sets in terms of the expectation of a single loss function, rather than a comparison of two loss functions. To do this, we simply introduce the "standardized" loss function $\tilde{L}$, defined as

$$\tilde{L}(f, z) := L(f, z) - L(f_{\text{ref}}, z). \tag{B.1}$$

Above, recall from Section 2 that $L(f, z)$ denotes $L(f, (y, x_1, x_2))$ for $z = (y, x_1, x_2)$. Because we assume f_{ref} is prespecified and fixed, we omit notation for f_{ref} in the definition of $\tilde{L}$. We can now write

$$\begin{aligned}
\mathcal{R}(\epsilon) &= \{f_{\text{ref}}\} \cup \{f \in \mathcal{F} : \mathbb{E}L(f, Z) \leq \mathbb{E}L(f_{\text{ref}}, Z) + \epsilon\} \\
&= \{f_{\text{ref}}\} \cup \left\{ f \in \mathcal{F} : \mathbb{E}\tilde{L}(f, Z) \leq \epsilon \right\},
\end{aligned}$$

and, similarly,

$$\hat{\mathcal{R}}(\epsilon) = \{f_{\text{ref}}\} \cup \left\{ f \in \mathcal{F} : \hat{\mathbb{E}}\tilde{L}(f, Z) \leq \epsilon \right\}.$$

With this definition, the following lemma allows us to limit the probability that a given model $f_1 \in \mathcal{R}(\epsilon)$ is excluded from an empirical Rashomon set.

Lemma 23 *For $\epsilon \in \mathbb{R}$ and $\delta \in (0,1)$, let $\epsilon_1' := \epsilon + 2B_{ref}\sqrt{\frac{\log(\delta^{-1})}{2n}}$, and let $f_1 \in \mathcal{R}(\epsilon)$ denote a specific, possibly unknown prediction model. If f_1 satisfies Assumption 2, then*

$$\mathbb{P}\{f_1 \in \hat{\mathcal{R}}(\epsilon_1')\} \geq 1 - \delta.$$

Proof If f_{ref} and f_1 are the same function, then the result holds trivially. Otherwise, the proof follows from Hoeffding's inequality (Theorem 2 of Hoeffding, 1963). First, note that if f_1 satisfies Assumption 2, then $\tilde{L}(f_1)$ is bounded within an interval of length $2B_{\text{ref}}$. Applying this in line B.3, below, we see that

$$\begin{aligned}
\mathbb{P}\{f_1 \notin \hat{\mathcal{R}}(\epsilon_1')\} &= \mathbb{P}\left[\hat{\mathbb{E}}\tilde{L}(f_1, Z) > \epsilon_1'\right] && \text{from } f_1 \notin \{f_{\text{ref}}\} \\
&= \mathbb{P}\left[\hat{\mathbb{E}}\tilde{L}(f_1, Z) - \epsilon > 2B_{\text{ref}}\sqrt{\frac{\log(\delta^{-1})}{2n}}\right] && \text{from definition of } \epsilon_1' \\
& && \text{(B.2)} \\
&\leq \mathbb{P}\left[\hat{\mathbb{E}}\tilde{L}(f_1, Z) - \mathbb{E}\tilde{L}(f_1, Z) > 2B_{\text{ref}}\sqrt{\frac{\log(\delta^{-1})}{2n}}\right] && \text{from } \mathbb{E}\tilde{L}(f_1, Z) \leq \epsilon \\
&\leq \exp\left\{-\frac{2n}{(2B_{\text{ref}})^2}\left[2B_{\text{ref}}\sqrt{\frac{\log(\delta^{-1})}{2n}}\right]^2\right\} && \text{from Hoeffding's inequality} \\
& && \text{(B.3)} \\
&= \delta. && \text{(B.4)}
\end{aligned}$$

For the inequality used in Line B.3, see Theorem 2 of Hoeffding, 1963. ■

B.2. Lemma to Transform Between Bounds

The following lemma will help us translate from bounds for variables to bounds for differences and ratios of those variables. We will apply this lemma to transform from bounds on empirical losses to bounds on empirical model reliance, defined either in terms of a ratio or in terms of a difference.

Lemma 24 *Let $X, Z, \mu_X, \mu_Z, k_X, k_Z \in \mathbb{R}$ be constants satisfying $|Z - \mu_Z| \leq k_Z$ and $|X - \mu_X| \leq k_X$, then*

$$|(Z - X) - (\mu_Z - \mu_X)| \leq q_{\textit{difference}}(k_Z, k_X), \tag{B.5}$$

where $q_{\textit{difference}}$ is the function

$$q_{\textit{difference}}(k_Z, k_X) := k_Z + k_X. \tag{B.6}$$

Further, if there exists constants b_{orig} and B_{switch} such that $0 < b_{orig} \leq X, \mu_X$ and $Z, \mu_Z \leq B_{switch} < \infty$, then

$$\left| \frac{Z}{X} - \frac{\mu_Z}{\mu_X} \right| \leq q_{ratio}(k_Z, k_X), \tag{B.7}$$

where q_{ratio} is the function

$$q_{ratio}(k_Z, k_X) := \frac{B_{switch}}{b_{orig}} - \frac{B_{switch} - k_Z}{b_{orig} + k_X}. \tag{B.8}$$

Proof Showing Eq B.5,

$$\begin{aligned} |(Z - X) - (\mu_Z - \mu_X)| &\leq |Z - \mu_Z| + |\mu_X - X| \\ &\leq k_Z + k_X. \end{aligned}$$

Showing Eq B.7, let $A_Z = \max(Z, \mu_Z)$, $a_X = \min(X, \mu_X)$, $d_Z = |Z - \mu_Z|$, and $d_X = |X - \mu_X|$. This implies that $\max(X, \mu_X) = a_X + d_X$ and $\min(Z, \mu_Z) = A_Z - d_Z$. Thus, $\frac{Z}{X}$ and $\frac{\mu_Z}{\mu_X}$ are both bounded within the interval

$$\left[\frac{\min(Z, \mu_Z)}{\max(X, \mu_X)}, \frac{\max(Z, \mu_Z)}{\min(X, \mu_X)} \right] = \left[\frac{A_Z - d_Z}{a_X + d_X}, \frac{A_Z}{a_X} \right],$$

which implies

$$\left| \frac{Z}{X} - \frac{\mu_Z}{\mu_X} \right| \leq \frac{A_Z}{a_X} - \frac{A_Z - d_Z}{a_X + d_X}. \tag{B.9}$$

Taking partial derivatives of the right-hand side, we get

$$\begin{aligned} \frac{\partial}{\partial a_X} \left(\frac{A_Z}{a_X} - \frac{A_Z - d_Z}{a_X + d_X} \right) &= \frac{-A_Z}{a_X^2} + \frac{A_Z - d_Z}{(a_X + d_X)^2} \leq 0, \\ \frac{\partial}{\partial A_Z} \left(\frac{A_Z}{a_X} - \frac{A_Z - d_Z}{a_X + d_X} \right) &= \frac{1}{a_X} - \frac{1}{a_X + d_X} \geq 0, \\ \frac{\partial}{\partial d_X} \left(\frac{A_Z}{a_X} - \frac{A_Z - d_Z}{a_X + d_X} \right) &= \frac{A_Z - d_Z}{(a_X + d_X)^2} > 0, \\ \text{and } \frac{\partial}{\partial d_Z} \left(\frac{A_Z}{a_X} - \frac{A_Z - d_Z}{a_X + d_X} \right) &= \frac{1}{a_X + d_X} > 0. \end{aligned}$$

So the right-hand side of B.9 is maximized when d_Z, d_X, and A_Z are maximized, and when a_X is minimized. Thus, in the case where $|Z - \mu_Z| \leq k_Z$; $|X - \mu_X| \leq k_X$; $0 < b_{\text{orig}} \leq X, \mu_X$; and $Z, \mu_Z \leq B_{\text{switch}} < \infty$, we have

$$\left|\frac{Z}{X} - \frac{\mu_Z}{\mu_X}\right| \leq \frac{A_Z}{a_X} - \frac{A_Z - d_Z}{a_X + d_X}$$
$$\leq \frac{B_{\text{switch}}}{b_{\text{orig}}} - \frac{B_{\text{switch}} - k_Z}{b_{\text{orig}} + k_X}.$$

■

B.3. Proof of Theorem 4

Proof We proceed in 4 steps.

B.3.1. Step 1: Show that $\mathbb{P}\left[\widehat{MR}(f_{+,\epsilon}) \leq \widehat{MCR}_+(\epsilon_{\text{OUT}})\right] \geq 1 - \frac{\delta}{3}$.

Consider the event that

$$\widehat{MR}(f_{+,\epsilon}) \leq \widehat{MCR}_+(\epsilon_{\text{out}}). \tag{B.10}$$

Eq B.10 will always hold if $f_{+,\epsilon} \in \hat{\mathcal{R}}(\epsilon_{\text{out}})$, since $\widehat{MCR}_+(\epsilon_{\text{out}})$ upper bounds the empirical model reliance for models in $\hat{\mathcal{R}}(\epsilon_{\text{out}})$ by definition. Applying the above reasoning in Line B.11, below, we get

$$\mathbb{P}\left[\widehat{MR}(f_{+,\epsilon}) > \widehat{MCR}_+(\epsilon_{\text{out}})\right] \leq \mathbb{P}\left[f_{+,\epsilon} \notin \hat{\mathcal{R}}(\epsilon_{\text{out}})\right] \tag{B.11}$$
$$\leq \frac{\delta}{3} \qquad \text{from } \epsilon_{\text{out}} \text{ definition and Lemma 23.} \tag{B.12}$$

B.3.2. Step 2: Conditional on $\widehat{MR}(f_{+,\epsilon}) \leq \widehat{MCR}_+(\epsilon_{\text{OUT}})$, Upper Bound $MR(f_{+,\epsilon})$ by $\widehat{MCR}_+(\epsilon_{\text{OUT}})$ Added to an Error Term.

When Eq B.10 holds we have,

$$\begin{aligned}
\widehat{MR}(f_{+,\epsilon}) &\leq \widehat{MCR}_+(\epsilon_{\text{out}}) \\
\widehat{MR}(f_{+,\epsilon}) &\leq \widehat{MCR}_+(\epsilon_{\text{out}}) + \{MR(f_{+,\epsilon}) - MR(f_{+,\epsilon})\} \\
MR(f_{+,\epsilon}) &\leq \widehat{MCR}_+(\epsilon_{\text{out}}) + [MR(f_{+,\epsilon}) - \widehat{MR}(f_{+,\epsilon})].
\end{aligned} \tag{B.13}$$

B.3.3. Step 3: Probabilistically Bound the Error Term from Step 2.

Next we show that the bracketed term in Line B.13 is less than or equal to $\mathcal{Q}_{\text{out}}$ with high probability. For $k \in \mathbb{R}$, let $q_{\text{difference}}$ and q_{ratio} be defined as in Eqs B.6 and B.8. Let

$q : \mathbb{R} \to \mathbb{R}$ be the function such that $q(k) = q_{\text{ratio}}\left(k, \frac{k}{\sqrt{2}}\right)$. Then

$$\begin{aligned}
\mathcal{Q}_{\text{out}} &= \frac{B_{\text{switch}}}{b_{\text{orig}}} - \frac{B_{\text{switch}} - B_{\text{ind}}\sqrt{\frac{\log(6\delta^{-1})}{n}}}{b_{\text{orig}} + B_{\text{ind}}\sqrt{\frac{\log(6\delta^{-1})}{2n}}} \\
&= q_{\text{ratio}}\left(B_{\text{ind}}\sqrt{\frac{\log(6\delta^{-1})}{n}}, B_{\text{ind}}\sqrt{\frac{\log(6\delta^{-1})}{2n}}\right) \\
&= q\left(B_{\text{ind}}\sqrt{\frac{\log(6\delta^{-1})}{n}}\right).
\end{aligned}$$

Applying this relation below, we have

$$\begin{aligned}
&\mathbb{P}\left[MR(f_{+,\epsilon}) - \widehat{MR}(f_{+,\epsilon}) > \mathcal{Q}_{\text{out}}\right] && \text{(B.14)} \\
&\leq \mathbb{P}\left[\left|MR(f_{+,\epsilon}) - \widehat{MR}(f_{+,\epsilon})\right| > q\left(B_{\text{ind}}\sqrt{\frac{\log(6\delta^{-1})}{n}}\right)\right] \\
&\leq \mathbb{P}\left[\left\{|\hat{e}_{\text{orig}}(f_{+,\epsilon}) - e_{\text{orig}}(f_{+,\epsilon})| > B_{\text{ind}}\sqrt{\frac{\log(6\delta^{-1})}{2n}}\right\}\right. \\
&\qquad \left.\bigcup\left\{|\hat{e}_{\text{switch}}(f_{+,\epsilon}) - e_{\text{switch}}(f_{+,\epsilon})| > B_{\text{ind}}\sqrt{\frac{\log(6\delta^{-1})}{n}}\right\}\right] && \text{from Lemma 24} \\
&\leq \mathbb{P}\left[|\hat{e}_{\text{orig}}(f_{+,\epsilon}) - e_{\text{orig}}(f_{+,\epsilon})| > B_{\text{ind}}\sqrt{\frac{\log\left(6\delta^{-1}\right)}{2n}}\right] \\
&\qquad + \mathbb{P}\left[|\hat{e}_{\text{switch}}(f_{+,\epsilon}) - e_{\text{switch}}(f_{+,\epsilon})| > B_{\text{ind}}\sqrt{\frac{\log\left(6\delta^{-1}\right)}{n}}\right] && \text{from the Union bound} \\
&\leq 2\exp\left\{-\frac{2n}{(B_{\text{ind}}-0)^2}\left[B_{\text{ind}}\sqrt{\frac{\log\left(6\delta^{-1}\right)}{2n}}\right]^2\right\} \\
&\qquad + 2\exp\left\{-\frac{n}{(B_{\text{ind}}-0)^2}\left[B_{\text{ind}}\sqrt{\frac{\log\left(6\delta^{-1}\right)}{n}}\right]^2\right\} && \text{from Hoeffding's bound for U-statistics} \\
& && \text{(B.15)} \\
&= \frac{2\delta}{6} + \frac{2\delta}{6} = \frac{2\delta}{3}. && \text{(B.16)}
\end{aligned}$$

In Line B.15, above, recall that $\hat{e}_{\text{orig}}(f_{+,\epsilon})$ and $\hat{e}_{\text{switch}}(f_{+,\epsilon})$ are both U-statistics. Note that $\mathbb{E}\left[\hat{e}_{\text{switch}}(f_{+,\epsilon})\right] = e_{\text{switch}}(f_{+,\epsilon})$ because $\hat{e}_{\text{switch}}(f_{+,\epsilon})$ is an average of terms, and each term has expectation equal to $e_{\text{switch}}(f_{+,\epsilon})$. For the same reason, $\mathbb{E}\left[\hat{e}_{\text{orig}}(f_{+,\epsilon})\right] = e_{\text{orig}}(f_{+,\epsilon})$. This allows us to apply Eq 5.7 of Hoeffding, 1963 (see also Eq 1 on page 201 of Serfling, 1980, in Theorem A) to obtain Line B.15.

Alternatively, if we instead define model reliance as $MR_{\text{difference}}(f) = e_{\text{switch}}(f) - e_{\text{orig}}(f)$ (see Appendix A.5), define empirical model reliance as $\widehat{MR}_{\text{difference}}(f) := \hat{e}_{\text{switch}}(f) - \hat{e}_{\text{orig}}(f)$, and define

$$\mathcal{Q}_{\text{out,difference}} := \left(1 + \frac{1}{\sqrt{2}}\right) B_{\text{ind}} \sqrt{\frac{\log(6\delta^{-1})}{n}} = q_{\text{difference}}\left(B_{\text{ind}}\sqrt{\frac{\log(6\delta^{-1})}{n}}, B_{\text{ind}}\sqrt{\frac{\log(6\delta^{-1})}{2n}}\right),$$

then the same proof holds without Assumption 3 if we replace MR, $\widehat{MR}$, $\mathcal{Q}_{\text{out}}$ respectively with $MR_{\text{difference}}$, $\widehat{MR}_{\text{difference}}$, $\mathcal{Q}_{\text{out,difference}}$, and redefine $q : \mathbb{R} \to \mathbb{R}$ as the function $q(k) = q_{\text{difference}}\left(k, \frac{k}{\sqrt{2}}\right)$.

Eqs B.14-B.16 also hold if we replace $\hat{e}_{\text{switch}}$ throughout with $\hat{e}_{\text{divide}}$, including in Assumption 3, since the same bound can be used for both $\hat{e}_{\text{switch}}$ and $\hat{e}_{\text{divide}}$ (Eq 5.7 of Hoeffding, 1963; see also Theorem A on page 201 of Serfling, 1980).

B.3.4. Step 4: Combine Results to Show Eq 4.2

Finally, we connect the above results to show Eq 4.2. We know from Eq B.12 that Eq B.10 holds with high probability. Eq B.10 implies Eq B.13, which bounds $MCR_+(\epsilon) = MR(f_{+,\epsilon})$ up to a bracketed residual term. We also know from Eq B.16 that, with high probability, the residual term in Eq B.13 is less than $\mathcal{Q}_{\text{out}} = q\left(B_{\text{ind}}\sqrt{\frac{\log(6\delta^{-1})}{n}}\right)$. Putting this together, we can show Eq 4.2:

$$\begin{aligned}
&\mathbb{P}\left(MCR_+(\epsilon) > \widehat{MCR}_+(\epsilon_{\text{out}}) + \mathcal{Q}_{\text{out}}\right) \\
&\quad = \mathbb{P}\left(MR(f_{+,\epsilon}) > \widehat{MCR}_+(\epsilon_{\text{out}}) + \mathcal{Q}_{\text{out}}\right) \\
&\quad \leq \mathbb{P}\left[\left(\widehat{MR}(f_{+,\epsilon}) > \widehat{MCR}_+(\epsilon_{\text{out}})\right) \bigcup \left(MR(f_{+,\epsilon}) - \widehat{MR}(f_{+,\epsilon}) > \mathcal{Q}_{\text{out}}\right)\right] \quad \text{from Step 2} \\
&\quad \leq \mathbb{P}\left[\widehat{MR}(f_{+,\epsilon}) > \widehat{MCR}_+(\epsilon_{\text{out}})\right] + \mathbb{P}\left[MR(f_{+,\epsilon}) - \widehat{MR}(f_{+,\epsilon}) > \mathcal{Q}_{\text{out}}\right] \\
&\quad \leq \frac{\delta}{3} + \frac{2\delta}{3} = \delta. \qquad \text{from Steps 1 \& 3}
\end{aligned} \tag{B.17}$$

This completes the proof for Eq 4.2. For Eq 4.3 we can use the same approach, shown below for completeness. Analogous to Eq B.12, we have

$$\mathbb{P}\left[\widehat{MR}(f_{-,\epsilon}) < \widehat{MCR}_-(\epsilon_{\text{out}})\right] \leq \frac{\delta}{3}.$$

Analogous to Eq B.13, when $\widehat{MR}(f_{-,\epsilon}) \geq \widehat{MR}(\hat{f}_{-,\epsilon_{\text{out}}})$ we have

$$\begin{aligned}
\widehat{MR}(f_{-,\epsilon}) &\geq \widehat{MCR}_-(\epsilon_{\text{out}}) \\
\widehat{MR}(f_{-,\epsilon}) &\geq \widehat{MCR}_-(\epsilon_{\text{out}}) + \{MR(f_{-,\epsilon}) - MR(f_{-,\epsilon})\} \\
MR(f_{-,\epsilon}) &\geq \widehat{MCR}_-(\epsilon_{\text{out}}) - \left[\widehat{MR}(f_{-,\epsilon}) - MR(f_{-,\epsilon})\right].
\end{aligned}$$

Analogous to Eq B.16, we have

$$\mathbb{P}\left[\widehat{MR}(f_{-,\epsilon}) - MR(f_{-,\epsilon}) > q\left(B_{\text{ind}}\sqrt{\frac{\log(6\delta^{-1})}{n}}\right)\right] \leq \frac{2\delta}{3}. \tag{B.18}$$

Finally, analogous to Eq B.17, we have

$$\begin{aligned}
&\mathbb{P}\left(MCR_-(\epsilon) < \widehat{MCR}_-(\epsilon_{\text{out}}) - \mathcal{Q}_{\text{out}}\right) \\
&\quad\leq \mathbb{P}\left[\left(\widehat{MR}(f_{-,\epsilon}) < \widehat{MCR}_-(\epsilon_{\text{out}})\right) \bigcup \left(\widehat{MR}(f_{-,\epsilon}) - MR(f_{-,\epsilon}) > q\left(B_{\text{ind}}\sqrt{\frac{\log(6\delta^{-1})}{n}}\right)\right)\right] \qquad \text{(B.19)} \\
&\quad\leq \mathbb{P}\left[\widehat{MR}(f_{-,\epsilon}) < \widehat{MCR}_-(\epsilon_{\text{out}})\right] + \mathbb{P}\left[\widehat{MR}(f_{-,\epsilon}) - MR(f_{-,\epsilon}) > q\left(B_{\text{ind}}\sqrt{\frac{\log(6\delta^{-1})}{n}}\right)\right] \\
&\quad\leq \frac{\delta}{3} + \frac{2\delta}{3} = \delta.
\end{aligned}$$

Again, the same proof holds without Assumption 3 if we replace MR, $\widehat{MR}$, $\mathcal{Q}_{\text{out}}$ respectively with $MR_{\text{difference}}$, $\widehat{MR}_{\text{difference}}$, $\mathcal{Q}_{\text{out,difference}}$, and redefine q as the function satisfying $q(k) = q_{\text{difference}}\left(k, \frac{k}{\sqrt{2}}\right)$ in Eqs B.18 & B.19. ■

B.4. Proof of Corollary 22

Proof By definition, $MR(f_{-,\epsilon_{\text{best}}}) \leq MR(f^\star) \leq MR(f_{+,\epsilon_{\text{best}}})$. Applying this relation in Line B.20, below, we see

$$\begin{aligned}
&\mathbb{P}\left(MR(f^\star) \in \left[\widehat{MCR}_-\left(\epsilon_{\text{best}}\right) - \mathcal{Q}_{\text{best}}, \quad \widehat{MCR}_+\left(\epsilon_{\text{best}}\right) + \mathcal{Q}_{\text{best}}\right]\right) \\
&\quad= 1 - \mathbb{P}\left\{MR(f^\star) < \widehat{MCR}_-\left(\epsilon_{\text{best}}\right) - \mathcal{Q}_{\text{best}} \bigcup MR(f^\star) > \widehat{MCR}_+\left(\epsilon_{\text{best}}\right) + \mathcal{Q}_{\text{best}}\right\} \\
&\quad\geq 1 - \mathbb{P}\left\{MR(f^\star) < \widehat{MCR}_-\left(\epsilon_{\text{best}}\right) - \mathcal{Q}_{\text{best}}\right\} - \mathbb{P}\left\{MR(f^\star) > \widehat{MCR}_+\left(\epsilon_{\text{best}}\right) + \mathcal{Q}_{\text{best}}\right\} \\
&\quad\geq 1 - \mathbb{P}\left\{MR(f_{-,\epsilon}) < \widehat{MCR}_-\left(\epsilon_{\text{best}}\right) - \mathcal{Q}_{\text{best}}\right\} - \mathbb{P}\left\{MR(f_{+,\epsilon}) > \widehat{MCR}_+\left(\epsilon_{\text{best}}\right) + \mathcal{Q}_{\text{best}}\right\} \qquad \text{(B.20)} \\
&\quad\geq 1 - \frac{\delta}{2} - \frac{\delta}{2} \qquad \text{from Theorem 4.} \qquad \text{(B.21)}
\end{aligned}$$

To apply Theorem 4 in Line B.21, above, we note that $\mathcal{Q}_{\text{best}}$ and ϵ_{best} are equivalent to the definitions of $\mathcal{Q}_{\text{out}}$ and ϵ_{out} in Theorem 4, but with δ replaced by $\frac{\delta}{2}$.

Alternatively, if we define model reliance as $MR_{\text{difference}}(f) = e_{\text{switch}}(f) - e_{\text{orig}}(f)$ (see Appendix A.5), and define empirical model reliance as $\widehat{MR}_{\text{difference}}(f) = \hat{e}_{\text{switch}}(f) - \hat{e}_{\text{orig}}(f)$, then let

$$\mathcal{Q}_{\text{best,difference}} := \left(1 + \frac{1}{\sqrt{2}}\right) B_{\text{ind}}\sqrt{\frac{\log(12\delta^{-1})}{n}}.$$

The term $\mathcal{Q}_{\text{best,difference}}$ is equivalent to $\mathcal{Q}_{\text{out,difference}}$ but with δ replaced with $\frac{\delta}{2}$. Under this difference-based definition of model reliance, Theorem 4 holds without Assumption 3 if we replace $\mathcal{Q}_{\text{out}}$ with $\mathcal{Q}_{\text{out,difference}}$ (see Section B.3), and so we can apply this altered version of Theorem 4 in Line B.21. Thus, Theorem 22 also holds without Assumption 3 if we replace MR, $\widehat{MR}$, and $\mathcal{Q}_{\text{best}}$ respectively with $MR_{\text{difference}}$, $\widehat{MR}_{\text{difference}}$, and $\mathcal{Q}_{\text{best,difference}}$. ■

B.5. Proof of Theorems 5 & 6

We begin by proving Theorem 5, along with related results. We then apply these results to show Theorem 6.

B.5.1. Proof of Theorem 5, and Other Limits on Estimation Error, Based on Covering Number

The following theorem uses the covering number based on r-margin-expectation-covers to jointly bound empirical losses for any function $f \in \mathcal{F}$. Theorem 5 in the main text follows directly from Eq B.25, below.

Theorem 25 *If Assumptions 1, 2 and 3 hold for all $f \in \mathcal{F}$, then for any $r > 0$*

$$\mathbb{P}_D\left[\sup_{f\in\mathcal{F}} |\hat{e}_{orig}(f) - e_{orig}(f)| > B_{ind}\sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F}, r))}{2n}} + 2r\right] \leq \delta, \tag{B.22}$$

$$\mathbb{P}_D\left[\sup_{f\in\mathcal{F}} \left|\hat{\mathbb{E}}\tilde{L}(f, Z) - \mathbb{E}\tilde{L}(f, Z)\right| > 2B_{ref}\sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F}, r))}{2n}} + 2r\right] \leq \delta, \tag{B.23}$$

$$\mathbb{P}_D\left[\sup_{f\in\mathcal{F}} |\hat{e}_{switch}(f) - e_{switch}(f)| > B_{ind}\sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F}, r))}{n}} + 2r\right] \leq \delta, \tag{B.24}$$

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}} \left|\frac{\hat{e}_{orig}(f)}{\hat{e}_{switch}(f)} - \frac{e_{orig}(f)}{e_{switch}(f)}\right| > \mathcal{Q}_4\right] \leq \delta, \tag{B.25}$$

$$\mathbb{P}_D\left[\sup_{f\in\mathcal{F}} |\{\hat{e}_{switch}(f) - \hat{e}_{orig}(f)\} - \{e_{switch}(f) - e_{orig}(f)\}| > \mathcal{Q}_{4,difference}\right] \leq \delta, \tag{B.26}$$

where

$$\mathcal{Q}_4 := q_{ratio}\left(B_{ind}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}(\mathcal{F}, r\sqrt{2}))}{n}} + 2r\sqrt{2}, B_{ind}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}(\mathcal{F}, r))}{2n}} + 2r\right), \tag{B.27}$$

$$\mathcal{Q}_{4,difference} := q_{difference}\left(B_{ind}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}(\mathcal{F}, r\sqrt{2}))}{n}} + 2r\sqrt{2}, B_{ind}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}(\mathcal{F}, r))}{2n}} + 2r\right), \tag{B.28}$$

and q_{ratio} and $q_{difference}$ are defined as in Lemma 24. For Eq B.26, the result is unaffected if we remove Assumption 3.

B.5.2. Proof of Eq B.22

Proof Let $\mathcal{G}_r$ be a r-margin-expectation-cover for $\mathcal{F}$ of size $\mathcal{N}(\mathcal{F}, r)$. Let D_p denote the population distribution, let D_s be the sample distribution, and let $D^\star$ be the uniform mixture of D_p and D_s, i.e., for any $z \in \mathcal{Z}$,

$$\mathbb{P}_{D^\star}(Z \leq z) = \frac{1}{2}\mathbb{P}_{D_p}(Z \leq z) + \frac{1}{2}\mathbb{P}_{D_s}(Z \leq z). \tag{B.29}$$

Unless otherwise stated, we take expectations and probabilities with respect to D_p. Since $\mathcal{G}_r$ is a r-margin-expectation-cover, we know that for any $f \in \mathcal{F}$ we can find a function $g \in \mathcal{G}_r$ such that $\mathbb{E}_{D^\star}|L(g,Z) - L(f,Z)| = \mathbb{E}_{D^\star}\left|\tilde{L}(g,Z) - \tilde{L}(f,Z)\right| \leq r$, and

$$\begin{aligned}
\left|\hat{\mathbb{E}}L(f,Z) - \mathbb{E}L(f,Z)\right| &= \left|\hat{\mathbb{E}}L(f,Z) - \mathbb{E}L(f,Z) + \left\{\hat{\mathbb{E}}L(g,Z) - \hat{\mathbb{E}}L(g,Z)\right\} + \{\mathbb{E}L(g,Z) - \mathbb{E}L(g,Z)\}\right| \qquad \text{(B.30)}\\
&\leq \left|\hat{\mathbb{E}}L(g,Z) - \mathbb{E}L(g,Z)\right| + |\mathbb{E}L(g,Z) - \mathbb{E}L(f,Z)| + \left|\hat{\mathbb{E}}L(f,Z) - \hat{\mathbb{E}}L(g,Z)\right| \\
&\leq \left|\hat{\mathbb{E}}L(g,Z) - \mathbb{E}L(g,Z)\right| + \mathbb{E}_{D_p}|L(g,Z) - L(f,Z)| + \mathbb{E}_{D_s}|L(f,Z) - L(g,Z)| \\
&= \left|\hat{\mathbb{E}}L(g,Z) - \mathbb{E}L(g,Z)\right| + 2\mathbb{E}_{D^\star}|L(g,Z) - L(f,Z)| \\
&\leq \left|\hat{\mathbb{E}}L(g,Z) - \mathbb{E}L(g,Z)\right| + 2r.
\end{aligned}$$

Applying the above relation in Line B.31 below, we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\hat{\mathbb{E}}L(f,Z) - \mathbb{E}L(f,Z)\right| > B_{\text{ind}}\sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F},r))}{2n}} + 2r\right) \\
&\quad= \mathbb{P}\left(\exists f \in \mathcal{F} :\ \left|\hat{\mathbb{E}}L(f,Z) - \mathbb{E}L(f,Z)\right| > B_{\text{ind}}\sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F},r))}{2n}} + 2r\right) \\
&\quad\leq \mathbb{P}\left(\exists g \in \mathcal{G}_r :\ \left|\hat{\mathbb{E}}L(g,Z) - \mathbb{E}L(g,Z)\right| + 2r > B_{\text{ind}}\sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F},r))}{2n}} + 2r\right) \qquad \text{(B.31)} \\
&\quad= \mathbb{P}\left(\bigcup_{g\in\mathcal{G}_r}\left|\hat{\mathbb{E}}L(g,Z) - \mathbb{E}L(g,Z)\right| > B_{\text{ind}}\sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F},r))}{2n}}\right) \\
&\quad\leq \sum_{g\in\mathcal{G}_r}\mathbb{P}\left(\left|\hat{\mathbb{E}}L(g,Z) - \mathbb{E}L(g,Z)\right| > B_{\text{ind}}\sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F},r))}{2n}}\right) \qquad \text{from the Union bound} \\
&\quad\leq \mathcal{N}(\mathcal{F},r)\,2\exp\left[-\frac{2n}{(B_{\text{ind}})^2}\left\{B_{\text{ind}}\sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F},r))}{2n}}\right\}^2\right] \qquad \text{from Hoeffding's inequality} \\
&\qquad\qquad \text{(B.32)} \\
&\quad= \delta. \qquad \text{(B.33)}
\end{aligned}$$

To apply Hoeffding's inequality (Theorem 2 of Hoeffding, 1963) in Line B.32, above, we use the fact that $L(g,Z)$ is bounded within an interval of length B_{ind}. ■

B.5.3. Proof of Eq B.23

Proof The proof for Eq B.23 is nearly identical to the proof for Eq B.22. Simply replacing L and B_{ind} respectively with $\tilde{L}$ and $(2B_{\text{ref}})$ in Eqs B.30-B.33 yields a valid proof for Eq B.23. ■

B.5.4. Proof of Eq B.24

Proof Let F_D denote the cumulative distribution function for a distribution D. Let $\tilde{D}_p$ be the distribution such that

$$F_{\tilde{D}_p}(Y = y, X_1 = x_1, X_2 = x_2) = F_{D_p}(Y = y, X_2 = x_2) F_{D_p}(X_1 = x_1).$$

Let $\tilde{D}_s$ be the distribution satisfying

$$\mathbb{P}_{\tilde{D}_s}(Y = y, X_1 = x_1, X_2 = x_2) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \mathbf{1}(\mathbf{y}_{[j]} = y, \mathbf{X}_{1[i,\cdot]} = x_1, \mathbf{X}_{2[j,\cdot]} = x_2).$$

Let $\tilde{D}^\star$ be the uniform mixture of $\tilde{D}_p$ and $\tilde{D}_s$, as in Eq B.29. Replacing e_{orig}, $\hat{e}_{\text{orig}}$, D_p, D_s, and $D^\star$ respectively with e_{switch}, $\hat{e}_{\text{switch}}$, $\tilde{D}_p$, $\tilde{D}_s$, and $\tilde{D}^\star$, we can follow the same steps as in the proof for Eq B.22. For any $f \in \mathcal{F}$, we know that there exists a function $g \in \mathcal{G}_r$ satisfying $\mathbb{E}_{\tilde{D}^\star} |L(g, Z) - L(f, Z)| \leq r$, which implies

$$\begin{aligned}
|\hat{e}_{\text{switch}}(f) - e_{\text{switch}}(f)| &= |\hat{e}_{\text{switch}}(f) - e_{\text{switch}}(f) + \{\hat{e}_{\text{switch}}(g) - \hat{e}_{\text{switch}}(g)\} + \{e_{\text{switch}}(g) - e_{\text{switch}}(g)\}| \\
&\leq |\hat{e}_{\text{switch}}(g) - e_{\text{switch}}(g)| + \mathbb{E}_{\tilde{D}_p} |L(g, Z) - L(f, Z)| + \mathbb{E}_{\tilde{D}_s} |L(f, Z) - L(g, Z)| \\
&= |\hat{e}_{\text{switch}}(g) - e_{\text{switch}}(g)| + 2\mathbb{E}_{\tilde{D}^\star} |L(g, Z) - L(f, Z)| \\
&\leq |\hat{e}_{\text{switch}}(g) - e_{\text{switch}}(g)| + 2r.
\end{aligned}$$

As a result,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f \in \mathcal{F}} |\hat{e}_{\text{switch}}(f) - e_{\text{switch}}(f)| > B_{\text{ind}} \sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F}, r))}{n}} + 2r \right) \\
&\quad \leq \mathbb{P}\left(\exists g \in \mathcal{G}_r : \; |\hat{e}_{\text{switch}}(g) - e_{\text{switch}}(g)| + 2r > B_{\text{ind}} \sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F}, r))}{n}} + 2r \right) \\
&\quad \leq \sum_{g \in \mathcal{G}_r} \mathbb{P}\left(|\hat{e}_{\text{switch}}(g) - e_{\text{switch}}(g)| > B_{\text{ind}} \sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F}, r))}{n}} \right) \\
&\quad \leq \mathcal{N}(\mathcal{F}, r)\, 2 \exp\left[-\frac{n}{(B_{\text{ind}} - 0)^2} \left\{ B_{\text{ind}} \sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F}, r))}{n}} \right\}^2 \right] \qquad \text{(B.34)} \\
&\quad = \delta.
\end{aligned}$$

In Line B.34, above, we apply Eq 5.7 of Hoeffding, 1963 (see also Eq 1 on page 201 of Serfling, 1980, in Theorem A), in the same way as in Eq B.15. ■

B.5.5. Proof for Eq B.25

Proof We apply Lemma 24 and Eq B.27 in Line B.36, below, to obtain

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}}\left|\frac{\hat{e}_{\text{orig}}(f)}{\hat{e}_{\text{switch}}(f)}-\frac{e_{\text{orig}}(f)}{e_{\text{switch}}(f)}\right|>\mathcal{Q}_4\right] \tag{B.35}$$

$$=\mathbb{P}\left(\exists f\in\mathcal{F}:\ \left|\frac{\hat{e}_{\text{orig}}(f)}{\hat{e}_{\text{switch}}(f)}-\frac{e_{\text{orig}}(f)}{e_{\text{switch}}(f)}\right|>\mathcal{Q}_4\right)$$

$$\leq\mathbb{P}\left(\left\{\exists f\in\mathcal{F}:\ |\hat{e}_{\text{orig}}(f)-e_{\text{orig}}(f)|>B_{\text{ind}}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}\left(\mathcal{F},r\right))}{2n}}+2r\right\}\right. \tag{B.36}$$

$$\left.\bigcup\left\{\exists f\in\mathcal{F}:\ |\hat{e}_{\text{switch}}(f)-e_{\text{switch}}(f)|>B_{\text{ind}}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}\left(\mathcal{F},r\sqrt{2}\right))}{n}}+2r\sqrt{2}\right\}\right)$$

$$=\mathbb{P}\left(\sup_{f\in\mathcal{F}}|\hat{e}_{\text{orig}}(f)-e_{\text{orig}}(f)|>B_{\text{ind}}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}\left(\mathcal{F},r\right))}{2n}}+2r\right)$$

$$+\mathbb{P}\left(\sup_{f\in\mathcal{F}}|\hat{e}_{\text{switch}}(f)-e_{\text{switch}}(f)|>B_{\text{ind}}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}\left(\mathcal{F},r\sqrt{2}\right))}{n}}+2r\sqrt{2}\right)$$

$$\leq\frac{\delta}{2}+\frac{\delta}{2}. \qquad \text{from Eqs B.22 and B.24} \tag{B.37}$$

■

B.5.6. Proof for Eq B.26

Proof Finally, to show B.26, we apply the same steps as in Eqs B.35 through B.37. We apply Eq B.28 & Lemma 24 to obtain

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}}|\{\hat{e}_{\text{switch}}(f)-\hat{e}_{\text{orig}}(f)\}-\{e_{\text{switch}}(f)-e_{\text{orig}}(f)\}|>\mathcal{Q}_{4,\text{difference}}\right]$$

$$\leq\mathbb{P}\left(\left\{\exists f\in\mathcal{F}:\ |\hat{e}_{\text{orig}}(f)-e_{\text{orig}}(f)|>B_{\text{ind}}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}\left(\mathcal{F},r\right))}{2n}}+2r\right\}\right.$$

$$\left.\bigcup\left\{\exists f\in\mathcal{F}:\ |\hat{e}_{\text{switch}}(f)-e_{\text{switch}}(f)|>B_{\text{ind}}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}\left(\mathcal{F},r\sqrt{2}\right))}{n}}+2r\sqrt{2}\right\}\right)$$

$$\leq\frac{\delta}{2}+\frac{\delta}{2}.$$

■

B.5.7. Implementing Theorem 25 to Show Theorem 6

Proof Consider the event that

$$\exists \hat{f}_{+,\epsilon_{\text{in}}} \in \underset{f \in \hat{\mathcal{R}}(\epsilon_{\text{in}})}{\arg\max} \widehat{MR}(f) \text{ such that } MCR_+(\epsilon) < MR\left(\hat{f}_{+,\epsilon_{\text{in}}}\right). \tag{B.38}$$

A brief outline of our proof for Eq 4.6 is as follows. We expect Eq B.38 to be unlikely due to the fact that $\epsilon_{\text{in}} < \epsilon$. If Eq B.38 does not hold, then the only way that $MCR_+(\epsilon) < \widehat{MCR}_+(\epsilon_{\text{in}}) - \mathcal{Q}_{\text{in}}$ holds is if there exists $\hat{f}_{+,\epsilon_{\text{in}}} \in \arg\max_{f \in \hat{\mathcal{R}}(\epsilon_{\text{in}})} \widehat{MR}(f)$ which has an empirical MR that differs from its population-level MR by at least $\mathcal{Q}_{\text{in}}$.

To show that Eq B.38 is unlikely, we apply Theorem 25:

$$\begin{aligned}
&\mathbb{P}\left(\exists \hat{f}_{+,\epsilon_{\text{in}}} \in \underset{f \in \hat{\mathcal{R}}(\epsilon_{\text{in}})}{\arg\max} \widehat{MR}(f) : MCR_+(\epsilon) < MR\left(\hat{f}_{+,\epsilon_{\text{in}}}\right)\right) \\
&\leq \mathbb{P}\left(\exists f \in \hat{\mathcal{R}}(\epsilon_{\text{in}}) : MCR_+(\epsilon) < MR(f)\right) \\
&= \mathbb{P}\left(\exists f \in \hat{\mathcal{R}}(\epsilon_{\text{in}}) \backslash f_{\text{ref}} : MCR_+(\epsilon) < MR(f)\right) && \text{by } MCR_+(\epsilon) \geq MR(f_{\text{ref}}) \\
&\leq \mathbb{P}\left(\exists f \in \hat{\mathcal{R}}(\epsilon_{\text{in}}) \backslash f_{\text{ref}} : \mathbb{E}\tilde{L}(f, Z) > \epsilon\right) && \text{by } MCR_+(\epsilon) \text{ Def} \\
&= \mathbb{P}\left(\exists f \in \mathcal{F}, \mathbb{E}\tilde{L}(f, Z) > \epsilon : \hat{\mathbb{E}}\tilde{L}(f, Z) \leq \epsilon_{\text{in}}\right) && \text{by } \hat{\mathcal{R}}(\epsilon) \text{ Def} \\
&= \mathbb{P}\Bigg(\exists f \in \mathcal{F}, \mathbb{E}\tilde{L}(f, Z) > \epsilon : \\
&\qquad \hat{\mathbb{E}}\tilde{L}(f, Z) - \epsilon \leq -2B_{\text{ref}}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}(\mathcal{F}, r))}{2n}} - 2r\Bigg) && \text{by } \epsilon_{\text{in}} \text{ Def}
\end{aligned} \tag{B.39}$$

$$\begin{aligned}
&\leq \mathbb{P}\Bigg(\exists f \in \mathcal{F}, \mathbb{E}\tilde{L}(f, Z) > \epsilon : \\
&\qquad \hat{\mathbb{E}}\tilde{L}(f, Z) - \mathbb{E}\tilde{L}(f, Z) \leq -2B_{\text{ref}}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}(\mathcal{F}, r))}{2n}} - 2r\Bigg) \\
&\leq \mathbb{P}\left(\sup_{f \in \mathcal{F}} |\hat{\mathbb{E}}\tilde{L}(f, Z) - \mathbb{E}\tilde{L}(f, Z)| \geq 2B_{\text{ref}}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}(\mathcal{F}, r))}{2n}} + 2r\right) \\
&= \frac{\delta}{2} && \text{by Thm 25.}
\end{aligned} \tag{B.40}$$

If Eq B.38 does not hold, we have

$$\begin{aligned}
MCR_+(\epsilon) &\geq MR\left(\hat{f}_{+,\epsilon_{\text{in}}}\right) && \text{for all } \hat{f}_{+,\epsilon_{\text{in}}} \in \underset{f\in\hat{\mathcal{R}}(\epsilon_{\text{in}})}{\arg\max}\, \widehat{MR}(f) \\
&= \widehat{MR}\left(\hat{f}_{+,\epsilon_{\text{in}}}\right) - \left\{\widehat{MR}\left(\hat{f}_{+,\epsilon_{\text{in}}}\right) - MR\left(\hat{f}_{+,\epsilon_{\text{in}}}\right)\right\} && \text{for all } \hat{f}_{+,\epsilon_{\text{in}}} \in \underset{f\in\hat{\mathcal{R}}(\epsilon_{\text{in}})}{\arg\max}\, \widehat{MR}(f) \\
&= \widehat{MCR}_+\left(\epsilon_{\text{in}}\right) - \left\{\widehat{MR}\left(\hat{f}_{+,\epsilon_{\text{in}}}\right) - MR\left(\hat{f}_{+,\epsilon_{\text{in}}}\right)\right\} && \text{for all } \hat{f}_{+,\epsilon_{\text{in}}} \in \underset{f\in\hat{\mathcal{R}}(\epsilon_{\text{in}})}{\arg\max}\, \widehat{MR}(f) \\
&\geq \widehat{MCR}_+\left(\epsilon_{\text{in}}\right) - \sup_{f\in\mathcal{F}} |\widehat{MR}(f) - MR(f)|. && \text{(B.41)}
\end{aligned}$$

Let q_{ratio} and $q_{\text{difference}}$ be defined as in Lemma 24. Then

$$\begin{aligned}
\mathcal{Q}_{\text{in}} &= \frac{B_{\text{switch}}}{b_{\text{orig}}} - \frac{B_{\text{switch}} - \left\{B_{\text{ind}}\sqrt{\frac{\log(8\delta^{-1}\mathcal{N}\left(\mathcal{F},r\sqrt{2}\right))}{n}} + 2r\sqrt{2}\right\}}{b_{\text{orig}} + \left\{B_{\text{ind}}\sqrt{\frac{\log(8\delta^{-1}\mathcal{N}(\mathcal{F},r))}{2n}} + 2r\right\}} \\
&= q_{\text{ratio}}\left(B_{\text{ind}}\sqrt{\frac{\log(8\delta^{-1}\mathcal{N}\left(\mathcal{F},r\sqrt{2}\right))}{n}} + 2r\sqrt{2}, B_{\text{ind}}\sqrt{\frac{\log(8\delta^{-1}\mathcal{N}\left(\mathcal{F},r\right))}{2n}} + 2r\right) && \text{(B.42)}
\end{aligned}$$

Theorem 25 implies that the sup term in Eq B.41 is less than $\mathcal{Q}_{\text{in}}$ with probability at least $1 - \frac{\delta}{2}$. Now, examining the left-hand side of Eq 4.6, we see

$$\begin{aligned}
&\mathbb{P}\left(MCR_+(\epsilon) < \widehat{MCR}_+\left(\epsilon_{\text{in}}\right) - \mathcal{Q}_{\text{in}}\right) \\
&\quad\leq \mathbb{P}\left[\left\{\exists \hat{f}_{+,\epsilon_{\text{in}}} \in \underset{f\in\hat{\mathcal{R}}(\epsilon_{\text{in}})}{\arg\max}\, \widehat{MR}(f) :\ MCR_+(\epsilon) < MR\left(\hat{f}_{+,\epsilon_{\text{in}}}\right)\right\}\right. \\
&\qquad\qquad \left.\bigcup \left\{\sup_{f\in\mathcal{F}} |\widehat{MR}(f) - MR(f)| > \mathcal{Q}_{\text{in}}\right\}\right] && \text{from Eq B.41} \\
&\quad\leq \mathbb{P}\left[\exists \hat{f}_{+,\epsilon_{\text{in}}} \in \underset{f\in\hat{\mathcal{R}}(\epsilon_{\text{in}})}{\arg\max}\, \widehat{MR}(f) :\ MCR_+(\epsilon) < MR\left(\hat{f}_{+,\epsilon_{\text{in}}}\right)\right] \\
&\qquad\qquad + \mathbb{P}\left[\sup_{f\in\mathcal{F}} |\widehat{MR}(f) - MR(f)| > \mathcal{Q}_{\text{in}}\right] && \text{from the Union bound} \\
&\quad= \frac{\delta}{2} + \frac{\delta}{2} && \text{from Eq B.40, Eq B.42, \& Theorem 25.} \quad \text{(B.43)}
\end{aligned}$$

This completes the proof for Eq 4.6.

Alternatively, if we have defined model reliance as $MR(f) = e_{\text{switch}}(f) - e_{\text{orig}}(f)$ (see Appendix A.5), with $\widehat{MR}(f) = \hat{e}_{\text{switch}}(f) - \hat{e}_{\text{orig}}(f)$, and

$$\mathcal{Q}_{\text{in,difference}} = B_{\text{ind}} \left\{ \sqrt{\frac{\log(8\delta^{-1}\mathcal{N}\left(\mathcal{F}, r\sqrt{2}\right))}{n}} + \sqrt{\frac{\log(8\delta^{-1}\mathcal{N}\left(\mathcal{F}, r\right))}{2n}} \right\} + 2r(\sqrt{2}+1)$$
$$= q_{\text{difference}} \left(B_{\text{ind}} \sqrt{\frac{\log(8\delta^{-1}\mathcal{N}\left(\mathcal{F}, r\sqrt{2}\right))}{n}} + 2r\sqrt{2}, B_{\text{ind}} \sqrt{\frac{\log(8\delta^{-1}\mathcal{N}\left(\mathcal{F}, r\right))}{2n}} + 2r \right),$$

then same proof of Eq 4.6 holds without Assumption 3 if we replace $\mathcal{Q}_{\text{in}}$ with $\mathcal{Q}_{\text{in,difference}}$, and apply Eq B.26 in Eq B.43.

For Eq 4.7 we can use the same approach. Consider the event that

$$\exists \hat{f}_{-,\epsilon_{\text{in}}} \in \underset{f \in \hat{\mathcal{R}}(\epsilon_{\text{in}})}{\arg\min} \widehat{MR}(f) : \; MCR_{-}(\epsilon) > MR\left(\hat{f}_{-,\epsilon_{\text{in}}}\right). \tag{B.44}$$

Applying steps analogous to those used to derive Eq B.40, we have

$$\mathbb{P}\left(\exists \hat{f}_{-,\epsilon_{\text{in}}} \in \underset{f \in \hat{\mathcal{R}}(\epsilon_{\text{in}})}{\arg\min} \widehat{MR}(f) : \; MCR_{-}(\epsilon) > MR\left(\hat{f}_{-,\epsilon_{\text{in}}}\right) \right)$$
$$\leq \quad \mathbb{P}\left(\exists f \in \mathcal{F}, \mathbb{E}\tilde{L}(f,Z) > \epsilon : \hat{\mathbb{E}}\tilde{L}(f,Z) \leq \epsilon_{\text{in}} \right) \quad \leq \quad \frac{\delta}{2}.$$

Analogous to B.41, when Eq B.44 does not hold, we have have

$$\begin{aligned}
MCR_{-}(\epsilon) &\leq MR\left(\hat{f}_{-,\epsilon_{\text{in}}}\right) && \text{for all } \hat{f}_{-,\epsilon_{\text{in}}} \in \underset{f \in \hat{\mathcal{R}}(\epsilon_{\text{in}})}{\arg\min} \widehat{MR}(f) \\
&= \widehat{MR}\left(\hat{f}_{-,\epsilon_{\text{in}}}\right) + \left\{ MR\left(\hat{f}_{-,\epsilon_{\text{in}}}\right) - \widehat{MR}\left(\hat{f}_{-,\epsilon_{\text{in}}}\right) \right\} && \text{for all } \hat{f}_{-,\epsilon_{\text{in}}} \in \underset{f \in \hat{\mathcal{R}}(\epsilon_{\text{in}})}{\arg\min} \widehat{MR}(f) \\
&= \widehat{MCR}_{-}\left(\epsilon_{\text{in}}\right) + \left\{ MR\left(\hat{f}_{-,\epsilon_{\text{in}}}\right) - \widehat{MR}\left(\hat{f}_{-,\epsilon_{\text{in}}}\right) \right\} && \text{for all } \hat{f}_{-,\epsilon_{\text{in}}} \in \underset{f \in \hat{\mathcal{R}}(\epsilon_{\text{in}})}{\arg\min} \widehat{MR}(f) \\
&\leq \widehat{MCR}_{-}\left(\epsilon_{\text{in}}\right) + \sup_{f \in \mathcal{F}} |MR(f) - \widehat{MR}(f)|
\end{aligned}$$

Finally, analogous to Eq B.43,

$$\begin{aligned}
&\mathbb{P}\left(MCR_-(\epsilon) > \widehat{MCR}\left(\epsilon_{\text{in}}\right) + \mathcal{Q}_{\text{in}}\right) \\
&\quad \leq \mathbb{P}\left[\left\{\exists \hat{f}_{-,\epsilon_{\text{in}}} \in \underset{f\in\hat{\mathcal{R}}(\epsilon_{\text{in}})}{\arg\min}\, \widehat{MR}(f) :\, MCR_-(\epsilon) > MR\left(\hat{f}_{-,\epsilon_{\text{in}}}\right)\right\}\right. \\
&\qquad\qquad \left.\bigcup \left\{\sup_{f\in\mathcal{F}} |\widehat{MR}(f) - MR(f)| > \mathcal{Q}_{\text{in}}\right\}\right] \\
&\quad \leq \mathbb{P}\left[\exists \hat{f}_{-,\epsilon_{\text{in}}} \in \underset{f\in\hat{\mathcal{R}}(\epsilon_{\text{in}})}{\arg\min}\, \widehat{MR}(f) :\, MCR_-(\epsilon) > MR\left(\hat{f}_{-,\epsilon_{\text{in}}}\right)\right] \\
&\qquad\qquad + \mathbb{P}\left[\sup_{f\in\mathcal{F}} |\widehat{MR}(f) - MR(f)| > \mathcal{Q}_{\text{in}}\right] \\
&\quad = \frac{\delta}{2} + \frac{\delta}{2}.
\end{aligned} \tag{B.45}$$

Under the difference-based definition of model reliance (see Appendix A.5), the same proof for Eq 4.7 holds without Assumption 3 if we replace MR, $\widehat{MR}$, & $\mathcal{Q}_{\text{in}}$ respectively with $MR_{\text{difference}}$, $\widehat{MR}_{\text{difference}}$, & $\mathcal{Q}_{\text{in,difference}}$, and apply Eq B.26 in Eq B.45. ■

B.6. Proof of Proposition 7, and Corollary for a Unique Best-in-class Model.

We first introduce a lemma to describe the performance of any individual model in the population ϵ-Rashomon set.

Lemma 26 *Let $\epsilon_1' := 2B_{ref}\sqrt{\frac{\log(\delta^{-1})}{2n}}$, and let the functions $\hat{\phi}_-$ and $\hat{\phi}_+$ be defined as in Proposition 7. Given a function $f_1 \in \mathcal{R}(\epsilon)$, if Assumption 2 holds for f_1, then*

$$\mathbb{P}\left\{\phi(f_1) \in \left[\hat{\phi}_-(\epsilon_1'), \hat{\phi}_+(\epsilon_1')\right]\right\} \geq 1 - \delta.$$

Proof Consider the event that

$$\phi(f_1) \in \left[\hat{\phi}_-(\epsilon_1'), \hat{\phi}_+(\epsilon_1')\right]. \tag{B.46}$$

Eq B.46 will always hold if $f_1 \in \hat{\mathcal{R}}(\epsilon_1')$, since the interval $\left[\hat{\phi}_-(\epsilon_1'), \hat{\phi}_+(\epsilon_1')\right]$ contains $\phi(f)$ for any $f \in \hat{\mathcal{R}}(\epsilon_1')$ by definition. Thus,

$$\begin{aligned}
\mathbb{P}\left\{\phi(f_1) \notin [\left[\hat{\phi}_-(\epsilon_1'), \hat{\phi}_+(\epsilon_1')\right]\right\} &\leq \mathbb{P}\left\{f_1 \notin \hat{\mathcal{R}}(\epsilon_1')\right\} \\
&\leq \delta \qquad \text{from Lemma 23.}
\end{aligned}$$

■

B.6.1. Proof of Proposition 7

Proof Let $f_{-,\epsilon,\phi} \in \arg\min_{f\in\mathcal{R}(\epsilon)} \phi(f)$ and $f_{+,\epsilon,\phi} \in \arg\max_{f\in\mathcal{R}(\epsilon)} \phi(f)$ respectively denote functions that attain the lowest and highest values of $\phi(f)$ among models $f \in \mathcal{R}(\epsilon)$. Applying the definitions of $f_{-,\epsilon,\phi}$ and $f_{+,\epsilon,\phi}$ in Line B.47, below, we have

$$
\begin{aligned}
&\mathbb{P}\Big(\{\phi(f) : f \in \mathcal{R}(\epsilon)\} \not\subset [\hat{\phi}_-(\epsilon'), \hat{\phi}_+(\epsilon')] \Big) \\
&\quad = \mathbb{P}\Big([\phi(f_{-,\epsilon,\phi}), \phi(f_{+,\epsilon,\phi})] \not\subset [\hat{\phi}_-(\epsilon'), \hat{\phi}_+(\epsilon')] \Big) \qquad \text{(B.47)} \\
&\quad = \mathbb{P}\Big(\phi(f_{-,\epsilon,\phi}) \notin [\hat{\phi}_-(\epsilon'), \hat{\phi}_+(\epsilon')] \quad \bigcup \quad \phi(f_{+,\epsilon,\phi}) \notin [\hat{\phi}_-(\epsilon'), \hat{\phi}_+(\epsilon')] \Big) \\
&\quad \leq \mathbb{P}\Big(\phi(f_{-,\epsilon,\phi}) \notin [\hat{\phi}_-(\epsilon'), \hat{\phi}_+(\epsilon')] \Big) + \mathbb{P}\Big(\phi(f_{+,\epsilon,\phi}) \notin [\hat{\phi}_-(\epsilon'), \hat{\phi}_+(\epsilon')] \Big) \\
&\quad \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta \qquad \text{from Lemma 26, and the definition of } \epsilon' = \epsilon + 2B_{\text{ref}}\sqrt{\frac{\log(2\delta^{-1})}{2n}}.
\end{aligned}
$$

■

B.6.2. Corollary for a Unique Best-in-Class Model

When the best-in-class model is unique, it can be described by the corollary below.

Corollary 27 *Let* $\hat{\phi}_-(\epsilon'_0) := \min_{f\in\hat{\mathcal{R}}(\epsilon'_1)} \phi(f)$ *and* $\hat{\phi}_+(\epsilon'_1) := \max_{f\in\hat{\mathcal{R}}(\epsilon'_1)} \phi(f)$, *where* $\epsilon'_0 := 2B_{ref}\sqrt{\frac{\log(\delta^{-1})}{2n}}$. *Let* $f^\star \in \arg\min_{f\in\mathcal{F}} e_{orig}(f)$ *be the prediction model that uniquely attains the lowest possible expected loss. If* $f^\star$ *satisfies Assumption 2, then*

$$\mathbb{P}\{\phi(f^\star) \in [\hat{\phi}_-(\epsilon'_1), \hat{\phi}_+(\epsilon'_1)]\} \geq 1 - \delta.$$

Proof Since $f^\star \in \mathcal{R}(0)$, Corollary 27 follows immediately from Lemma 26.

Notice that by assuming $f^\star$ is unique, we can use the threshold $\epsilon'_0 := 2B_{\text{ref}}\sqrt{\frac{\log(\delta^{-1})}{2n}}$, which is lower than the threshold of $\epsilon' = \epsilon + 2B_{\text{ref}}\sqrt{\frac{\log(2\delta^{-1})}{2n}}$ with $\epsilon = 0$, as in Proposition 7. In this way, assuming uniqueness allows a stronger statement than the one in Proposition 7. ■

B.7. Absolute Losses versus Relative Losses in the Definition of the Rashomon Set

In this paper we primarily define Rashomon sets as the models that perform well *relative* to a reference model f_{ref}. We can also study an alternate formulation of Rashomon sets by replacing the relative loss $\tilde{L}$ with the non-standardized loss L throughout. This results in a new interpretation of the Rashomon set $\mathcal{R}(\epsilon_{\text{abs}}, f_{\text{ref}}, \mathcal{F}) = \{f_{\text{ref}}\} \cup \{f \in \mathcal{F} : \mathbb{E}L(f, Z) \leq \epsilon_{\text{abs}}\}$ as the union of f_{ref} and the subset of models with *absolute* loss L no higher than ϵ_{abs}, for $\epsilon_{\text{abs}} > 0$. The process of computing empirical MCR is largely unaffected by whether L or $\tilde{L}$ is used, as it is simple to transform from one optimization problem to the other.

We still require the explicit inclusion of f_{ref} in empirical and population Rashomon sets to ensure that they are nonempty. However, in many cases, this inclusion becomes redundant when interpreting a Rashomon set (e.g., when $\epsilon \geq 0$, and $\mathbb{E}L(f_{\text{ref}}, Z) \leq \epsilon_{\text{abs}}$).

Under the replacement of $\tilde{L}$ with L, we also replace Assumption 2 with Assumption 1 (whenever this is not redundant), and replace $2B_{\text{ref}}$ with B_{ind} in the definitions of ϵ_{out}, ϵ_{best}, ϵ_{in}, ϵ' and ϵ'_1 in Theorem 4, Corollary 22, Theorem 6, Proposition 7, and Corollary 27. This is because the motivation for the $2B_{\text{ref}}$ term is that $\tilde{L}(f_1)$ is bounded within an interval of length $2B_{\text{ref}}$ when f_1 satisfies Assumption 2. However, under Assumption 1, $L(f_1)$ is bounded within an interval of length B_{ind}.

B.8. Proof of Proposition 15

Proof To show Eq 7.1 we start with $e_{\text{orig}}(f_\beta)$,

$$\begin{aligned} e_{\text{orig}}(f_\beta) &= \mathbb{E}[\{Y - X_1'\beta_1 - X_2'\beta_2\}^2] \\ &= \mathbb{E}[\{(Y - X_2'\beta_2) - X_1'\beta_1\}^2] \\ &= \mathbb{E}[(Y - X_2'\beta_2)^2] - 2\mathbb{E}[(Y - X_2'\beta_2)X_1']\beta_1 + \beta_1'\mathbb{E}[X_1X_1']\beta_1. \end{aligned}$$

For $e_{\text{switch}}(f_\beta)$, we can follow the same steps as above:

$$\begin{aligned} e_{\text{switch}}(f_\beta) &= \mathbb{E}_{Y^{(b)},X_1^{(a)},X_2^{(b)}}[\{Y^{(b)} - X_1^{(a)'}\beta_1 - X_2^{(b)'}\beta_2\}^2] \\ &= \mathbb{E}[(Y^{(b)} - X_2^{(b)'}\beta_2)^2] - 2\mathbb{E}[Y^{(b)} - X_2^{(b)'}\beta_2]\mathbb{E}[X_1^{(a)'}]\beta_1 + \beta_1'\mathbb{E}[X_1^{(a)}X_1^{(a)'}]\beta_1. \end{aligned}$$

Since $(Y^{(b)}, X_1^{(b)}, X_2^{(b)})$ and $(Y^{(a)}, X_1^{(a)}, X_2^{(a)})$ each have the same distribution as (Y, X_1, X_2), we can omit the superscript notation to show Eq 7.1:

$$\begin{aligned} e_{\text{switch}}(f_\beta) &= \mathbb{E}[(Y - X_2'\beta_2)^2] - 2\mathbb{E}[Y - X_2'\beta_2]\mathbb{E}[X_1']\beta_1 + \beta_1'\mathbb{E}[X_1X_1']\beta_1 \\ e_{\text{switch}}(f_\beta) &= e_{\text{orig}}(f_\beta) - 2\mathbb{E}[Y - X_2'\beta_2]\mathbb{E}[X_1']\beta_1 + 2\mathbb{E}[(Y - X_2'\beta_2)X_1']\beta_1 \\ e_{\text{switch}}(f_\beta) &= e_{\text{orig}}(f_\beta) + 2\text{Cov}(Y - X_2'\beta_2, X_1)\beta_1 \\ e_{\text{switch}}(f_\beta) &= e_{\text{orig}}(f_\beta) + 2\text{Cov}(Y, X_1)\beta_1 - 2\beta_2\text{Cov}(X_2, X_1)\beta_1. \end{aligned}$$

Dividing both sides by $e_{\text{orig}}(f_\beta)$ gives the desired result.

Next, we can use a similar approach to show Eq 7.2:

$$\begin{aligned} \hat{e}_{\text{switch}}(f_\beta) &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}(\mathbf{y}_{[j]} - \mathbf{X}_{2[j,\cdot]}\beta_2 - \mathbf{X}_{1[i,\cdot]}\beta_1)^2 \\ n(n-1)\hat{e}_{\text{switch}}(f_\beta) &= \sum_{i=1}^{n}\sum_{j\neq i}\left\{(\mathbf{y}_{[j]} - \mathbf{X}_{2[j,\cdot]}\beta_2)^2 - 2(\mathbf{y}_{[j]} - \mathbf{X}_{2[j,\cdot]}\beta_2)(\mathbf{X}_{1[i,\cdot]}\beta_1) + (\mathbf{X}_{1[i,\cdot]}\beta_1)^2\right\} \\ &= (n-1)\sum_{i=1}^{n}(\mathbf{y}_{[i]} - \mathbf{X}_{2[i,\cdot]}\beta_2)^2 \\ &\quad - 2\left\{\sum_{i=1}^{n}\sum_{j\neq i}(\mathbf{X}_{1[i,\cdot]}\beta_1)(\mathbf{y}_{[j]} - \mathbf{X}_{2[j,\cdot]}\beta_2)\right\} + (n-1)\sum_{i=1}^{n}(\mathbf{X}_{1[i,\cdot]}\beta_1)^2. \end{aligned} \tag{B.48}$$

Focusing on the term in braces,

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j\neq i}(\mathbf{X}_{1[i,\cdot]}\beta_1)(\mathbf{y}_{[j]}-\mathbf{X}_{2[j,\cdot]}\beta_2)\\
&\qquad=\sum_{i=1}^{n}\sum_{j=1}^{n}(\mathbf{X}_{1[i,\cdot]}\beta_1)(\mathbf{y}_{[j]}-\mathbf{X}_{2[j,\cdot]}\beta_2)-\sum_{i=1}^{n}(\mathbf{X}_{1[i,\cdot]}\beta_1)(\mathbf{y}_{[i]}-\mathbf{X}_{2[i,\cdot]}\beta_2)\\
&\qquad=\sum_{i=1}^{n}(\mathbf{X}_{1[i,\cdot]}\beta_1)\sum_{j=1}^{n}(\mathbf{y}_{[j]}-\mathbf{X}_{2[j,\cdot]}\beta_2)-\sum_{i=1}^{n}(\mathbf{X}_{1[i,\cdot]}\beta_1)(\mathbf{y}_{[i]}-\mathbf{X}_{2[i,\cdot]}\beta_2)\\
&\qquad=\left\{(\mathbf{X}_1\beta_1)'\mathbf{1}_n\right\}\left\{\mathbf{1}_n'(\mathbf{y}-\mathbf{X}_2\beta_2)\right\}-(\mathbf{X}_1\beta_1)'(\mathbf{y}-\mathbf{X}_2\beta_2) \qquad \text{(B.49)}\\
&\qquad=(\mathbf{X}_1\beta_1)'(\mathbf{1}_n\mathbf{1}_n'-\mathbf{I}_n)(\mathbf{y}-\mathbf{X}_2\beta_2).
\end{aligned}$$

Plugging this into Eq B.48, and applying the sample linear algebra representations as in Eq B.49, we get

$$\begin{aligned}
n(n-1)\hat{e}_{\text{switch}}(f_\beta) &= (n-1)\|\mathbf{y}-\mathbf{X}_2\beta_2\|_2^2\\
&\quad-2(\mathbf{X}_1\beta_1)'(\mathbf{1}_n\mathbf{1}_n'-\mathbf{I}_n)(\mathbf{y}-\mathbf{X}_2\beta_2)\\
&\quad+(n-1)\|\mathbf{X}_1\beta_1\|_2^2\\
n\hat{e}_{\text{switch}}(f_\beta) &= \|\mathbf{y}-\mathbf{X}_2\beta_2\|_2^2\\
&\quad-2(\mathbf{X}_1\beta_1)'\mathbf{W}(\mathbf{y}-\mathbf{X}_2\beta_2)\\
&\quad+\|\mathbf{X}_1\beta_1\|_2^2\\
&=\mathbf{y}'\mathbf{y}-2\mathbf{y}'\mathbf{X}_2\beta_2+\beta_2'\mathbf{X}_2'\mathbf{X}_2\beta_2\\
&\quad-2\beta_1'\mathbf{X}_1'\mathbf{W}\mathbf{y}+2\beta_1'\mathbf{X}_1'\mathbf{W}\mathbf{X}_2\beta_2\\
&\quad+\beta_1'\mathbf{X}_1'\mathbf{X}_1\beta_1\\
&=\mathbf{y}'\mathbf{y}-2\begin{bmatrix}\mathbf{X}_1'\mathbf{W}\mathbf{y}\\ \mathbf{X}_2'\mathbf{y}\end{bmatrix}'\beta+\beta'\begin{bmatrix}\mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{W}\mathbf{X}_2\\ \mathbf{X}_2'\mathbf{W}\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2\end{bmatrix}\beta.
\end{aligned}$$

■

B.9. Proof of Proposition 19

Proof First we consider $e_{\text{orig}}(f_0)$. We briefly recall that the notation $f_0(t,c)$ refers to the *true* conditional expectation function for *both* potential outcomes Y_1, Y_0, rather than the expectation for Y_0 alone.

Under the assumption that $(Y_1, Y_0) \perp T|C$, we have $f_0(t,c) = \mathbb{E}(Y|C=c, T=t) = \mathbb{E}(Y_t|C=c)$. Applying this, we see that

$$\begin{aligned} e_{\text{orig}}(f_0) &= \mathbb{E}L(f_0,(Y,T,C)) \\ &= \mathbb{E}L(f_0,(Y_T,T,C)) \\ &= \mathbb{E}_T\mathbb{E}_{C|T}\mathbb{E}_{Y_T|C}[\{Y_T-\mathbb{E}(Y_T|C)\}^2] \\ &= \mathbb{E}_T\mathbb{E}_{C|T}\text{Var}(Y_T|C) \\ &= q\mathbb{E}_{C|T=0}\text{Var}(Y_0|C)+p\mathbb{E}_{C|T=1}\text{Var}(Y_1|C), \end{aligned} \tag{B.50}$$

where $p := \mathbb{P}(T=1)$ and $q := \mathbb{P}(T=0)$.

Now we consider $e_{\text{switch}}(f_0)$. Let $(Y_0^{(a)}, Y_1^{(a)}, T^{(a)}, C^{(a)})$ and $(Y_0^{(b)}, Y_1^{(b)}, T^{(b)}, C^{(b)})$ be a pair of independent random variable vectors, each with the same distribution as (Y_0, Y_1, T, C). Then

$$\begin{aligned} e_{\text{switch}}(f_0) &= \mathbb{E}_{T^{(b)},T^{(a)},C^{(b)},Y^{(b)}_{T^{(b)}}}[\{Y^{(b)}_{T^{(b)}}-f_0(T^{(a)},C^{(b)})\}^2] \\ &= \mathbb{E}_{T^{(b)},T^{(a)},C^{(b)},Y^{(b)}_{T^{(b)}}}[\{Y^{(b)}_{T^{(b)}}-\mathbb{E}(Y_{T^{(a)}}|C=C^{(b)})\}^2] \\ &= \mathbb{E}_{T^{(b)},T^{(a)}}\mathbb{E}_{C^{(b)}|T^{(b)}}\mathbb{E}_{Y^{(b)}_{T^{(b)}}|C^{(b)}}[\{Y^{(b)}_{T^{(b)}}-\mathbb{E}(Y_{T^{(a)}}|C=C^{(b)})\}^2]. \end{aligned}$$

First we expand the outermost expectation, over $T^{(b)}, T^{(a)}$:

$$\begin{aligned} &e_{\text{switch}}(f_0) \\ &= \sum_{i,j\in\{0,1\}} \mathbb{P}(T^{(b)}=i,T^{(a)}=j)\mathbb{E}_{C^{(b)}|T^{(b)}=i}\mathbb{E}_{Y^{(b)}_i|C^{(b)}}[\{Y^{(b)}_i-\mathbb{E}(Y_j|C=C^{(b)})\}^2]. \end{aligned} \tag{B.51}$$

Since $T^{(b)} \perp T^{(a)}$, we can write

$$\begin{aligned} \mathbb{P}(T^{(b)}=i,T^{(a)}=j) &= \mathbb{P}(T^{(b)}=i)\mathbb{P}(T^{(a)}=j) \\ &= p^{i+j}q^{2-i-j}. \end{aligned}$$

Plugging this into Eq B.51 we get

$$e_{\text{switch}}(f_0) = \sum_{i,j\in\{0,1\}} p^{i+j}q^{2-i-j}\mathbb{E}_{C^{(b)}|T^{(b)}=i}\mathbb{E}_{Y^{(b)}_i|C^{(b)}}[\{Y^{(b)}_i-\mathbb{E}(Y_j|C=C^{(b)})\}^2].$$

Since $(Y_0^{(b)}, Y_1^{(b)}, C^{(b)}, T^{(b)})$ are the only random variables remaining, we can omit the superscript notation (e.g., replace $C^{(b)}$ with C) to get

$$\begin{aligned} e_{\text{switch}}(f_0) &= \sum_{i,j\in\{0,1\}} p^{i+j}q^{2-i-j}\mathbb{E}_{C|T=i}\mathbb{E}_{Y_i|C}[\{Y_i-\mathbb{E}(Y_j|C)\}^2] \\ &=: \sum_{i,j\in\{0,1\}} A_{ij}, \end{aligned}$$

where $A_{ij} = p^{i+j}q^{2-i-j}\mathbb{E}_{C|T=i}\mathbb{E}_{Y_i|C}[\{Y_i - \mathbb{E}(Y_j|C)\}^2]$. First, we consider A_{00} and A_{11}:

$$\begin{aligned} A_{00} &= q^2\mathbb{E}_{C|T=0}\mathbb{E}_{Y_0|C}[\{Y_0 - \mathbb{E}(Y_0|C)\}^2] \\ &= q^2\mathbb{E}_{C|T=0}\text{Var}(Y_0|C), \end{aligned}$$

and, similarly, $A_{11} = p^2\mathbb{E}_{C|T=1}\text{Var}(Y_1|C)$.

Next we consider A_{01} and A_{10}:

$$\begin{aligned} A_{01}: &= pq\mathbb{E}_{C|T=0}\mathbb{E}_{Y_0|C}[\{Y_0 - \mathbb{E}(Y_1|C)\}^2] \\ &= pq\mathbb{E}_{C|T=0}\left(\mathbb{E}(Y_0^2|C) - 2\mathbb{E}(Y_0|C)\mathbb{E}(Y_1|C) + \mathbb{E}(Y_1|C)^2\right) \\ &= pq\mathbb{E}_{C|T=0}\left(\text{Var}(Y_0|C) + \mathbb{E}(Y_0|C)^2 - 2\mathbb{E}(Y_0|C)\mathbb{E}(Y_1|C) + \mathbb{E}(Y_1|C)^2\right) \\ &= pq\mathbb{E}_{C|T=0}\left(\text{Var}(Y_0|C) + [\mathbb{E}(Y_1|C) - \mathbb{E}(Y_0|C)]^2\right) \\ &= pq\mathbb{E}_{C|T=0}\left(\text{Var}(Y_0|C) + \text{CATE}(C)^2\right), \end{aligned}$$

and, following the same steps,

$$A_{10} = pq\mathbb{E}_{C|T=1}\left(\text{Var}(Y_1|C) + \text{CATE}(C)^2\right).$$

Plugging in A_{00}, A_{01}, A_{10}, and A_{11} we get

$$\begin{aligned} e_{\text{switch}}(f_0) &= \{A_{00} + A_{11}\} \\ &\quad + [A_{01} + A_{10}] \\ &= \left\{q^2\mathbb{E}_{C|T=0}\text{Var}(Y_0|C) + p^2\mathbb{E}_{C|T=1}\text{Var}(Y_1|C)\right\} \\ &\quad + \left[pq\mathbb{E}_{C|T=0}\left(\text{Var}(Y_0|C) + \text{CATE}(C)^2\right) + pq\mathbb{E}_{C|T=1}\left(\text{Var}(Y_1|C) + \text{CATE}(C)^2\right)\right] \\ &= \left\{q(q+p)\mathbb{E}_{C|T=0}\text{Var}(Y_0|C) + p(p+q)\mathbb{E}_{C|T=1}\text{Var}(Y_1|C)\right\} \quad \text{(B.52)} \\ &\quad + pq\left[\mathbb{E}_{C|T=0}\left(\text{CATE}(C)^2\right) + \mathbb{E}_{C|T=1}\left(\text{CATE}(C)^2\right)\right] \quad \text{(B.53)} \\ &= \{e_{\text{orig}}(f_0)\} \quad \text{(B.54)} \\ &\quad + \text{Var}(T)\left[\mathbb{E}_{C|T=0}\left(\text{CATE}(C)^2\right) + \mathbb{E}_{C|T=1}\left(\text{CATE}(C)^2\right)\right]. \quad \text{(B.55)} \end{aligned}$$

In Lines B.52 and B.53, we consolidate terms involving $\mathbb{E}_{C|T=0}\text{Var}(Y_0|C)$ and $\mathbb{E}_{C|T=1}\text{Var}(Y_1|C)$. In Line B.54, we use $p+q=1$ to reduce Line B.52 to the right-hand side of Eq B.50. Finally, in Line B.55, we use $qp = \text{Var}(T)$. Dividing both sides by $e_{\text{orig}}(f_0) = \mathbb{E}_{T,C}Var(Y|T,C)$ gives the desired result. ■

Appendix C. Proofs for Computational Results

Almost all of the proofs in this section are unchanged if we replace $\hat{e}_{\text{switch}}(f)$ with $\hat{e}_{\text{divide}}(f)$ in our definitions of $\hat{h}_{-,\gamma}$, $\hat{h}_{+,\gamma}$, $\hat{g}_{-,\gamma}$, $\hat{g}_{+,\gamma}$, and $\widehat{MR}$. The only exception is in Appendix C.3.

Throughout the following proofs, we will make use of the fact that, for constants $a, b, c, d \in \mathbb{R}$ satisfying $a \geq c$, the relation $a + b \leq c + d$ implies

$$\begin{aligned} a + b &\leq c + d \\ a - c &\leq d - b \\ 0 &\leq d - b \qquad \text{since } 0 \leq a - c \\ b &\leq d. \end{aligned} \tag{C.1}$$

We also make use of the fact that for any $\gamma_1, \gamma_2 \in \mathbb{R}$, the definitions of $\hat{g}_{+,\gamma_1}$ and $\hat{g}_{-,\gamma_1}$ imply

$$\hat{h}_{+,\gamma_1}(\hat{g}_{+,\gamma_1}) \leq \hat{h}_{+,\gamma_1}(\hat{g}_{+,\gamma_2}), \quad \text{and} \quad \hat{h}_{-,\gamma_1}(g_{-,\gamma_1}) \leq \hat{h}_{-,\gamma_1}(g_{-,\gamma_2}). \tag{C.2}$$

Finally, for any two values $\gamma_1, \gamma_2 \in \mathbb{R}$, we make use of the fact that

$$\begin{aligned} \hat{h}_{+,\gamma_1}(f) &= \hat{e}_{\text{orig}}(f) + \gamma_1 \hat{e}_{\text{switch}}(f) \\ &= \hat{e}_{\text{orig}}(f) + \gamma_2 \hat{e}_{\text{switch}}(f) + \{\gamma_1 \hat{e}_{\text{switch}}(f) - \gamma_2 \hat{e}_{\text{switch}}(f)\} \\ &= \hat{h}_{+,\gamma_2}(f) + (\gamma_1 - \gamma_2)\hat{e}_{\text{switch}}(f), \end{aligned} \tag{C.3}$$

and, by the same steps,

$$\hat{h}_{-,\gamma_1}(f) = \hat{h}_{-,\gamma_2}(f) + (\gamma_1 - \gamma_2)\hat{e}_{\text{orig}}(f). \tag{C.4}$$

C.1. Proof of Lemma 9 (Lower Bound for MR)

Proof We prove Lemma 9 in 2 parts.

C.1.1. Part 1: Showing Eq 6.1 Holds for All $f \in \mathcal{F}$ Satisfying $\hat{e}_{\text{ORIG}}(f) \leq \epsilon_{\text{ABS}}$.

If $\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma}) \geq 0$, then for any function $f \in \mathcal{F}$ satisfying $\hat{e}_{\text{orig}}(f) \leq \epsilon_{\text{abs}}$ we know that

$$\begin{aligned} \frac{1}{\epsilon_{\text{abs}}} &\leq \frac{1}{\hat{e}_{\text{orig}}(f)} \\ \frac{\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma})}{\epsilon_{\text{abs}}} &\leq \frac{\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma})}{\hat{e}_{\text{orig}}(f)}. \end{aligned} \tag{C.5}$$

Now, for any $f \in \mathcal{F}$ satisfying $\hat{e}_{\text{orig}}(f) \leq \epsilon_{\text{abs}}$, the definition of $\hat{g}_{-,\gamma}$ implies that

$$\begin{aligned} \hat{h}_{-,\gamma}(f) &\geq \hat{h}_{-,\gamma}\left(\hat{g}_{-,\gamma}\right) \\ \gamma \hat{e}_{\text{orig}}(f) + \hat{e}_{\text{switch}}(f) &\geq \hat{h}_{-,\gamma}\left(\hat{g}_{-,\gamma}\right) \\ \gamma + \frac{\hat{e}_{\text{switch}}(f)}{\hat{e}_{\text{orig}}(f)} &\geq \frac{\hat{h}_{-,\gamma}\left(\hat{g}_{-,\gamma}\right)}{\hat{e}_{\text{orig}}(f)} \\ \gamma + \frac{\hat{e}_{\text{switch}}(f)}{\hat{e}_{\text{orig}}(f)} &\geq \frac{\hat{h}_{-,\gamma}\left(\hat{g}_{-,\gamma}\right)}{\epsilon_{\text{abs}}} \qquad \text{from Eq C.5} \\ \widehat{MR}(f) &\geq \frac{\hat{h}_{-,\gamma}\left(\hat{g}_{-,\gamma}\right)}{\epsilon_{\text{abs}}} - \gamma. \end{aligned}$$

C.1.2. Part 2: Showing that, if $f = \hat{g}_{-,\gamma}$, and at Least One of the Inequalities in Condition 8 Holds with Equality, then Eq 6.1 Holds with Equality.

We consider each of the two inequalities in Condition 8 separately. If $\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma}) = 0$, then

$$\begin{aligned} 0 &= \gamma \hat{e}_{\text{orig}}(\hat{g}_{-,\gamma}) + \hat{e}_{\text{switch}}(\hat{g}_{-,\gamma}) \\ \frac{-\hat{e}_{\text{switch}}(\hat{g}_{-,\gamma})}{\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma})} &= \gamma. \end{aligned}$$

As a result

$$\frac{\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma})}{\epsilon_{\text{abs}}} - \gamma = \frac{0}{\epsilon_{\text{abs}}} - \left\{ \frac{-\hat{e}_{\text{switch}}(\hat{g}_{-,\gamma})}{\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma})} \right\} = \widehat{MR}(\hat{g}_{-,\gamma}).$$

Alternatively, if $\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma}) = \epsilon_{\text{abs}}$, then

$$\frac{\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma})}{\epsilon_{\text{abs}}} - \gamma = \frac{\gamma \hat{e}_{\text{orig}}(\hat{g}_{-,\gamma}) + \hat{e}_{\text{switch}}(\hat{g}_{-,\gamma})}{\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma})} - \gamma = \gamma + \frac{\hat{e}_{\text{switch}}(\hat{g}_{-,\gamma})}{\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma})} - \gamma = \widehat{MR}(\hat{g}_{-,\gamma}).$$

■

C.2. Proof of Lemma 10 (Monotonicity for MR Lower Bound Binary Search)

Proof We prove Lemma 10 in 3 parts.

C.2.1. Part 1: $\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma})$ is Monotonically Increasing in γ.

Let $\gamma_1, \gamma_2 \in \mathbb{R}$ satisfy $\gamma_1 < \gamma_2$. We have assumed that $0 < \hat{e}_{\text{orig}}(f)$ for any $f \in \mathcal{F}$. Thus, for any $f \in \mathcal{F}$ we have

$$\begin{aligned} \gamma_1 \hat{e}_{\text{orig}}(f) + \hat{e}_{\text{switch}}(f) &< \gamma_2 \hat{e}_{\text{orig}}(f) + \hat{e}_{\text{switch}}(f) \\ \hat{h}_{-,\gamma_1}(f) &< \hat{h}_{-,\gamma_2}(f). \end{aligned} \tag{C.6}$$

Applying this, we have

$$\begin{aligned} \hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_1}) &\leq \hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_2}) && \text{from Eq C.2} \\ &\leq \hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_2}) && \text{from Eq C.6.} \end{aligned}$$

This result is analogous to Lemma 3 from Dinkelbach (1967).

C.2.2. Part 2: $\hat{e}_{\text{ORIG}}(\hat{g}_{-,\gamma})$ is Monotonically Decreasing in γ.

Let $\gamma_1, \gamma_2 \in \mathbb{R}$ satisfy $\gamma_1 < \gamma_2$. Then

$$\begin{aligned} \hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_1}) &\leq \hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_2}) && \text{from Eq C.2} \\ \hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_1}) + (\gamma_1 - \gamma_2)\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_1}) &\leq \hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_2}) + (\gamma_1 - \gamma_2)\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_2}) && \text{from Eq C.4} \\ (\gamma_1 - \gamma_2)\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_1}) &\leq (\gamma_1 - \gamma_2)\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_2}) && \text{from Eqs C.1 \& C.2} \\ \hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_1}) &\geq \hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_2}). \end{aligned}$$

C.2.3. Part 3: $\left\{\frac{\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma})}{\epsilon_{\text{ABS}}} - \gamma\right\}$ is Monotonically Decreasing in γ in the Range Where $\hat{e}_{\text{ORIG}}(\hat{g}_{-,\gamma}) \leq \epsilon_{\text{ABS}}$, and Increasing Otherwise.

Suppose $\gamma_1 < \gamma_2$ and $\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_1}), \hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_2}) \leq \epsilon_{\text{abs}}$. Then, from Eq C.2,

$$\begin{aligned}
&\hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_2}) \leq \hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_1}) \\
&\hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_2}) \leq \hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_1}) + (\gamma_2 - \gamma_1)\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_1}) && \text{from Eq C.4} \\
&\hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_2}) \leq \hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_1}) + (\gamma_2 - \gamma_1)\epsilon_{\text{abs}} && \text{from } \hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_1}) \leq \epsilon_{\text{abs}} \\
&\frac{\hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_2})}{\epsilon_{\text{abs}}} - \gamma_2 \leq \frac{\hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_1})}{\epsilon_{\text{abs}}} - \gamma_1.
\end{aligned}$$

Similarly, if $\gamma_1 < \gamma_2$ and $\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_1}), \hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_2}) \geq \epsilon_{\text{abs}}$. Then, from Eq C.2

$$\begin{aligned}
&\hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_1}) \leq \hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_2}) \\
&\hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_1}) \leq \hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_2}) + (\gamma_1 - \gamma_2)\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_2}) && \text{from Eq C.4} \\
&\hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_1}) \leq \hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_2}) + (\gamma_1 - \gamma_2)\epsilon_{\text{abs}} && \text{from } \hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_1}) \geq \epsilon_{\text{abs}} \\
&\frac{\hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_1})}{\epsilon_{\text{abs}}} - \gamma_1 \leq \frac{\hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_2})}{\epsilon_{\text{abs}}} - \gamma_2.
\end{aligned}$$

■

C.3. Proof of Proposition 11 (Nonnegative Weights for MR Lower Bound Binary Search)

Proof Let $\gamma_1 := \frac{1}{n-1}$. First we show that there exists a function $\hat{g}_{-,\gamma_1}$ minimizing $\hat{h}_{-,\gamma_1}$ such that $\widehat{MR}(\hat{g}_{-,\gamma_1}) = 1$. Let D_s denote the sample distribution of the data, and let D_m be the distribution satisfying

$$\begin{aligned}
\mathbb{P}_{D_m}\{(Y, X_1, X_2) = (y, x_1, x_2)\} &= \mathbb{P}_{D_s}\{(Y, X_2) = (y, x_2)\} \times \mathbb{P}_{D_s}(X_1 = x_1) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} 1(\mathbf{y}_{[i]} = y \text{ and } \mathbf{X}_{2[i]} = x_2)\sum_{j=1}^{n} 1(\mathbf{X}_{1[j]} = x_1).
\end{aligned}$$

From $\gamma_1 = \frac{1}{n-1}$ and Eq 6.2, we see that

$$\begin{aligned}
\hat{h}_{-,\gamma_1}(f) &= \sum_{i=1}^{n}\sum_{j=1}^{n} L\{f, (\mathbf{y}_{[i]}, \mathbf{X}_{1[j]}, \mathbf{X}_{2[i]})\} \times \left\{\frac{\gamma_1 1(i=j)}{n} + \frac{1(i \neq j)}{n(n-1)}\right\} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} L\{f, (\mathbf{y}_{[i]}, \mathbf{X}_{1[j]}, \mathbf{X}_{2[i]})\} \times \left\{\frac{1}{n(n-1)}\right\}. \\
&\propto \mathbb{E}_{D_m} L\{f, (Y, X_1, X_2)\}.
\end{aligned}$$

Thus, from Condition 2 of Proposition 11, we know there exists a function $\hat{g}_{-,\gamma_1}$ that minimizes $\hat{h}_{-,\gamma_1}$ with $\hat{g}_{-,\gamma_1}(x_1^{(a)}, x_2) = \hat{g}_{-,\gamma_1}(x_1^{(b)}, x_2)$ for any $x_1^{(a)}, x_1^{(b)} \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$. Condition 1 of Proposition 11 then implies that $L\{\hat{g}_{-,\gamma_1}, (y, x_1^{(a)}, x_2)\} = L\{\hat{g}_{-,\gamma_1}, (y, x_1^{(b)}, x_2)\}$

for any $x_1^{(a)}, x_1^{(b)} \in \mathcal{X}_1$, $x_2 \in \mathcal{X}_2$, and $y \in \mathcal{Y}$. We apply this result in Line C.7, below, to show that loss of model $\hat{g}_{-,\gamma_1}$ is unaffected by permuting X_1 within our sample:

$$\begin{aligned}
\hat{e}_{\text{switch}}(\hat{g}_{-,\gamma_1}) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n-1}\sum_{j\neq i} L\{\hat{g}_{-,\gamma_1}, (\mathbf{y}_{[i]}, \mathbf{X}_{1[j]}, \mathbf{X}_{2[i]})\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n-1}\sum_{j\neq i} L\{\hat{g}_{-,\gamma_1}, (\mathbf{y}_{[i]}, \mathbf{X}_{1[i]}, \mathbf{X}_{2[i]})\} \qquad \text{(C.7)} \\
&= \frac{1}{n}\sum_{i=1}^{n} L\{\hat{g}_{-,\gamma_1}, (\mathbf{y}_{[i]}, \mathbf{X}_{1[i]}, \mathbf{X}_{2[i]})\} \\
&= \hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_1}).
\end{aligned}$$

It follows that $\widehat{MR}(\hat{g}_{-,\gamma_1}) = 1$. To show the result of Proposition 11, let $\gamma_2 = 0$. For any function $\hat{g}_{-,\gamma_2}$ minimizing $\hat{h}_{-,\gamma_2}$, we know that

$$\begin{aligned}
\hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_2}) &\leq \hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_1}) && \text{from the definition of } \hat{g}_{-,\gamma_2} \\
0 + \hat{e}_{\text{switch}}(\hat{g}_{-,\gamma_2}) &\leq 0 + \hat{e}_{\text{switch}}(\hat{g}_{-,\gamma_1}) && \text{from } \gamma_2 = 0 \text{ and the definition of } \hat{h}_{-,\gamma_2}. \qquad \text{(C.8)}
\end{aligned}$$

From $\gamma_2 \leq \gamma_1$, and Part 2 of Lemma 10, we know that

$$\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_2}) \geq \hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_1}). \qquad \text{(C.9)}$$

Combining Eqs C.8 and C.9, we have

$$\widehat{MR}(\hat{g}_{-,\gamma_2}) = \frac{\hat{e}_{\text{switch}}(\hat{g}_{-,\gamma_2})}{\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_2})} \leq \frac{\hat{e}_{\text{switch}}(\hat{g}_{-,\gamma_1})}{\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_1})} = \widehat{MR}(\hat{g}_{-,\gamma_1}) = 1. \qquad \text{(C.10)}$$

Since $\hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_2}) = \hat{e}_{\text{switch}}(\hat{g}_{-,\gamma_2}) \geq 0$ by definition, Condition 8 holds for γ_2, ϵ_{abs} and $\hat{g}_{-,\gamma_2}$ if and only if $\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_2}) \leq \epsilon_{\text{abs}}$. This, combined with Eq C.10, completes the proof.

The same result does not necessarily hold if we replace $\hat{e}_{\text{switch}}$ with $\hat{e}_{\text{divide}}$ in our definitions of $\hat{h}_{-,\gamma}$, $\widehat{MR}$, and $\widehat{MCR}_-$. This is because $\hat{e}_{\text{divide}}$ does not correspond to the expectation over a distribution in which X_1 is independent of X_2 and Y, due to the fixed pairing structure used in $\hat{e}_{\text{divide}}$. Thus, Condition 2 of Proposition 11 will not apply. ■

C.4. Proof of Lemma 13 (Upper Bound for MR)

Proof We prove Lemma 13 in 2 parts.

C.4.1. Part 1: Showing Eq 6.4 Holds for All $f \in \mathcal{F}$ Satisfying $\hat{e}_{\text{ORIG}}(f) \leq \epsilon_{\text{ABS}}$.

If $\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma}) \geq 0$, then for any function $f \in \mathcal{F}$ satisfying $\hat{e}_{\text{orig}}(f) \leq \epsilon_{\text{abs}}$ we know that

$$\begin{aligned}
\frac{1}{\epsilon_{\text{abs}}} &\leq \frac{1}{\hat{e}_{\text{orig}}(f)} \\
\frac{\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma})}{\epsilon_{\text{abs}}} &\leq \frac{\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma})}{\hat{e}_{\text{orig}}(f)}. \qquad \text{(C.11)}
\end{aligned}$$

Now, if $\gamma \leq 0$, then for any $f \in \mathcal{F}$ satisfying $\hat{e}_{\text{orig}}(f) \leq \epsilon_{\text{abs}}$, the definition of $\hat{g}_{+,\gamma}$ implies

$$\begin{aligned}
\hat{h}_{+,\gamma}(f) &\geq \hat{h}_{+,\gamma}\left(\hat{g}_{+,\gamma}\right) \\
\hat{e}_{\text{orig}}(f) + \gamma \hat{e}_{\text{switch}}(f) &\geq \hat{h}_{+,\gamma}\left(\hat{g}_{+,\gamma}\right) \\
1 + \gamma \frac{\hat{e}_{\text{switch}}(f)}{\hat{e}_{\text{orig}}(f)} &\geq \frac{\hat{h}_{+,\gamma}\left(\hat{g}_{+,\gamma}\right)}{\hat{e}_{\text{orig}}(f)} \\
1 + \gamma \frac{\hat{e}_{\text{switch}}(f)}{\hat{e}_{\text{orig}}(f)} &\geq \frac{\hat{h}_{+,\gamma}\left(\hat{g}_{+,\gamma}\right)}{\epsilon_{\text{abs}}} && \text{from Eq C.11} \\
1 + \gamma \widehat{MR}(f) &\geq \frac{\hat{h}_{+,\gamma}\left(\hat{g}_{+,\gamma}\right)}{\epsilon_{\text{abs}}} \\
\widehat{MR}(f) &\leq \left\{ \frac{\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma})}{\epsilon_{\text{abs}}} - 1 \right\} \gamma^{-1}.
\end{aligned}$$

C.4.2. Part 2: Showing that if $f = \hat{g}_{+,\gamma}$, and at Least One of the Enequalities in Condition 12 Holds with Equality, then Eq 6.4 Holds with Equality.

We consider each of the two inequalities in Condition 12 separately. If $\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma}) = 0$, then

$$\begin{aligned}
0 &= \hat{e}_{\text{orig}}(\hat{g}_{+,\gamma}) + \gamma \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma}) \\
-\gamma \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma}) &= \hat{e}_{\text{orig}}(\hat{g}_{+,\gamma}) \\
-\frac{\hat{e}_{\text{switch}}(\hat{g}_{+,\gamma})}{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma})} &= \frac{1}{\gamma}.
\end{aligned}$$

As a result,

$$\left\{ \frac{\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma})}{\epsilon_{\text{abs}}} - 1 \right\} \gamma^{-1} = \left\{ \frac{0}{\epsilon_{\text{abs}}} - 1 \right\} \left\{ -\frac{\hat{e}_{\text{switch}}(\hat{g}_{+,\gamma})}{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma})} \right\} = \widehat{MR}(\hat{g}_{+,\gamma}).$$

Alternatively, if $\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma}) = \epsilon_{\text{abs}}$, then

$$\begin{aligned}
\left\{ \frac{\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma})}{\epsilon_{\text{abs}}} - 1 \right\} \gamma^{-1} = \left\{ \frac{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma}) + \gamma \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma})}{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma})} - 1 \right\} \gamma^{-1} &= \left\{ 1 + \gamma \frac{\hat{e}_{\text{switch}}(\hat{g}_{+,\gamma})}{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma})} - 1 \right\} \gamma^{-1} \\
&= \widehat{MR}(\hat{g}_{+,\gamma}).
\end{aligned}$$

■

C.5. Proof of Lemma 14 (Monotonicity for MR Upper Bound Binary Search)

Proof We prove Lemma 14 in 3 parts.

C.5.1. Part 1: $\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma})$ is Monotonically Increasing in γ.

Let $\gamma_1, \gamma_2 \in \mathbb{R}$ satisfy $\gamma_1 < \gamma_2$. We have assumed that $0 \leq \hat{e}_{\text{switch}}(f)$ for any $f \in \mathcal{F}$. Thus, for any $f \in \mathcal{F}$ we have

$$\begin{aligned} \hat{e}_{\text{orig}}(f) + \gamma_1 \hat{e}_{\text{switch}}(f) &< \hat{e}_{\text{orig}}(f) + \gamma_2 \hat{e}_{\text{switch}}(f) \\ \hat{h}_{+,\gamma_1}(f) &< \hat{h}_{+,\gamma_2}(f). \end{aligned} \tag{C.12}$$

Applying this, we have

$$\begin{aligned} \hat{h}_{+,\gamma_1}(\hat{g}_{+,\gamma_1}) &\leq \hat{h}_{+,\gamma_1}(\hat{g}_{+,\gamma_2}) && \text{from Eq C.2} \\ &< \hat{h}_{+,\gamma_2}(\hat{g}_{+,\gamma_2}) && \text{from Eq C.12.} \end{aligned}$$

C.5.2. Part 2: $\hat{e}_{\text{ORIG}}(\hat{g}_{+,\gamma})$ is Monotonically Decreasing in γ for $\gamma \leq 0$, and Condition 12 Holds for $\gamma = 0$ and $\epsilon_{\text{ABS}} \geq \min_{f \in \mathcal{F}} \hat{e}_{\text{ORIG}}(f)$.

Let $\gamma_1, \gamma_2 \in \mathbb{R}$ satisfy $\gamma_1 < \gamma_2 \leq 0$. Then

$$\begin{aligned} \hat{h}_{+,\gamma_1}(\hat{g}_{+,\gamma_1}) &\leq \hat{h}_{+,\gamma_1}(\hat{g}_{+,\gamma_2}) && \text{from Eq C.2} \\ \hat{h}_{+,\gamma_2}(\hat{g}_{+,\gamma_1}) + (\gamma_1 - \gamma_2)\hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_1}) &\leq \hat{h}_{+,\gamma_2}(\hat{g}_{+,\gamma_2}) + (\gamma_1 - \gamma_2)\hat{e}_{\text{switch}}(g_{\gamma_2}) && \text{from Eq C.3} \\ (\gamma_1 - \gamma_2)\hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_1}) &\leq (\gamma_1 - \gamma_2)\hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_2}) && \text{from Eqs C.1 \& C.2} \\ \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_1}) &\geq \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_2}) \\ \gamma_2 \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_1}) &\leq \gamma_2 \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_2}) && \text{from } \gamma_2 \leq 0. \end{aligned} \tag{C.13}$$

Now we are equipped to show the result that $\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma})$ is monotonically decreasing in γ for $\gamma \leq 0$:

$$\begin{aligned} \hat{h}_{+,\gamma_2}(\hat{g}_{+,\gamma_2}) &\leq \hat{h}_{+,\gamma_2}(\hat{g}_{+,\gamma_1}) && \text{from Eq C.2} \\ \hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_2}) + \gamma_2 \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_2}) &\leq \hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_1}) + \gamma_2 \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_1}) \\ \hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_2}) &\leq \hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_1}) && \text{from Eqs C.1 \& C.13.} \end{aligned} \tag{C.14}$$

To show that Condition 12 holds for $\gamma = 0$ and $\min_{f \in \mathcal{F}} \hat{e}_{\text{orig}}(f) \leq \epsilon_{\text{abs}}$, we first note that $h_{0,+}(g_{0,+}) = \hat{e}_{\text{orig}}(g_{0,+})$, which is positive by assumption. Second, we note that

$$\hat{e}_{\text{orig}}(g_{0,+}) = h_{0,+}(g_{0,+}) = \min_{f \in \mathcal{F}} h_{0,+}(f) = \min_{f \in \mathcal{F}} \hat{e}_{\text{orig}}(f) \leq \epsilon_{\text{abs}}.$$

C.5.3. Part 3: $\left\{ \frac{\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma})}{\epsilon_{\text{ABS}}} - 1 \right\} \gamma^{-1}$ is Monotonically Increasing in γ in the Range Where $\hat{e}_{\text{ORIG}}(\hat{g}_{+,\gamma}) \leq \epsilon_{\text{ABS}}$ and $\gamma < 0$, and Decreasing in the Range Where $\hat{e}_{\text{ORIG}}(\hat{g}_{+,\gamma}) > \epsilon_{\text{ABS}}$ and $\gamma < 0$.

To prove the first result, suppose that $\gamma_1 < \gamma_2 < 0$ and $\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_1}), \hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_2}) \leq \epsilon_{\text{abs}}$. This implies

$$\begin{aligned} \frac{1}{\gamma_2} &< \frac{1}{\gamma_1} \\ \frac{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_1}) - \epsilon_{\text{abs}}}{\gamma_2} &> \frac{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_1}) - \epsilon_{\text{abs}}}{\gamma_1}. \end{aligned} \tag{C.15}$$

Then, starting with Eq C.2,

$$\begin{aligned}
\hat{h}_{+,\gamma_2}(\hat{g}_{+,\gamma_2}) &\leq \hat{h}_{+,\gamma_2}(\hat{g}_{+,\gamma_1}) \\
\hat{h}_{+,\gamma_2}(\hat{g}_{+,\gamma_2}) &\leq \gamma_2 \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_1}) + \hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_1}) \\
\frac{\hat{h}_{+,\gamma_2}(\hat{g}_{+,\gamma_2}) - \epsilon_{\text{abs}}}{\gamma_2} &\geq \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_1}) + \frac{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_1}) - \epsilon_{\text{abs}}}{\gamma_2} \\
&\geq \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_1}) + \frac{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_1}) - \epsilon_{\text{abs}}}{\gamma_1} \qquad \text{from Eq C.15} \\
&= \frac{\hat{h}_{+,\gamma_1}(\hat{g}_{+,\gamma_1}) - \epsilon_{\text{abs}}}{\gamma_1}.
\end{aligned}$$

Dividing both sides of the above equation by ϵ_{abs} proves that $\left\{\frac{\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma})}{\boldsymbol{\epsilon_{\text{abs}}}} - 1\right\}\gamma^{-1}$ is monotonically increasing in γ in the range where $\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma}) \leq \epsilon_{\text{abs}}$ and $\gamma < 0$.

To prove the second result we proceed in the same way. Suppose that $\gamma_1 < \gamma_2 < 0$ and $\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_1}), \hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_2}) \geq \epsilon_{\text{abs}}$. This implies

$$\begin{aligned}
\frac{1}{\gamma_2} &< \frac{1}{\gamma_1} \\
\frac{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_2}) - \epsilon_{\text{abs}}}{\gamma_2} &< \frac{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_2}) - \epsilon_{\text{abs}}}{\gamma_1}.
\end{aligned} \tag{C.16}$$

Then, starting with Eq C.2,

$$\begin{aligned}
\hat{h}_{+,\gamma_1}(\hat{g}_{+,\gamma_1}) &\leq \hat{h}_{+,\gamma_1}(\hat{g}_{+,\gamma_2}) \\
\hat{h}_{+,\gamma_1}(\hat{g}_{+,\gamma_1}) &\leq \gamma_1 \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_2}) + \hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_2}) \\
\frac{\hat{h}_{+,\gamma_1}(\hat{g}_{+,\gamma_1}) - \epsilon_{\text{abs}}}{\gamma_1} &\geq \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_2}) + \frac{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_2}) - \epsilon_{\text{abs}}}{\gamma_1} \\
&\geq \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_2}) + \frac{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_2}) - \epsilon_{\text{abs}}}{\gamma_2} \qquad \text{from Eq C.16} \\
&= \frac{\hat{h}_{+,\gamma_1}(\hat{g}_{+,\gamma_2}) - \epsilon_{\text{abs}}}{\gamma_2}.
\end{aligned}$$

Diving both sides of the above equation by ϵ_{abs} proves that $\left[\left\{\frac{\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma})}{\epsilon_{\text{abs}}} - 1\right\}\gamma^{-1}\right]$ is monotonically decreasing in γ in the range where $\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma}) > \epsilon_{\text{abs}}$ and $\gamma < 0$. ■

C.6. Proof of Remark 16 (Tractability of Empirical MCR for Linear Model Classes)

Proof To show Remark 16, we apply Proposition 15 to see that

$$\begin{aligned}
&\xi_{\text{orig}}\hat{e}_{\text{orig}}(f_\beta) + \xi_{\text{switch}}\hat{e}_{\text{switch}}(f_\beta) \\
&\quad= \frac{\xi_{\text{orig}}}{n}\,||\mathbf{y} - \mathbf{X}\beta||_2^2 + \xi_{\text{switch}}\hat{e}_{\text{switch}}(f_\beta) \\
&\quad= \frac{\xi_{\text{orig}}}{n}\left(\mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta\right) \\
&\qquad\qquad + \frac{\xi_{\text{switch}}}{n}\left\{\mathbf{y}'\mathbf{y} - 2\begin{bmatrix} \mathbf{X}_1'\mathbf{W}\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{bmatrix}'\beta + \beta'\begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{W}\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{W}\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix}\beta\right\} \\
&\quad\propto_\beta -2\mathbf{q}'\beta + \beta'\mathbf{Q}\beta.
\end{aligned}$$

■

C.7. Proof of Lemma 17 (Loss Upper Bound for Linear Models)

Proof Under the conditions in Lemma 17 and Eq 7.5, we can construct an upper bound on $L(f_\beta, (y, x)) = (y - x'\beta)^2$ by either maximizing or minimizing $x'\beta$. First, we consider the maximization problem

$$\max_{\beta, x \in \mathbb{R}^p} x'\beta \text{ subject to } x'\mathbf{M}_{\text{lm}}^{-1}x \leq r_{\mathcal{X}} \text{ and } \beta'\mathbf{M}_{\text{lm}}\beta \leq r_{\text{lm}}. \tag{C.17}$$

We can see that both constraints hold with equality at the solution to this problem. Next, we apply the change of variables $\tilde{x} = \frac{1}{\sqrt{r_{\mathcal{X}}}}\mathbf{D}^{\frac{-1}{2}}\mathbf{U}'x$ and $\tilde{\beta} = \frac{1}{\sqrt{r_{\text{lm}}}}\mathbf{D}^{\frac{1}{2}}\mathbf{U}'\beta$, where $\mathbf{U}\mathbf{D}\mathbf{U}' = \mathbf{M}_{\text{lm}}$ is the eigendecomposition of $\mathbf{M}_{\text{lm}}$. We obtain

$$\max_{\tilde{\beta}, \tilde{x} \in \mathbb{R}^p} \tilde{x}'\tilde{\beta}\sqrt{r_{\mathcal{X}} r_{\text{lm}}} \text{ subject to } \tilde{x}'\tilde{x} = 1 \text{ and } \tilde{\beta}'\tilde{\beta} = 1,$$

which has an optimal objective value equal to $\sqrt{r_{\mathcal{X}} r_{\text{lm}}}$. By negating the objective in Eq C.17, we see that the minimum possible value of $x'\beta$, subject to the constraints in Eq 7.5 and Lemma 17, is found at $-\sqrt{r_{\mathcal{X}} r_{\text{lm}}}$. Thus, we know that

$$L(f, (y, x_1, x_2)) \leq \max\left[\left\{\left(\min_{y \in \mathcal{Y}} y\right) - \sqrt{r_{\mathcal{X}} r_{\text{lm}}}\right\}^2, \left\{\left(\max_{y \in \mathcal{Y}} y\right) + \sqrt{r_{\mathcal{X}} r_{\text{lm}}}\right\}^2\right],$$

for any $(y, x_1, x_2) \in (\mathcal{Y} \times \mathcal{X}_1 \times \mathcal{X}_2)$. ■

C.8. Proof of Lemma 18 (Loss Upper Bound for Regression in a RKHS)

This proofs follows a similar structure as the proof in Section C.7. From the assumptions of Lemma 18, we know from Eq 7.7 that the largest possible output from a model $f_\alpha \in \mathcal{F}_{\mathbf{D}, r_k}$

is

$$\begin{aligned}
&\mu + \max_{x\in\mathbb{R}^p,\alpha\in\mathbb{R}^R} \sum_{i=1}^{R} k(x, \mathbf{D}_{[i,\cdot]})\alpha_{[i]} && \text{subject to } v(x)'\mathbf{K}_{\mathbf{D}}^{-1}v(x) \le r_{\mathbf{D}} \text{ and } \alpha'\mathbf{K}_{\mathbf{D}}\alpha \le r_k \\
=&\mu + \max_{x\in\mathbb{R}^p,\alpha\in\mathbb{R}^R} v(x)'\alpha && \text{subject to } v(x)'\mathbf{K}_{\mathbf{D}}^{-1}v(x) \le r_{\mathbf{D}} \text{ and } \alpha'\mathbf{K}_{\mathbf{D}}\alpha \le r_k \\
\le&\mu + \max_{\mathbf{z},\alpha\in\mathbb{R}^R} \mathbf{z}'\alpha && \text{subject to } \mathbf{z}'\mathbf{K}_{\mathbf{D}}^{-1}\mathbf{z} \le r_{\mathbf{D}} \text{ and } \alpha'\mathbf{K}_{\mathbf{D}}\alpha \le r_k.
\end{aligned}$$

The above problem can be solved in the same way as Eq C.17, and has a solution at $\left(\mu + \sqrt{r_{\mathbf{D}} r_k}\right)$. The smallest possible model output will similarly be lower bounded by $-\left(\mu + \sqrt{r_{\mathbf{D}} r_k}\right)$. Thus, B_{ind} is less than or equal to

$$\max\left[\left\{\min_{y\in\mathcal{Y}}(y) - (\mu + \sqrt{r_{\mathbf{D}} r_k})\right\}^2, \left\{\max_{y\in\mathcal{Y}}(y) + (\mu + \sqrt{r_{\mathbf{D}} r_k})\right\}^2\right].$$

References

André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

Kellie J Archer and Ryan V Kimes. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4):2249–2260, 2008.

Razia Azen, David V Budescu, and Benjamin Reiser. Criticality of predictors in multiple regression. *British Journal of Mathematical and Statistical Psychology*, 54(2):201–225, 2001.

Katherine Beckett, Kris Nyrop, and Lori Pfingst. Race, drugs, and policing: understanding disparities in drug delivery arrests. *Criminology*, 44(1):105–137, 2006.

Irene V Blair, Charles M Judd, and Kristine M Chapleau. The influence of afrocentric facial features in criminal sentencing. *Psychological science*, 15(10):674–679, 2004.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge university press, 2004.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Leo Breiman et al. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

M Luz Calle and Víctor Urrea. Letter to the editor: stability of random forest importance measures. *Briefings in bioinformatics*, 12(1):86–89, 2010.

Hugh A Chipman, Edward I George, Robert E McCulloch, et al. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

Alexandra Chouldechova. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Beau Coker, Cynthia Rudin, and Gary King. A theory of statistical inference for ensuring the robustness of scientific results. *arXiv preprint arXiv:1804.08646*, 2018.

Sam Corbett-Davies, Emma Pierson, Avi Feller, and Sharad Goel. A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear. *The Washington Post*, October 2016. URL `https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?utm_term=.e896ff1e4107`.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.

Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: theory and experiments with learning systems. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 598–617. IEEE, 2016.

Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

Olga V Demler, Michael J Pencina, and Ralph B D'Agostino Sr. Misuse of delong test to compare aucs for nested models. *Statistics in medicine*, 31(23):2577–2587, 2012.

Iván Díaz, Alan Hubbard, Anna Decker, and Mitchell Cohen. Variable importance and prediction methods for longitudinal problems with missing variables. *PloS one*, 10(3): e0120031, 2015.

Werner Dinkelbach. On nonlinear fractional programming. *Management science*, 13(7): 492–498, 1967.

Jiayun Dong and Cynthia Rudin. Variable importance clouds: A way to explore variable importance for the set of good models. *arXiv preprint arXiv:1901.03209*, 2019.

R Dorfman. A note on the delta-method for finding variance formulae. *The Biometric Bulletin*, 1(129-137):92, 1938.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

Muriel Gevrey, Ioannis Dimopoulos, and Sovan Lek. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling*, 160(3):249–264, 2003.

Baptiste Gregorutti, Bertrand Michel, and Philippe Saint-Pierre. Grouped variable importance with random forests and application to multiple functional data analysis. *Computational Statistics & Data Analysis*, 90:15–35, 2015.

Baptiste Gregorutti, Bertrand Michel, and Philippe Saint-Pierre. Correlation and variable importance in random forests. *Statistics and Computing*, 27(3):659–678, 2017.

Alexander Hapfelmeier, Torsten Hothorn, Kurt Ulm, and Carolin Strobl. A new variable importance measure for random forests with missing data. *Statistics and Computing*, 24 (1):21–34, 2014.

T Hastie, R Tibshirani, and J Friedman. *The elements of statistical learning 2nd edition*. New York: Springer, 2009.

Karl G Heider. The Rashomon effect: when ethnographers disagree. *American Anthropologist*, 90(1):73–81, 1988.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The annals of mathematical statistics*, pages 293–325, 1948.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL `https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1963.10500830`.

Giles Hooker. Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732, 2007.

Reiner Horst and Nguyen V Thoai. Dc programming: overview. *Journal of Optimization Theory and Applications*, 103(1):1–43, 1999.

Faisal Kamiran, Indrė Žliobaitė, and Toon Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and information systems*, 35(3):613–644, 2013.

Jalil Kazemitabar, Arash Amini, Adam Bloniarz, and Ameet S Talwalkar. Variable importance using decision trees. In *Advances in Neural Information Processing Systems*, pages 425–434, 2017.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica*, May 2016. URL `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`.

Guillaume Lecué. *Interplay between concentration, complexity and geometry in learning theory with applications to high dimensional data analysis*. PhD thesis, Université Paris-Est, 2011.

Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

Benjamin Letham, Portia A Letham, Cynthia Rudin, and Edward P Browne. Prediction uncertainty and optimal experimental design for learning dynamical systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 26(6):063110, 2016.

Gilles Louppe, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. Understanding variable importances in forests of randomized trees. In *Advances in neural information processing systems*, pages 431–439, 2013.

Kristian Lum and William Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Lucas Mentch and Giles Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research*, 17(1):841–881, 2016.

John Monahan and Jennifer L Skeem. Risk assessment in criminal sentencing. *Annual review of clinical psychology*, 12:489–513, 2016.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2018, page 1931. NIH Public Access, 2018.

Daniel Nevo and Ya'acov Ritov. Identifying a minimal class of models for high-dimensional data. *The Journal of Machine Learning Research*, 18(1):797–825, 2017.

Julian D Olden, Michael K Joy, and Russell G Death. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, 178(3):389–397, 2004.

Jaehyun Park and Stephen Boyd. General heuristics for nonconvex quadratically constrained quadratic programming. *arXiv preprint arXiv:1703.07870*, 2017.

Raymond Paternoster and Robert Brame. Reassessing race disparities in maryland capital cases. *Criminology*, 46(4):971–1008, 2008.

Sarah Picard-Fritsche, Michael Rempel, Jennifer A. Tallon, Julian Adler, and Natalie Reyes. Demystifying risk assessment, key principles and controversies. Technical report, 2017. Available at `https://www.courtinnovation.org/publications/demystifying-risk-assessment-key-principles-and-controversies`.

Imre Pólik and Tamás Terlaky. A survey of the s-lemma. *SIAM review*, 49(3):371–418, 2007.

Rajeev Ramchand, Rosalie Liccardo Pacula, and Martin Y Iguchi. Racial differences in marijuana-users' risk of arrest in the united states. *Drug and alcohol dependence*, 84(3): 264–272, 2006.

Friedrich Recknagel, Mark French, Pia Harkonen, and Ken-Ichi Yabunaka. Artificial neural network approach for modelling and prediction of algal blooms. *Ecological Modelling*, 96 (1-3):11–28, 1997.

Wendy D Roth and Jal D Mehta. The Rashomon effect: combining positivist and interpretivist approaches in the analysis of contested events. *Sociological Methods & Research*, 31(2):131–173, 2002.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, May 2019.

Cynthia Rudin, Caroline Wang, and Beau Coker. The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, 2019. accepted.

Michele Scardi and Lawrence W Harding. Developing an empirical model of phytoplankton primary production: a neural network case study. *Ecological modelling*, 120(2):213–223, 1999.

Lesia Semenova and Cynthia Rudin. A study in rashomon curves and volumes: a new perspective on generalization and model simplicity in machine learning. *arXiv preprint arXiv:1908.01755*, 2019.

Robert J Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 1980.

Cassia Spohn. Thirty years of sentencing reform: the quest for a racially neutral sentencing process. *Criminal justice*, 3:427–501, 2000.

Alexander Statnikov, Nikita I Lytkin, Jan Lemeire, and Constantin F Aliferis. Algorithms for discovery of multiple markov boundaries. *Journal of Machine Learning Research*, 14 (Feb):499–566, 2013.

Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25, 2007.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC bioinformatics*, 9 (1):307, 2008.

Elizabeth A Stuart. Matching methods for causal inference: a review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

Laura Toloşi and Thomas Lengauer. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, 27(14):1986–1994, 2011.

Theja Tulabandhula and Cynthia Rudin. Robust optimization using machine learning for uncertainty sets. *arXiv preprint arXiv:1407.1097*, 2014.

U.S. Department of Justice - Civil Rights Devision. Investigation of the Baltimore City Police Department, August 2016. Available at `https://www.justice.gov/crt/file/883296/download`.

Mark J van der Laan. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1), 2006.

Jay M Ver Hoef. Who invented the delta method? *The American Statistician*, 66(2): 124–127, 2012.

Huazhen Wang, Fan Yang, and Zhiyuan Luo. An experimental study of the intrinsic stability of random forest variable importance measures. *BMC bioinformatics*, 17(1):60, 2016.

Brian D Williamson, Peter B Gilbert, Noah Simon, and Marco Carone. Nonparametric variable importance assessment using machine learning techniques. *bepress (unpublished preprint)*, 2017.

Jingtao Yao, Nicholas Teng, Hean-Lee Poh, and Chew Lim Tan. Forecasting and analysis of marketing data using neural networks. *J. Inf. Sci. Eng.*, 14(4):843–862, 1998.

Ruoqing Zhu, Donglin Zeng, and Michael R Kosorok. Reinforcement learning trees. *Journal of the American Statistical Association*, 110(512):1770–1784, 2015.