Floodgate: inference for model-free variable importance

Lu Zhang and Lucas Janson

Department of Statistics, Harvard University

Abstract

Many modern applications seek to understand the relationship between an outcome variable Y and a covariate X in the presence of a (possibly high-dimensional) confounding variable Z. Although much attention has been paid to testing *whether* Y depends on X given Z, in this paper we seek to go beyond testing by inferring the *strength* of that dependence. We first define our estimand, the minimum mean squared error (mMSE) gap, which quantifies the conditional relationship between Y and X in a way that is deterministic, model-free, interpretable, and sensitive to nonlinearities and interactions. We then propose a new inferential approach called *floodgate* that can leverage any working regression function chosen by the user (allowing, e.g., it to be fitted by a state-of-the-art machine learning algorithm or be derived from qualitative domain knowledge) to construct asymptotic confidence bounds, and we apply it to the mMSE gap. We additionally show that floodgate's accuracy (distance from confidence bound to estimand) is adaptive to the error of the working regression function. We then show we can apply the same floodgate principle to a different measure of variable importance when Y is binary. Finally, we demonstrate floodgate's performance in a series of simulations and apply it to data from the UK Biobank to infer the strengths of dependence of platelet count on various groups of genetic mutations.

Keywords. Variable importance, effect size, model-X, heterogeneous treatment effects, heritability.

1 Introduction

1.1 Problem Statement

Scientists looking to better-understand the relationship between a response variable Y of interest and a covariate X in the presence of confounding variables $Z = (Z_1, \dots, Z_{p-1})$ often start by asking *how important* X is in this relationship. Although this question is sometimes simplified by statisticians to the binary question of 'is X important or not?', a more informative and useful inferential goal is to provide inference (i.e., confidence bounds) for an interpretable real-valued measure of variable importance (MOVI). The canonical approach of assuming a parametric model for $Y \mid X, Z$ will usually provide obvious MOVI candidates in terms of the model parameters, but the simple models for which it is known how to construct confidence intervals (e.g., low-dimensional or ultra-sparse generalized linear models) often provide at best very coarse approximations to the true $Y \mid X, Z$ (as evidenced by the marked predictive outperformance of nonparametric machine learning methods in many domains), resulting in undercoverage due to violated assumptions *and* lost power due to insufficient capacity to capture complex relationships. This raises the motivating question for this paper: **what is an interpretable, sensitive, and model-free measure of variable importance and how can we provide valid and narrow confidence bounds for it?**

1.2 Our contribution

The main contribution of this paper is to introduce *floodgate*, a method for inference of the minimum mean squared error (mMSE) gap, which satisfies the following high-level objectives which we believe are fairly universal for the task at hand.

(Sensitivity) The mMSE gap is strictly positive unless $\mathbb{E}[Y \mid X, Z] \overset{a.s.}{=} \mathbb{E}[Y \mid Z]$, and is large whenever X explains a lot of the variance in Y not already explained by Z alone, making it sensitive to arbitrary nonlinearities and interactions in Y's relationship with X.

(Interpretability) The mMSE gap has simple predictive, explanatory, and causal interpretations for Y's relationship with X, is a functional of *only* the joint distribution of (Y, X, Z), and is exactly zero when $Y \perp\!\!\!\perp X \mid Z$.

(Validity) We first prove floodgate's asymptotic validity assuming the user knows the distribution of $X \mid Z$, but with essentially no other assumptions (in particular we require no smoothness, sparsity, or other constraints on $\mathbb{E}[Y \mid X, Z]$ that would ensure its learnability at *any* geometric rate). However, to emphasize that the floodgate idea is not tied to such assumptions, we also provide a version of floodgate valid under double-robustness-type assumptions.

(Accuracy) Floodgate derives accuracy from flexibility by allowing the user to estimate $\mathbb{E}[Y \mid X, Z]$ in whatever way they like, and we prove that the accuracy of inference is adaptive to the mean squared error (MSE) of that estimate.

In a bit more detail, we (in Section 2) define the mMSE gap as an interpretable and model-free MOVI (Section 2.1) and present a method, *floodgate*, to construct asymptotic lower confidence bounds for it that provides the user absolute latitude to leverage any domain knowledge or advanced machine learning algorithms to make those bounds as tight as possible (Section 2.2). We consider upper confidence bounds (Section 2.3), address computational considerations (Section 2.4), theoretically characterize the width of floodgate's confidence bounds (Section 2.5), and briefly address some immediate generalizations (Section 2.6).

We then proceed to extensions of floodgate (Section 3), first presenting an alternative MOVI that we can similarly construct asymptotic confidence bounds for when Y is binary (Section 3.1). Second, we present a modification of floodgate that, for certain models, allows asymptotic inference even when X's distribution is only known up to a parametric model (Section 3.2) and apply it to multivariate Gaussian (Section 3.2.1) and discrete Markov chain (Section 3.2.2) covariate models.

Finally we demonstrate floodgate's performance and support our theory with simulations (Section 4) and an application to data from the UK Biobank (Section 5). We end with a discussion of the future research directions opened by this work (Section 6). All proofs are deferred to the appendix.

1.3 Related work

Many existing works consider *marginal* variable importance, i.e., not accounting for the presence of Z in the relationship between Y and X (Hirschfeld, 1935; Gretton et al., 2005, 2007; Székely et al., 2007; Székely and Rizzo, 2013; Heller et al., 2013; Shao and Zhang, 2014; Wang et al., 2017; Chatterjee, 2021; Deb and Sen, 2021), including some that measure that importance via differences in conditional means in a way resembling our mMSE gap (Shao and Zhang, 2014). Such approaches address a very different statistical question, and so we focus our literature review on works that, like us, consider conditional variable importance.

The standard approach to conditional statistical inference in regression is to assume a parametric model for $Y \mid X, Z$, often a generalized linear model (GLM) or cousin thereof. With $Y \mid X, Z$ so parameterized, it is usually straightforward to define a parametric MOVI and a large body of literature is available to provide asymptotic inference for such parametric MOVIs (see, for example, Bühlmann et al. (2013); Nickl et al. (2013); Zhang and Zhang (2014); Van de Geer et al. (2014); Javanmard and Montanari (2014); Bühlmann et al. (2015); Dezeure et al. (2017); Zhang and Cheng (2017)). However, when the parametric $Y \mid X, Z$ model is misspecified even slightly, the associated parametric MOVI becomes ill-defined, reducing its interpretability. Furthermore, many $Y \mid X, Z$ models are too simple to capture or detect nonlinearities that may be present in real-world data sets.

One approach to addressing the shortcomings of parametric inference is to generalize the parameters of common parametric models to be well-defined in a much larger nonparametric model class. For example, under mild moment conditions one can generalize the parameters in a linear model for $Y \mid X, Z$ as parameters in the least-squares *projection* to a linear model of any $Y \mid X, Z$ distribution (Berk et al., 2013; Taylor et al., 2014; Buja and Brown, 2014; Buja et al., 2015; Rinaldo et al., 2019; Lee et al., 2016; Buja et al., 2019a,b). Such a linear projection MOVI can be hard to interpret because it will in general have a non-zero value even when $Y \perp\!\!\!\perp X \mid Z$; see Appendix B for a simple example. Another example of a generalized parameter is the expected conditional covariance functional $\mathbb{E}[\mathrm{Cov}(Y, X \mid Z)]$ (see, for example, Robins et al. (2008, 2009); Li et al. (2011); Robins et al. (2017); Newey and Robins (2018); Shah and Peters (2020); Chernozhukov et al. (2018a); Liu et al. (2019); Katsevich and Ramdas (2020)), which represents a generalization of the linear coefficient in a *partially* linear model. $\mathbb{E}[\mathrm{Cov}(Y, X \mid Z)]$ always equals zero when $Y \perp\!\!\!\perp X \mid Z$, but it shares the shortcoming of linear projection MOVIs that it lacks sensitivity to capture nonlinearities or interactions in Y's relationship with X. That is, both MOVIs mentioned in this paragraph will assign any non-null variable that influences Y nonlinearly or through interactions with other covariates a value that can severely underrate that variable's true importance, and can even assign a variable the MOVI value zero when Y is a deterministic non-constant function of it.

A second approach has been to infer model-free MOVIs defined through machine learning algorithms fitted to part of the data itself (Lei et al., 2018; Fisher et al., 2019; Watson and Wright, 2019). By leveraging the expressiveness of machine learning, such a MOVI can be made sensitive to nonlinearities and interactions but is itself *random* and depends both on the data and the choice of machine learning algorithm. This poses a challenge for interpretability and in particular for replicability, since even *identical* analyses run on two independent data sets that are *identically-distributed* will provide inferences for *different* MOVI values.

Another line of work (Castro et al., 2009; Štrumbelj and Kononenko, 2014; Owen and Prieur, 2017; Lundberg et al., 2020; Covert et al., 2020; Williamson and Feng, 2020) considers MOVIs based on the classical form of the Shapley value (Shapley, 1953; Charnes et al., 1988), which in general assigns a non-zero MOVI value to covariates X with $Y \perp\!\!\!\perp X \mid Z$, making it hard to interpret its value mechanistically or causally (though it has some appealing properties for a *predictive* interpretation).

An interesting new proposal for a model-free MOVI was made in Azadkia and Chatterjee (2019). Their MOVI has the distinction that it equals zero if and only if $Y \perp\!\!\!\perp X \mid Z$ *and* it attains the maximum value 1 if Y is almost surely a measurable function of X given Z. More recently, Huang et al. (2020) proposed a larger class of MOVIs satisfying the same properties. However, both papers focus on consistent estimators and do not provide confidence bounds for their MOVIs.

As we will detail in Section 2.1, the MOVI we provide inference for, the mMSE gap, does not suffer from the drawbacks of the MOVIs described in the previous paragraphs, and indeed the same MOVI has been considered before. In the sensitivity analysis literature it is called the "total-effect index" (Saltelli et al., 2008) but to our knowledge its inference (confidence lower- or upper-bounds) is not considered there. In one of the Shapley value papers (Covert et al., 2020) a generalization of the mMSE gap is used as the input to the Shapley value calculation, but again inferential results (for the mMSE gap or its Shapley version) are not considered in that paper. Otherwise, Williamson et al. (2019) appears to be the first to consider inference for the mMSE gap (this inference is then used with neural networks in Feng et al. (2018)), but the asymptotic normality theory their coverage guarantee relies on fails at the boundary of the parameter space, i.e., the important case of when the mMSE gap is zero, or the variable is unimportant. A recent follow-up work (Williamson et al., 2020) addresses this limitation by combining estimators on two disjoint subsets of the data (though their inference still requires the *group* mMSE gap of the entire covariate vector to be positive). Our different approach avoids altogether this issue when the mMSE gap is zero so that our inference is valid for any value of the mMSE gap (group or otherwise), and although we also use data splitting, we do so in a way that seems to lead to significantly reduced variance (and hence more accurate inference) relative to Williamson et al. (2020), as we show in Section 4.4.

1.4 Notation

For two random variables A and B defined on the same probability space, let $P_{A \mid B}$ denote the conditional distribution of $A \mid B$. Denote the $(1-\alpha)$th quantile of the standard normal distribution by z_α. Let $\chi^2(P\|Q)$ denote the χ^2 divergence $\int_\Omega (\frac{dP}{dQ}-1)^2 dQ$ between two distributions P, Q on the probability space Ω. Let $[n]$ denote the set $\{1, \ldots, n\}$.

2 Methodology

2.1 Measuring variable importance with the mMSE gap

We begin by defining the MOVI that we will provide inference for in this paper.

Definition 2.1 (Minimum mean squared error gap)**.** *The* minimum mean squared error (mMSE) gap *for variable X is defined as*

$$\mathcal{I}^2 = \mathbb{E}\left[(Y - \mathbb{E}[Y \mid Z])^2\right] - \mathbb{E}\left[(Y - \mathbb{E}[Y \mid X, Z])^2\right] \tag{2.1}$$

whenever all the above expectations exist.

We will at times refer to either $\mathcal{I}^2$ or $\mathcal{I}$ as the mMSE gap when it causes no confusion. Although the same MOVI has been used before (see Section 1.3), we provide here a number of equivalent definitions/interpretations which we have not seen presented together before.

- Equation (2.1) has a direct *predictive* interpretation as the increase in the achievable or minimum MSE for predicting Y when X is removed.

- The mMSE gap can also be interpreted as the decrease in the *explainable variance* of Y without X:

$$\mathcal{I}^2 = \mathrm{Var}(\mathbb{E}[Y \mid X, Z]) - \mathrm{Var}(\mathbb{E}[Y \mid Z]). \tag{2.2}$$

- When X is viewed as a treatment level for Y and Z is a set of measured confounders, $\mathcal{I}$ can be seen as an *expected squared treatment effect*:

$$\mathcal{I}^2 = \frac{1}{2}\mathbb{E}_{x_1, x_2, Z}\left[(\mathbb{E}[Y \mid X = x_1, Z] - \mathbb{E}[Y \mid X = x_2, Z])^2\right]. \tag{2.3}$$

  where x_1 and x_2 are independently drawn from $P_{X|Z}$ in the outer expectation.

- We can also rewrite the mMSE gap as:

$$\mathcal{I}^2 = \mathbb{E}\left[(\mathbb{E}[Y \mid Z] - \mathbb{E}[Y \mid X, Z])^2\right] \tag{2.4}$$

  and interpret $\mathcal{I}$ as the ℓ_2 distance between the two regression functions $\mathbb{E}[Y \mid Z]$ and $\mathbb{E}[Y \mid X, Z]$.

- Lastly, we remark that $\mathcal{I}^2$ also admits a very compact (if less immediately interpretable) expression:

$$\mathcal{I}^2 = \mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid X, Z] \mid Z)]. \tag{2.5}$$

In light of these multiple alternative expressions, we find the mMSE gap remarkably interpretable. Note that it only requires the existence of some low-order conditional and unconditional moments of Y to be well-defined, and its value is invariant to any fixed translation of Y and to the replacement of X or Z by any fixed bijective function of itself. Furthermore, the mMSE gap is zero if and only if $\mathbb{E}[Y \mid X, Z] \overset{a.s.}{=} \mathbb{E}[Y \mid Z]$, and in particular it is exactly zero when $Y \perp\!\!\!\perp X \mid Z$ and strictly positive if $\mathbb{E}[Y \mid X, Z]$ depends at all on X, allowing it to fully capture arbitrary nonlinearities and interactions in $\mathbb{E}[Y \mid X, Z]$.

Note that $\mathcal{I}$ has the same units as Y, which can help interpretation when Y's units are meaningful (much like it does for the average treatment effect in causal inference). However, if a unitless quantity is preferred, such as for comparison between MOVIs across Ys with different units, we can also measure variable importance by and extend our methodology to a standardized version of $\mathcal{I}^2$, namely, $\mathcal{I}^2/\mathrm{Var}(Y)$. In fact, with some more work, we can even extend our inferential results to a version of the mMSE gap which is invariant to transformations of Y, or versions that are zero if *and only if* $Y \perp\!\!\!\perp X \mid Z$; see Section 2.6 and Appendix F for details, with Appendix F.2 extending our results to the kernel partial correlation of Huang et al. (2020).

2.2 Floodgate: asymptotic lower confidence bounds for the mMSE gap

As can be seen by Equation (2.5), the mMSE gap is a nonlinear functional of the true regression function $\mu^\star(x,z) := \mathbb{E}[Y \mid X = x, Z = z]$. Hence if we had a sufficiently-well-behaved estimator $\hat{\mu}$ for $\mu^\star$ (e.g., asymptotically normal or consistent at a sufficiently-fast geometric rate), there would be a number of existing tools in the literature (e.g., the delta method, influence functions) that we could use to provide inference for the mMSE gap. But such estimation-accuracy assumptions are only known to hold for a very limited class of regression estimators, and in particular preclude most modern machine learning algorithms and methods that integrate hard-to-quantify domain knowledge, which are exactly the types of powerful regression estimators we would most like to leverage for accurate inference.

However, given the centrality of $\mu^\star$ in the definition of the mMSE gap, it seems we need to at least implicitly estimate it with some working regression function μ. And even if we avoid assumptions on μ's accuracy, if we want to provide rigorous inference then we ultimately still need *some* way to relate μ to $\mathcal{I}$, which is a function of $\mu^\star$. We address this issue in the context of constructing a lower confidence bound (LCB) for the mMSE gap. The key idea proposed in this paper is to use a functional, which we call a *floodgate*, to relate *any* μ to $\mathcal{I}$. In particular, we will shortly introduce a $f(\mu)$ such that for *any* μ,

(a) $f(\mu) \leq \mathcal{I}$

(b) we can construct a lower confidence bound L for $f(\mu)$.

Then by construction L will also constitute a valid LCB for $\mathcal{I}$. The term *floodgate* comes from metaphorically thinking of constructing a LCB as preventing flooding ($L > \mathcal{I}$, i.e., miscoverage) by keeping the water level (L) below a critical threshold ($\mathcal{I}$) under arbitrary weather conditions (μ, or more specifically, μ's error, which we may not expect to be able to control well). Then by controlling L below $\mathcal{I}$ for any μ, f acts as a floodgate, and we also use the same name for the inference procedure we derive from f.

In particular, for any (nonrandom) function $\mu : \mathbb{R}^p \to \mathbb{R}$, define

$$f(\mu) := \frac{\mathbb{E}\left[\mathrm{Cov}(\mu^\star(X,Z), \mu(X,Z) \mid Z)\right]}{\sqrt{\mathbb{E}\left[\mathrm{Var}(\mu(X,Z) \mid Z)\right]}}, \tag{2.6}$$

where by convention we define $0/0 = 0$ so that $f(\mu)$ remains well-defined when the denominator of (2.6) is zero. It is not hard to see that f tightly satisfies the lower-bounding property (a) and we formalize this in the following lemma which is proved in Appendix A.1.1.

Lemma 2.2. *For any μ such that $f(\mu)$ exists, $f(\mu) \leq \mathcal{I}$, with equality when $\mu = \mu^\star$.*

In order to establish property (b) of f, we first take a *model-X* approach (Janson, 2017; Candès et al., 2018): we assume we know $P_{X|Z}$ but avoid assumptions on $Y \mid X, Z$. We start with such a model-X assumption because its simplicity helps elucidate the key ideas underlying the floodgate method, but floodgate is not tied to such assumptions, and indeed we present alternative versions of floodgate that operate under different assumptions later in the paper (Section 3.2's version somewhat relaxes the assumed knowledge of $P_{X|Z}$ without requiring any new assumptions and Remark 2.3.1's version relies on a *double-robust* set of assumptions). That said, the model-X assumption is sometimes reasonable and has been used before in a number of applications (see Appendix D for elaboration and examples), including in genomics like in the application presented in Section 5, and we theoretically (Appendix E) and numerically (Section 4.5) characterize model-X floodgate's robustness to misspecification of $P_{X|Z}$. Knowing $P_{X|Z}$ and μ means that, given data $\{(X_i, Z_i, Y_i)\}_{i=1}^n$, we also know $\{V_i := \mathrm{Var}(\mu(X_i, Z_i) \mid Z_i)\}_{i=1}^n$ which are i.i.d. and unbiased for the squared denominator in (2.6). And if we rewrite the numerator as

$$\mathbb{E}\left[\mathrm{Cov}(\mu^\star(X,Z), \mu(X,Z) \mid Z)\right] = \mathbb{E}\left[Y\big(\mu(X,Z) - \mathbb{E}\left[\mu(X,Z) \mid Z\right]\big)\right], \tag{2.7}$$

then we see we also know $\{R_i := Y_i\big(\mu(X_i, Z_i) - \mathbb{E}\left[\mu(X, Z_i) \mid Z_i\right]\big)\}_{i=1}^n$ which are i.i.d. and unbiased for the numerator. Thus for any given μ, we can use sample means of R_i and V_i to asymptotically-normally estimate both expectations in Equation (2.6), and then combine said estimators through the delta method to get an estimator of $f(\mu)$ whose asymptotic normality facilitates an immediate asymptotic LCB. This strategy is spelled out in Algorithm 1 and Theorem 2.3 establishes its asymptotic coverage. We pause to mention a simple but important point: when $\mu(X, Z)$ does not depend on X at all, then $f(\mu) = 0$ and all the V_i and R_i are zero with probability 1, making floodgate's LCB computed in Algorithm 1 deterministically zero as well. This implies that when the regression algorithm for obtaining μ is sparse, in the sense that it only depends on a fraction of its inputs, then floodgate will produce LCBs of zero for many of the covariates. For those covariates, coverage will hold *deterministically*, and hence floodgate will have average coverage even higher than the nominal $1 - \alpha$, as observed in some simulations in Section 4.

Algorithm 1 Floodgate

Input: Data $\{(Y_i, X_i, Z_i)\}_{i=1}^n$, $P_{X|Z}$, a working regression function $\mu : \mathbb{R}^p \to \mathbb{R}$, and a confidence level $\alpha \in (0, 1)$.

Compute $R_i = Y_i\big(\mu(X_i, Z_i) - \mathbb{E}\left[\mu(X_i, Z_i) \mid Z_i\right]\big)$ and $V_i = \mathrm{Var}\left(\mu(X_i, Z_i) \mid Z_i\right)$ for each $i \in [n]$, and their sample mean $(\bar{R}, \bar{V})$ and sample covariance matrix $\hat{\Sigma}$, and compute $s^2 = \frac{1}{\bar{V}}\left[\left(\frac{\bar{R}}{2\bar{V}}\right)^2 \hat{\Sigma}_{22} + \hat{\Sigma}_{11} - \frac{\bar{R}}{\bar{V}}\hat{\Sigma}_{12}\right]$.

Output: Lower confidence bound $L_n^\alpha(\mu) = \max\left\{\frac{\bar{R}}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n}}, 0\right\}$, with the convention that $0/0 = 0$.

Theorem 2.3 (Floodgate validity)**.** *For any given working regression function $\mu : \mathbb{R}^p \to \mathbb{R}$ and i.i.d. data $\{(Y_i, X_i, Z_i)\}_{i=1}^n$, if $\mathbb{E}[Y^4]$, $\mathbb{E}[\mu^4(X, Z)] < \infty$, then $L_n^\alpha(\mu)$ from Algorithm 1 satisfies*

$$\liminf_{n\to\infty} \mathbb{P}\left(L_n^\alpha(\mu) \le \mathcal{I}\right) \ge 1 - \alpha.$$

The proof of Theorem 2.3 can be found in Appendix A.1.2. Fourth moments (as opposed to the usual second moments for the CLT) are required because the estimand itself involves the expectations of $Y\mu(X, Z)$ and $\mu^2(X, Z)$. With higher moment conditions, we can apply relatively recent Berry–Esseen-type results for the delta method (Pinelis et al., 2016) to strengthen the pointwise asymptotic coverage of Theorem 2.3 to have a rate of $n^{-1/2}$; see Appendix C for details. We note that in both Algorithm 1 and Theorem 2.3, Y can be everywhere replaced by $Y - g_0(Z)$ for any non-random function g_0 (e.g., $\mathbb{E}\left[\mu(X, Z) \mid Z = z\right]$ would be a natural choice), which can reduce the variance of the R_i terms and hence improve the LCB.

Remark 2.3.1 (Doubly robust floodgate)**.** *Although for ease of exposition we have presented Algorithm 1 and Theorem 2.3 under the model-X assumption that $P_{X|Z}$ is known exactly, we emphasize here that the underlying idea of floodgate is* not *tied to this assumption. To reiterate, the key conceptual contribution of this paper is to introduce a lower-bounding functional $f(\mu)$ for $\mathcal{I}$ such that $f(\mu)$ provides a tractable statistical target to obtain a LCB for. To underscore this point, we present here a version of floodgate following the same principle but that is valid under standard double-robust assumptions instead of the aforementioned model-X assumption. Consider the following functional that depends not only on a working* *regression function $\mu(x,z)$, but also some Q_y estimating the true $P_{Y|Z}$ and some Q_x estimating the true $P_{X|Z}$:*

$$f_{Q_y,Q_x}(\mu) := \frac{\mathbb{E}\left[(Y - \mathbb{E}_{Q_y}[Y \mid Z])(\mu(X,Z) - \mathbb{E}_{Q_x}[\mu(X,Z) \mid Z])\right]}{\sqrt{\mathbb{E}\left[(\mu(X,Z) - \mathbb{E}_{Q_x}[\mu(X,Z) \mid Z])^2\right]}}, \tag{2.8}$$

where $\mathbb{E}_{Q_x}$ (resp. $\mathbb{E}_{Q_y}$) denotes expectation with respect to Q_x (resp. Q_y) as opposed to the true data-generating distribution, and by convention we again define $0/0 = 0$. Given Q_y, Q_x, and μ, i.i.d. unbiased estimates analogous to R_i and V_i in Algorithm 1 of the numerator and squared denominator, respectively, of $f_{Q_y,Q_x}(\mu)$ can be computed from each data point under no assumptions whatsoever, thus allowing the exact same kind of LCB as in Algorithm 1 to be computed for $f_{Q_y,Q_x}(\mu)$. It now just remains to check that $f_{Q_y,Q_x}(\mu)$ lower-bounds $\mathcal{I}$.

Lemma 2.3. *For any μ, Q_y, Q_x such that Q_x is absolutely continuous with respect to $P_{X|Z}$ and $f_{Q_y,Q_x}(\mu)$ exists, we have that $f_{Q_y,Q_x}(\mu) \le \mathcal{I} + \Delta$, where*

$$\Delta = \sqrt{\mathbb{E}\left[(\mathbb{E}[Y \mid Z] - \mathbb{E}_{Q_y}[Y \mid Z])^2\right]\mathbb{E}\left[w_\mu(X,Z)\chi^2\left(Q_x \| P_{X|Z}\right)\right]} \tag{2.9}$$

and $w_\mu(X,Z) = \frac{(\mu(X,Z) - \mathbb{E}[\mu(X,Z) \mid Z])^2}{\mathbb{E}[(\mu(X,Z) - \mathbb{E}[\mu(X,Z) \mid Z])^2]}$ is non-negative, has mean 1, and does not depend on Q_y or Q_x, and we again define $0/0 = 0$. Furthermore, $f_{Q_y,P_{X|Z}}(\mu) = f(\mu)$ and thus $f_{Q_y,P_{X|Z}}(\mu^\star) = \mathcal{I}$ (for any Q_y).

The proof can be found in Appendix A.1.3. Lemma 2.3 says that $f_{Q_y,Q_x}(\mu)$ only fails to lower-bound $\mathcal{I}$ to an extent bounded by the square root of the product of two terms: the MSE of $\mathbb{E}_{Q_y}[Y \mid Z]$ and the weighted χ^2 error of Q_x. The same result also holds if we move $w_\mu(X,Z)$ in Equation (2.9) *from the second term to the first term; see Equation* (A.33)*. As the first term measures the error in modeling $Y \mid Z$ and the second term measures the error in modeling $X \mid Z$, the square root of their product Δ is exactly what we would expect to be bounded as $o(n^{-1/2})$ under standard double-robustness assumptions (see, e.g., Chernozhukov et al. (2018b)). And indeed, since the LCB for $f_{Q_y,Q_x}(\mu)$ will be $\Omega(n^{-1/2})$ below $f_{Q_y,Q_x}(\mu)$, $\Delta = o(n^{-1/2})$ implies asymptotic coverage exactly as in Theorem 2.3.*

Remark 2.3.2 (Floodgate's validity in high dimensions)**.** *Again for ease of exposition, Theorem 2.3 establishes floodgate's pointwise asymptotic coverage for a fixed μ and a fixed (and hence fixed-dimensional) distribution for (Y,X,Z). It is certainly of interest to also consider the high-dimensional regime where the data-generating distribution (including the covariate dimension p) and the working regression function μ both depend on n, but it turns out that this setting is actually not very different from the simpler setting of Theorem 2.3. To see this, first note that Theorem 2.3 relies only on Lemma 2.2 ($f(\mu) \le \mathcal{I}$) and a central limit theorem (CLT) applied to the 2-dimensional mean of the i.i.d. pairs (R_i, V_i). But Lemma 2.2 is non-asymptotic, and hence $f(\mu) \le \mathcal{I}$ still holds even if μ varies with n. And the pairs (R_i, V_i) remain i.i.d. and 2-dimensional even as μ and the distribution of (Y,X,Z) vary with n, so all that is needed for floodgate's validity is a 2-dimensional i.i.d. triangular array CLT, which only requires that the 2-dimensional random variables (R_i, V_i) remain "well-behaved". In Appendix C we show in fact an even stronger (non-asymptotic) result, which, similarly to Theorem 2.3, only requires certain moments of Y and $\mu(X,Z)$ to remain bounded (although the result in Appendix C requires a bound on higher moments than Theorem 2.3 so that recent Berry–Eseen-type results for the delta method can be applied to bound floodgate's undercoverage at a rate of $n^{-1/2}$). In fact, it is even sufficient to replace the bound on $\mu(X,Z)$'s absolute moment with a bound on that of its conditional residual $h(X,Z) := \mu(X,Z) - \mathbb{E}[\mu(X,Z) \mid Z]$. Note that h only really measures the contribution from the single covariate X to the whole working regression function μ, even when Z is high-dimensional. Hence, we believe that assuming that Y's and $h(X,Z)$'s moments do not explode, even in high dimensions (recall Y and $h(X,Z)$ remain 1-dimensional regardless of the dimension of the data), seems quite mild in practice. For instance, if $|Y|$ is a bounded random variable (as it often will be in practice), then as long as μ is winsorized at some level (which, as long as the level is at least as large as $|Y|$'s bound, can only improve μ's performance) (Rinaldo et al., 2019), then floodgate's asymptotic validity* *is automatically ensured in the most general high-dimensional regime. Even when Y is unbounded, we would usually not expect the moments of Y or $h(X, Z)$ to diverge. Indeed in Section 4.3 we conduct high-dimensional simulations with unbounded Y and μ fitted via various parametric and nonparametric machine learning algorithms, yet floodgate's coverage remains empirically valid regardless of the dimension.*

Remark 2.3.3 (Choosing μ)**.** *The final missing piece in our LCB procedure is the choice of μ. In terms of how to obtain a working regression function μ, the flexibility of our procedure thus far finally pays off: μ can be chosen in* any *way that does not depend on the data used for inference. Normally we expect this to be achieved through data-splitting, i.e., a set of data samples is divided into two independent parts, and one part is used to produce an estimate μ of $\mu^\star$ while floodgate is applied to the other part with input μ; we will explore this strategy in simulations in Section 4. But in general, μ can be derived from any independent source, including mechanistic models or data of a completely different type than that used in floodgate (see, for example, Bates et al. (2020) for an example of using a regression model fitted to a separate data set in the context of variable selection). The goal is to allow the user as much latitude as possible in choosing μ so that they can leverage every tool at their disposal, including modern machine learning algorithms and qualitative domain knowledge, to get as close to $\mu^\star$ as possible. We show in Section 2.5 that there is a direct relationship between the accuracy of μ and the accuracy of the resulting floodgate LCB.*

In fact, an interesting and surprising feature of floodgate (both f and Algorithm 1) is that it is invariant to certain transformations of μ, making floodgate work well even sometimes when μ is quite far from $\mu^\star$. In particular, everything about floodgate remains identical if μ is replaced by any member of the set $S_\mu = \{c\mu(\cdot,\cdot) + g(\cdot,\cdot) : c > 0,\ g(x,\cdot) = g(x',\cdot)\,\forall x, x'\}$. An immediate consequence is that if μ is a partially linear function in x, i.e., $\mu(x,z) = cx + g(z)$ for some c and g, then floodgate only depends on μ through the sign of c, making floodgate particularly forgiving for partially linear working models. To be precise, floodgate using $\mu(x,z) = cx + g(z)$ will perform identically *to floodgate using the* best *partially linear approximation to $\mu^\star$ as long as c has the same sign as the coefficient in that best approximation (regardless of c's magnitude or anything about g).*

2.3 Upper confidence bounds for the mMSE gap

Before continuing our study of floodgate LCBs, we first pause to address a natural question: what about an *upper* confidence bound (UCB)? One way to get a UCB is to follow a workflow similar to the previous subsection, as follows. For any working regression function ν for $\mathbb{E}[Y \mid Z]$, consider the functional

$$f^{\text{UCB}}(\nu) = \mathbb{E}\left[(Y - \nu(Z))^2\right].$$

Then f^{UCB} plays an analogous role to f in the opposite direction, in that for *any* ν, (a) $f^{\text{UCB}}(\nu) \geq \mathcal{I}^2$ and (b) we can construct a level α UCB $U_n^\alpha(\nu)$ for $f^{\text{UCB}}(\nu)$. Property (a) is immediate from the minimality of the first term and non-negativity of the second term in definition (2.1), while property (b) can be established without even making model-X assumptions: simply take the CLT-based UCB from the estimator $\frac{1}{n}\sum_{i=1}^n (Y_i - \nu(Z_i))^2$, which is unbiased for $f^{\text{UCB}}(\nu)$.

Unfortunately, there is no value of ν such that $f^{\text{UCB}}(\nu) = \mathcal{I}^2$ except in the noiseless setting where Y is a *deterministic* function of (X, Z). In particular, no matter how well ν is chosen and how large n is, $U_n^\alpha(\nu) - \mathcal{I}^2 \geq \mathbb{E}[\text{Var}(Y \mid X, Z)]$ with probability at least $1-\alpha$. This shortcoming is perhaps foreseeable given that $U_n^\alpha(\nu)$ never even uses the X_i, but it turns out to be unimprovable (even using model-X information), as we now prove in Theorem 2.4.

Theorem 2.4. *Fix a continuous joint distribution $P_{X,Z}$ for (X, Z), and let $\mathcal{F}$ denote the class of joint distributions F for (Y, X, Z) such that F is compatible with $P_{X,Z}$ and* $\text{Var}(Y) < \infty$*. Let $U(D_n)$ denote a scalar-valued function of the n i.i.d. samples $D_n = \{Y_i, X_i, Z_i\}_{i=1}^n$; if $U(D_n)$ outputs a UCB for the mMSE gap that is pointwise asymptotically valid for any $F \in \mathcal{F}$, i.e.,*

$$\inf_{F\in\mathcal{F}} \liminf_{n\to\infty} \mathbb{P}_F(U(D_n) \geq \mathcal{I}_F^2) \geq 1-\alpha,$$

then

$$\sup_{F\in\mathcal{F}} \limsup_{n\to\infty} \mathbb{P}_F\left(U(D_n) - \mathcal{I}_F^2 < \mathbb{E}_F\left[\mathrm{Var}_F\left(Y \mid X, Z\right)\right]\right) \leq \alpha, \tag{2.10}$$

where the subscript F denotes quantities computed with F as the data-generating distribution.

The proof of Theorem 2.4 can be found in Appendix A.2. Note that since we fix $P_{X,Z}$ at the beginning of the theorem statement, U is allowed to use model-X information. As just mentioned above, this theorem provides no cause for concern in the noiseless setting when $\mathbb{E}\left[\mathrm{Var}\left(Y \mid X, Z\right)\right] = 0$. However, in many applications we may expect $\mathbb{E}\left[\mathrm{Var}\left(Y \mid X, Z\right)\right]$ to be substantial, and the above theorem guarantees *any* pointwise asymptotically valid UCB must be conservative by this amount. The only way to overcome this problem would be to assume some sort of structure on $Y \mid X, Z$, such as smoothness or sparsity, in contrast to model-X floodgate which requires no information about $Y \mid X, Z$ and can certainly produce nontrivial LCBs and even achieve the parametric rate with sufficiently-accurate μ; see Section 2.5. Although it is disappointing that a better UCB is not achievable, we envision MOVI inference often being used to quantify *new* important relationships, in which case we expect it to be more useful to know a variable is *at least as* important as some LCB than to upper-bound its importance with a UCB. Given this perspective and the negative UCB result of Theorem 2.4, we return for the remainder of the paper to the study of using floodgate to obtain LCBs.

2.4 Computation

Astute readers may have noticed that the quantities R_i and V_i in Algorithm 1 involve conditional expectations/variances which, though in principle known due to the assumed model-X knowledge of $P_{X|Z}$, may be quite hard to compute in practice. In certain cases these conditional expectations can have simple or even closed-form expressions, such as when μ is a generalized linear model and $X \mid Z$ is Gaussian, but otherwise a more general approach is needed. Monte Carlo provides a natural solution: assume that we can sample K copies $\tilde{X}_i^{(k)}$ of X_i from $P_{X_i|Z_i}$ conditionally independently of X_i and Y_i and thus replace R_i and V_i, respectively, by the sample estimators

$$R_i^K = Y_i\left(\mu(X_i, Z_i) - \frac{1}{K}\sum_{k=1}^{K}\mu(\tilde{X}_i^{(k)}, Z_i)\right),$$

$$V_i^K = \frac{1}{K-1}\sum_{k=1}^{K}\left(\mu(\tilde{X}_i^{(k)}, Z_i) - \frac{1}{K}\sum_{k=1}^{K}\mu(\tilde{X}_i^{(k)}, Z_i)\right)^2.$$

Luckily the same guarantees hold for the Monte Carlo analogue of floodgate, even for fixed K.

Theorem 2.5. *Under the conditions of Theorem 2.3, for any given $K > 1$, $L_{n,K}^{\alpha}(\mu)$ computed by replacing R_i and V_i with R_i^K and V_i^K, respectively, in Algorithm 1 satisfies*

$$\liminf_{n\to\infty} \mathbb{P}\left(L_{n,K}^{\alpha}(\mu) \leq \mathcal{I}\right) \geq 1 - \alpha.$$

The proof can be found in Appendix A.3. In general we expect larger values of K to produce more accurate LCBs, but we found the difference between $K = 2$ and $K = \infty$ to be surprisingly small in our simulations and, of course, it will always be computationally faster to use smaller K. Although Theorem 2.5 is a pointwise result holding for any fixed $K > 1$, it can be generalized to a uniform result over all $K > 1$ with miscoverage bounded by a $n^{-1/2}$ rate using higher moment conditions and a variance lower bound assumption; see Appendix C for details.

2.5 Accuracy adaptivity to μ's mean squared error

Having established floodgate's validity and computational tractability, the natural next question is: how accurate is it, i.e., how close is the LCB to the mMSE gap? The answer depends on the accuracy of μ—the better that μ approximates $\mu^\star$, the more accurate the floodgate LCB is, as formalized in the following theorem.

Theorem 2.6 (Floodgate accuracy and adaptivity)**.** *For i.i.d. data* $\{(Y_i, X_i, Z_i)\}_{i=1}^n$ *such that* $\mathbb{E}[Y^{12}] < \infty$, $\mathrm{Var}\,(Y \mid X, Z) \geq \tau$ *a.s. for some* $\tau > 0$*, and a sequence of working regression functions* $\mu_n : \mathbb{R}^p \to \mathbb{R}$ *such that for some* C *and all* n *either* $\mathbb{E}\,[\mathrm{Var}\,(\mu_n(X,Z) \mid Z)] = 0$ *or* $\frac{\mathbb{E}[\mu_n^{12}(X,Z)]}{\mathbb{E}[\mathrm{Var}(\mu_n(X,Z) \mid Z)]^6} \leq C$*, the output of Algorithm 1 satisfies*

$$\mathcal{I} - L_n^\alpha(\mu_n) = O_p\left(\inf_{\mu \in S_{\mu_n}} \mathbb{E}\left[(\mu(X,Z) - \mu^\star(X,Z))^2\right] + n^{-1/2}\right), \tag{2.11}$$

where $S_{\mu_n} = \{c\mu_n(\cdot,\cdot) + g(\cdot,\cdot) : c > 0,\ g(x,\cdot) = g(x',\cdot)\,\forall x, x'\}$ *as defined in Remark 2.3.3.*

The proof can be found in Appendix A.4. The above condition that "$\mathbb{E}\,[\mathrm{Var}\,(\mu_n(X,Z) \mid Z)] = 0$ or $\frac{\mathbb{E}[\mu_n^{12}(X,Z)]}{\mathbb{E}[\mathrm{Var}(\mu_n(X,Z) \mid Z)]^6} \leq C$" is a scale-free moment condition on μ_n which says that $\mu_n(X,Z)$ can have no dependence on Z at all or have a non-vanishing conditional variance (given Z) relative to its higher moments. The high-order moments in our assumptions are likely a technical artifact of our proof (which actually proves a somewhat stronger result than stated in the theorem), and could perhaps be relaxed with a different approach. As it stands, these assumptions allow us to utilize the Berry–Esseen-type results in Appendix C.1.1 to handle the fact that μ_n varies with n.

We call the left-hand side of Equation (2.11) the *half-width* (by analogy with the *width* that would measure the accuracy of a two-sided confidence interval) and Theorem 2.6 shows it is *adaptive* to the accuracy of μ_n through the MSE of the best element of its equivalence class S_{μ_n}, up to a limit of the parametric or central limit theorem rate of $n^{-1/2}$. So in principle floodgate can achieve $n^{-1/2}$ accuracy if a member of S_{μ_n} converges very quickly to $\mu^\star$, but in general floodgate's accuracy decays gracefully with μ_n's accuracy. Note that the infimum in Equation (2.11) means that floodgate is *self-correcting* with respect to μ_n's conditional mean given Z, as explained in the second paragraph of Remark 2.3.3.

2.6 Straightforward generalizations

Before moving onto extensions, we briefly address a few relatively straightforward generalizations of floodgate.

Extending the mMSE gap The mMSE gap can be very naturally made invariant to the scale of Y and bounded between 0 and 1 by dividing it by $\mathrm{Var}\,(Y)$. And since $\mathrm{Var}\,(Y)$ can be easily and asymptotically-normally estimated under weaker conditions than already assumed for floodgate's validity in Theorem 2.5, it is straightforward to extend the floodgate procedure and its validity to perform inference on the scale-free version $\mathcal{I}_{\mathrm{sf}}^2 = \mathcal{I}^2/\mathrm{Var}\,(Y)$. We also consider two ways of extending the mMSE gap such that the key property of the MOVI in Azadkia and Chatterjee (2019) is satisfied, i.e., the MOVI equals zero if *and only if* $Y \perp\!\!\!\perp X \mid Z$. Details about defining the MOVIs and providing inference can be found in Appendices F.1 and F.2.

Inference for group variable importance In applications where a group of variables share a common interpretation or are too correlated to powerfully distinguish, it is often necessary to infer a measure of *group* importance instead of a MOVI. Luckily, when X is multivariate, the mMSE gap remains perfectly well-defined and interpretable and floodgate (both f and Algorithm 1) retain all the same inferential properties. Indeed, we apply floodgate to groups of variables in our genomics application in Section 5.

Transporting inference to other covariate distributions In some applications, the samples we collect may not be uniformly drawn from the population we are interested in studying. For instance, our data may come from a lab experiment with covariates randomized according to one distribution, while our interest lies in inference about a population outside the lab whose covariates follow a different distribution. As long as the samples at hand share a common conditional distribution $Y \mid X, Z$ with the target population, it is relatively straightforward to perform an importance-weighted version of floodgate that provides inference for the target population's mMSE gap. We provide the details in Appendix G.

Adjusting for selection When inference is required for many variables simultaneously, it is often preferable to focus attention on a subset of variables whose inferences appear particularly interesting. But if we only report the set of LCBs that are, say, farthest from zero, then our coverage guarantees will fail to hold for this set due to selection bias (this is not a defect of floodgate, but a property of nearly every non-selective inferential procedure). One way to address this may be to apply false coverage-statement rate adjustments (Benjamini and Yekutieli, 2005) to floodgate LCBs. The application is straightforward, and floodgate LCBs satisfy the monotone property required by Benjamini and Yekutieli (2005), although they do not in general satisfy the independence or positive regression dependence on a subset (PRDS) condition and hence would require a correction (Benjamini and Yekutieli, 2001) for strict guarantees to hold. We leave a more formal treatment of selection adjustment to future work, but note also some simple ways to perform benign selection.

First, if selection is performed using μ and/or independent data, then no adjustment is needed for validity. For instance, if floodgate is run by data-splitting, we could arbitrarily use the first half of the data (which is also used for choosing μ, but not for running floodgate) for selection, including selecting precisely the subset of variables that μ depends on. In fact, we can even perform a certain type of benign post-hoc data processing based on the floodgate data itself: if the floodgate data are used to construct a *transformation* of the floodgate LCBs such that every transformed LCB either shrinks or remains the same, then the transformed LCBs retain their marginal asymptotic validity. This is because any such transformation, even one depending on the data or LCBs themselves, can only *increase* coverage of each LCB by reducing it or leaving it unchanged; this is related to the screening procedure in Liu et al. (2021). This means, for instance, that if a selection procedure is applied to the floodgate data and used to zero out any unselected LCBs, then as long as the zeroed-out LCBs are reported alongside the rest, the marginal validity of all reported LCBs remains intact even though the same data was used to construct the LCBs and to perform the selection that transformed them.

3 Extensions

3.1 Beyond the mMSE gap

To demonstrate that the floodgate idea can be used beyond the mMSE gap, we consider the following MOVI.

Definition 3.1 (Mean absolute conditional mean gap)**.** *The* mean absolute conditional mean (MACM) gap *for variable X is defined as*

$$\mathcal{I}_{\ell_1} = \mathbb{E}\left[\left|\mathbb{E}\left[Y \mid Z\right] - \mathbb{E}\left[Y \mid X, Z\right]\right|\right] \tag{3.1}$$

whenever all the above expectations exist.

The subscript in $\mathcal{I}_{\ell_1}$ reflects its similarity to $\mathcal{I}^2 = \mathbb{E}\left[(\mathbb{E}\left[Y \mid Z\right] - \mathbb{E}\left[Y \mid X, Z\right])^2\right]$ except with the square replaced by the absolute value (also known as the ℓ_1 norm). Although we have not found a floodgate function to enable inference for arbitrary Y, the remainder of this subsection shows how to perform floodgate inference when Y is binary (coded as $Y \in \{-1, 1\}$). We note that when Y is binary, $\mathcal{I}_{\ell_1}$ is zero if and *only if* $Y \perp\!\!\!\perp X \mid Z$ holds (the "if" part holds for non-binary Y as well), since the expected value uniquely determines the distribution of a binary random variable.

In particular, for any (nonrandom) function $\mu : \mathbb{R}^p \to \mathbb{R}$, define

$$f_{\ell_1}(\mu) = 2\mathbb{P}\big(Y(\mu(\tilde{X}, Z) - \mathbb{E}[\mu(X,Z) \mid Z]) < 0\big) - 2\mathbb{P}\big(Y(\mu(X,Z) - \mathbb{E}[\mu(X,Z) \mid Z]) < 0\big) \tag{3.2}$$

where $\tilde{X} \sim P_{X|Z}$ and is conditionally independent of X and Y.

Lemma 3.2. *If $|Y| \overset{a.s.}{=} 1$, then for any μ such that $f_{\ell_1}(\mu)$ exists, $f_{\ell_1}(\mu) \le \mathcal{I}_{\ell_1}$, with equality when $\mu = \mu^\star$.*

Obtaining an LCB for $f_{\ell_1}(\mu)$ is even easier than it was for $f(\mu)$ because $f_{\ell_1}(\mu)$ is essentially just one expectation instead of a ratio of expectations, so a straightforward central limit theorem argument suffices; Algorithm 3 (presented in Appendix H) formalizes the procedure and Theorem 3.3 establishes its asymptotic coverage.

Theorem 3.3 (MACM gap floodgate validity)**.** *For any given working regression function $\mu : \mathbb{R}^p \to \mathbb{R}$ and i.i.d. data $\{(Y_i, X_i, Z_i)\}_{i=1}^n$, $L_n^\alpha(\mu)$ from Algorithm 3 satisfies*

$$\mathbb{P}\left(L_n^\alpha(\mu) \le \mathcal{I}_{\ell_1}\right) \ge 1 - \alpha - O(n^{-1/2}).$$

Theorem 3.3 is proved in Appendix A.5, and perhaps its most striking feature is its lack of assumptions, which follows from the boundedness of $f_{\ell_1}(\mu)$ and the R_i. Like f, f_{ℓ_1} is invariant to any transformation of μ that leaves $\operatorname{sign}(\mu(X,Z) - \mathbb{E}[\mu(X,Z) \mid Z])$ unchanged on a set of probability 1, making its validity immediately uniform over large classes of μ.

Although the boundedness of the R_i streamlines the coverage guarantees, their conditional probabilities make it somewhat more complicated to carry out efficient computation of Algorithm 3. In particular, the sharp boundary at zero inside the probabilities requires a certain degree of smoothness in μ and P to be able to estimate the R_i by Monte Carlo samples analogously to Section 2.4. We give precise sufficient conditions and a proof of their validity in Appendix H, and defer study of Algorithm 3's accuracy and robustness to future work.

3.2 Relaxing the assumptions by conditioning

In this section we show that we can relax the model-X assumption that $P_{X|Z}$ be known exactly and apply floodgate when only a *parametric model* is known for $P_{X|Z}$. This is inspired by Huang and Janson (2020) which similarly relaxes the assumptions of model-X knockoffs. We follow the same general principle of conditioning on a sufficient statistic of the parametric model for $P_{X|Z}$, but doing so in floodgate requires a somewhat different approach than Huang and Janson (2020). Note that this section's method and assumptions are also distinct from the double robust assumptions in Remark 2.3.1, further emphasizing that the key ideas underlying floodgate are not tied to any particular set of assumptions.

The approach we take in this section will involve computations on the entire matrix of observations, i.e., $(\boldsymbol{X}, \boldsymbol{Z}) \in \mathbb{R}^{n \times p}$ whose rows are the covariate samples (X_i, Z_i) and $\boldsymbol{y} \in \mathbb{R}^n$ whose entries are the response samples Y_i. Now suppose that we know a model $F_{X|Z}$ for $P_{X|Z}$ with a sufficient statistic functional for n independent (but not necessarily identically distributed) samples $\boldsymbol{X} \mid \boldsymbol{Z}$ given by $\mathcal{T}(\boldsymbol{X}, \boldsymbol{Z})$, whose random value we will denote simply by $\boldsymbol{T}$. We will assume that $\mathcal{T}$ is invariant to permutation of the rows of $(\boldsymbol{X}, \boldsymbol{Z})$ (as we would expect for any reasonable $\mathcal{T}$, since these rows are i.i.d.).

The key idea that allows us to perform floodgate inference without knowing the distribution of $\boldsymbol{X} \mid \boldsymbol{Z}$ is that, by definition of sufficiency, we *do* know the distribution of $\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{T}$. Leveraging this idea requires some adjustment to the floodgate procedure, and we start by defining a conditional analogue of f.

$$f_n^{\mathcal{T}}(\mu) := \frac{\mathbb{E}\left[\operatorname{Cov}(\mu^\star(X_i, Z_i), \mu(X_i, Z_i) \mid \boldsymbol{Z}, \boldsymbol{T})\right]}{\sqrt{\mathbb{E}\left[\operatorname{Var}(\mu(X_i, Z_i) \mid \boldsymbol{Z}, \boldsymbol{T})\right]}}, \tag{3.3}$$

again with the convention $0/0 = 0$. Note that $f_n^{\mathcal{T}}(\mu)$ does not depend on the choice of i thanks to $\mathcal{T}$'s permutation invariance, but it *does* depend on the sample size n. Nevertheless, it follows immediately from the proof of Lemma 2.2 that $f_n^{\mathcal{T}}(\mu) \leq f_n^{\mathcal{T}}(\mu^\star)$ for any nonrandom μ. On the other hand, $f_n^{\mathcal{T}}(\mu^\star) \neq \mathcal{I}$, but instead a different relationship that is nearly as useful holds:

$$f_n^{\mathcal{T}}(\mu^\star) \leq f(\mu^\star) = \mathcal{I},$$

due to the monotonicity of conditional variance.

With floodgate property (a) ($f_n^{\mathcal{T}}(\mu) \leq \mathcal{I}$) established, we now turn to property (b): the ability to construct a LCB for $f_n^{\mathcal{T}}(\mu)$. In an analogous way as for $f(\mu)$, we can compute n unbiased estimators of the numerator and the squared denominator, but these estimators are no longer i.i.d. because they are linked through $\boldsymbol{T}$, so we cannot immediately apply the central limit theorem or delta method as we did in Section 2.2. Our workaround is to split the data into n_2 *batches* of size n_1 and only condition on the sufficient statistic within each batch. This way, there is still independence between batches and we can apply the central limit theorem and delta method across batches. This strategy is spelled out in Algorithm 4 (see Appendix I for details) and Theorem 3.4 establishes its asymptotic coverage. We call this procedure *co-sufficient* floodgate because the term "co-sufficiency" describes sampling conditioned on a sufficient statistic (Stephens, 2012).

Theorem 3.4 (Co-sufficient floodgate validity)**.** *For any given working regression function $\mu : \mathbb{R}^p \to \mathbb{R}$, i.i.d. data $\{(X_i, Z_i, Y_i)\}_{i=1}^n$, and permutation-invariant sufficient statistic functional $\mathcal{T}$, if $\mathbb{E}[Y^4] < \infty$ and $\mathbb{E}[\mu^4(X,Z)] < \infty$, then $L_n^{\alpha,\mathcal{T}}(\mu)$ from Algorithm 4 satisfies*

$$\liminf_{n\to\infty} \mathbb{P}\left(L_n^{\alpha,\mathcal{T}}(\mu) \leq \mathcal{I}\right) \geq 1 - \alpha.$$

The proof can be found in Appendix A.6. Regarding computation, as in Section 2.4, we can replace the conditional expectations in Algorithm 4 with Monte Carlo estimates; see Appendix I.1 for details. For a given μ, we may worry that co-sufficient floodgate loses some accuracy relative to regular floodgate due to the gap between $f(\mu)$ and $f_n^{\mathcal{T}}(\mu)$, but in fact this gap is typically $O(n_2^{-1})$ for fixed-dimensional parametric models. We quantify this gap for multivariate Gaussian and discrete Markov chain covariate models in the following two subsections, showing that, at least in these two cases, co-sufficient floodgate relaxes the assumptions of model-X floodgate with only a minimal loss in accuracy.

3.2.1 Low-dimensional multivariate Gaussian model

In this section we let $\mathcal{B}_m = \{(m-1)n_2 + 1, \ldots, mn_2\}$.

Proposition 3.5. *Suppose samples $\{X, Z\}_{i=1}^n$ are i.i.d. multivariate Gaussian parameterized as $X_i \mid Z_i \sim \mathcal{N}\left((1, Z_i)\gamma, \sigma^2\right)$ for some $\gamma \in \mathbb{R}^p$ and $\sigma^2 > 0$, and $Z_i \sim \mathcal{N}(\boldsymbol{v}_0, \boldsymbol{\Sigma}_0)$. Assume σ^2 is known and the batch size n_2 satisfies $n_2 > p + 2$. Let $\mathcal{T}$ be the following sufficient statistic functional*

$$\boldsymbol{T}_m := \mathcal{T}(\boldsymbol{X}_m, \boldsymbol{Z}_m) = \left(\sum_{i\in\mathcal{B}_m} X_i, \sum_{i\in\mathcal{B}_m} X_i Z_i\right).$$

Then if $\mathbb{E}\left[\mu^4(X,Z)\right], \mathbb{E}\left[(\mu^\star)^4(X,Z)\right] < \infty$, we have

$$f(\mu) - f_n^{\mathcal{T}}(\mu) = O\left(\frac{p}{n_2 - p - 2}\right). \tag{3.4}$$

The proof can be found in Appendix I.2.1. Note the condition $n_2 > p + 2$ is not surprising as when the sample size is smaller than p, the sufficient statistic functional is degenerate, resulting in a zero value of $f_n^{\mathcal{T}}(\mu)$. The bound in (3.4) allows p to grow with n in general, but when p is fixed, it gives the rate of $O(n_2^{-1})$, as mentioned earlier in Section 3.2.

3.2.2 Discrete Markov chains

To present our second example model, we define some new notation. Consider a random variable W following a discrete Markov chain with K states with $X = W_j$, $Z = W_{\text{-}j}$, then the model parameters include the initial probability vector $\pi^{(1)} \in \mathbb{R}^K$ with $\pi^{(1)}_k = \mathbb{P}(W_1 = k)$ and the transition probability matrix $\Pi^{(j)} \in \mathbb{R}^{K\times K}$ (between W_{j-1} and $X = W_j$) with $\Pi^{(j)}_{k,k'} = \mathbb{P}(W_j = k' \mid W_{j-1} = k)$. Further denoting $q(k, k_1, k_2) = \mathbb{P}(W_j = k \mid W_{j-1} = k_1, W_{j+1} = k_2)$, we have

$$q(k, k_1, k_2) = \frac{\Pi^{(j)}_{k_1,k}\Pi^{(j+1)}_{k,k_2}}{\sum_{k=1}^{K} \Pi^{(j)}_{k_1,k}\Pi^{(j+1)}_{k,k_2}},$$

so that the conditional distribution of $\boldsymbol{X}_m \mid \boldsymbol{Z}_m$ can be compactly written down as

$$\mathbb{P}(\boldsymbol{X}_m \mid \boldsymbol{Z}_m) = \prod_{k,k_1,k_2\in[K]} (q(k, k_1, k_2))^{N(k,k_1,k_2)}, \tag{3.5}$$

where $N(k, k_1, k_2) = \sum_{i\in\mathcal{B}_m} \mathbb{1}_{\{X_i=k,W_{i,j-1}=k_1,W_{i,j+1}=k_1\}}$. Thus we finally conclude that $\{N(k, k_1, k_2)\}_{(k,k_1,k_2\in[K])}$ is sufficient, and we proceed with this sufficient statistic.

Proposition 3.6. *Consider the above discrete Markov chain model and define the sufficient statistic functional $\mathcal{T}$ as*

$$\boldsymbol{T}_m = \mathcal{T}(\boldsymbol{X}_m, \boldsymbol{Z}_m) = \{N(k, k_1, k_2)\}_{(k,k_1,k_2\in[K])}.$$

Then if for variable $X = W_j$, $K^2 \min\{\mathbb{P}(W_{j-1} = k_1, W_{j+1} = k_2)\}_{k_1,k_2\in[K]}\} \geq q_0 > 0$ holds and $\mathbb{E}\left[(\mu^\star)^2(X, Z)\right]$, $\mathbb{E}\left[\mu^2(X, Z)\right] < \infty$, we have

$$f(\mu) - f^{\mathcal{T}}_n(\mu) = O\left(\frac{K^3}{n_2}\right).$$

The proof can be found in Appendix I.2.2. Note that $\mathcal{T}$ here is not minimal sufficient and the above rate is cubic in K. The non-minimal sufficient statistic is adopted for the discrete Markov chain model in this paper since it is easier to work with and gives the desired rate in n_2, but we expect the rate in K could be improved by using the minimal sufficient statistic. Again, K is allowed to grow with n in general, but when it is fixed we get a rate of $O(n_2^{-1})$, as mentioned earlier in Section 3.2.

4 Simulations

Source code for conducting our simulation studies can be found at `https://github.com/LuZhangH/floodgate`.

4.1 Setup

In the following subsections, we conduct simulation studies to complement the main theoretical claims of the paper. We study the effects of the sample-splitting proportion (Section 4.2), covariate dimension (Section 4.3), and model misspecification (Section 4.5) on floodgate. Additional simulation studies on the effect of covariate dependence and sample size can be found in Appendix J.4. In Section 4.4, we numerically compare floodgate with the method proposed in Williamson et al. (2020). We also study the extensions to floodgate for the MACM gap (Section 4.6) and co-sufficient floodgate (Section 4.7). Each simulation study generates a set of covariates and performs floodgate inference on each in turn (i.e., treating each covariate as X and the rest as Z) before averaging its results (either coverage or half-width) over the covariates.

This paragraph describes the simulation setup for all but the simulation of Section 4.4. The covariates are sampled from a Gaussian autoregressive model of order 1 (AR(1)) with autocorrelation 0.3, except in Section J.4.6 where this value is varied over. The conditional distribution of $Y \mid X, Z$ is given by $\mu^\star(X, Z)$

plus standard Gaussian noise, and in each subsection we perform experiments with both a linear and a highly nonlinear model. The linear model is sparse with non-zero coefficients' locations independently uniformly drawn from among the covariates, and the non-zero coefficients' values having uniform random signs and identical magnitudes (5, unless stated otherwise) divided by $\sqrt{n}$. The nonlinear model combines zero'th-, first-, and second-order interactions between nonlinear (mostly trigonometric and polynomial) transformations of elementwise functions of a subset of covariates, and then multiplies this entire function by an amplitude (50, unless stated otherwise) divided by $\sqrt{n}$; see Appendix J.1 for details. Both models use $n = 1100$, $p = 1000$, and a sparsity of 30 unless stated otherwise.

In our implementations of floodgate, we split the sample into two equal parts (justified by the results of Section 4.2) and use the first half to fit μ. In most of the simulations, we consider four fitting algorithms (two linear, two nonlinear): the LASSO (Tibshirani, 1996), Ridge regression, Sparse Additive Models (SAM; (Ravikumar et al., 2009)), and Random Forests (Breiman, 2001); when the response is binary there are two additional fitting algorithms: logistic regression with an L1 penalty and an L2 penalty; see Appendix J.2 for implementation details of these algorithms. The Monte Carlo version of floodgate from Section 2.4 is not needed for the linear methods, and for the nonlinear methods, $K = 500$ is used.

Given the novelty of considering inference for the mMSE gap, it is challenging to compare floodgate to alternatives except in special cases. For instance, in low-dimensional Gaussian linear models the mMSE gap is a simple function of the coefficient and thus ordinary least squares (OLS) inference can be compared to floodgate; see Appendix J.3 for details of how it is made comparable. Thus, in the low-dimensional linear-$\mu^\star$ simulations of Sections 4.3 and J.4.6, we compare floodgate's inference to that of OLS, which acts as a sort of oracle since its inference relies on very strong knowledge of $Y \mid X, Z$ which floodgate does not rely on, and OLS is not valid without that knowledge (and does not apply in high dimensions). Another example is when we can assume the group mMSE gap of all of (X, Z) is bounded away from zero, in which case the method of Williamson et al. (2020) applies, so in Section 4.4 we compare their method with floodgate in such a setting.

Remark 4.1 (Floodgate's connection to conditional independence testing)**.** *Recall that $Y \perp\!\!\!\perp X \mid Z$ implies $\mathcal{I} = 0$, and hence rejecting $Y \perp\!\!\!\perp X \mid Z$ when $L_n^\alpha(\mu) > 0$ constitutes an asymptotically valid level-α conditional independence test (which could then be combined with a multiple testing procedure to perform variable selection). However, floodgate was explicitly designed to solve the harder problem of quantifying strength of dependence, as opposed to the conditional independence problem of whether any dependence exists at all. Due to the methodological constraints imposed by the more challenging nature of our problem, especially the need for data splitting, we do not expect this test derived from floodgate to be competitive with (and hence do not compare with) the many excellent conditional independence tests available in the literature (see, e.g., Candès et al. (2018); Huang and Janson (2020); Berrett et al. (2020); Liu et al. (2021); Barber and Janson (2020); Tansey et al. (2022); Fukumizu et al. (2008); Zhang et al. (2011); Wang et al. (2015); Shah and Peters (2020); Park and Muandet (2020); Huang et al. (2020)).*

We always take the significance level $\alpha = 0.05$, and all results are averaged over 64 independent replicates unless stated otherwise (although in most cases each plotted point is averaged over multiple covariates per replicate as well, since we apply floodgate to each covariate in turn in each replicate).

4.2 Effect of sample splitting proportion

As mentioned in Section 2.2, we can split a fixed sample size n into a first part of size n_e for estimating $\mu^\star$ and use the remaining $n - n_e$ samples for floodgate inference via Algorithm 1. The choice of n_e represents a tradeoff between higher accuracy in estimating $\mu^\star$ (larger n_e) and having more samples available for inference (smaller n_e).

In Figure 1, we vary the sample splitting proportion and plot the average half-widths of floodgate LCBs of non-null covariates under distributions with the linear and the nonlinear $\mu^\star$ described in Section 4.1. Corresponding coverage plots and additional plots with different simulation parameters can be found in Appendix J.4. Our main takeaway from these plots is that, while the optimal choice of splitting proportion

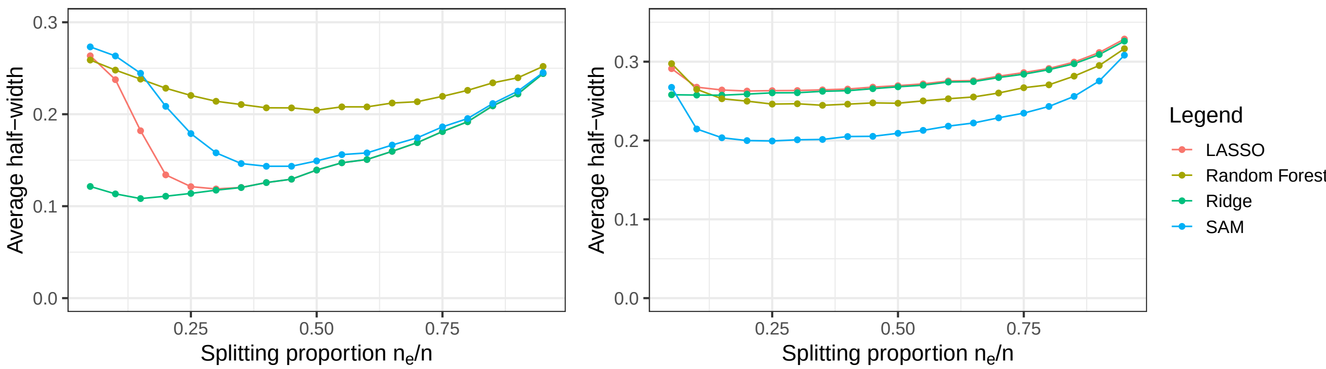

Figure 1: Average half-widths for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 4.2. The coefficient amplitude is chosen to be 10 for the left panel and the sample size n equals 3000 in the right panel; see Section 4.1 for remaining details. Standard errors are below 0.005 (left) and 0.006 (right).

varies between distributions and algorithms, the choice of 0.5 seems to frequently achieve a half-width close to the optimum. Acknowledging that in some circumstances a more informed choice than 0.5 can be made, we nevertheless choose 0.5 as the default splitting proportion throughout the rest of our simulations.

In addition to displaying the dynamics of sample splitting proportion, these plots also demonstrate two other phenomena. First, the linear algorithms (LASSO and Ridge) dominate when $\mu^\star$ is linear, and the nonlinear algorithms (SAM and Random Forest) dominate when $\mu^\star$ is nonlinear. Second, Ridge has smaller half-width than LASSO for all sample splitting proportions, which can be explained by floodgate's invariance to (partially-)linear μ: all that matters is getting the sign of the coefficient right, and setting a coefficient to zero guarantees a zero LCB. So the LASSO suffers from being a sparse estimator, although in practice we may still prefer it because of the corresponding computational savings of only having to run floodgate on a subset of covariates.

4.3 Effect of covariate dimension

To understand the dependence of dimension on floodgate, we perform simulations varying the dimension. In particular, in the first panel of Figure 2, we vary the covariate dimension and plot the average half-widths of floodgate LCBs of non-null covariates when $\mu^\star$ is linear. This setting enables comparison with OLS because it is linear and low-dimensional, so we also include a curve for OLS.

The main takeaway is that floodgate's accuracy is relatively unaffected by dimension, and although for very low dimensions (where OLS is known to be essentially optimal) it is less accurate than OLS, for a good choice of n_e floodgate's half-widths are at most about 50% larger than OLS's and actually narrower than OLS's when $p \approx n/2$. A similar message is found with nonlinear $\mu^\star$ in the second panel of Figure 2, except OLS no longer applies and in this case the nonlinear algorithms outperform the linear ones in floodgate. Coverage plots corresponding to Figure 2 and additional plots with different simulation parameters can be found in Appendix J.4.

4.4 Comparison with Williamson et al. (2020)

Although Williamson et al. (2020)'s method (which we refer to as W20b) is only valid when the group mMSE gap of all the covariates is bounded away from zero, we can compare it with floodgate in that setting. We use W20b according to that paper's instructions for ensuring validity for any value of $\mathcal{I}$ (as long as the group mMSE gap for all the variables put together is bounded away from zero), which seems most comparable to floodgate. That is, we implement the sample-split and cross-fitted version using the default function `vimp_rsquared` in the W20b authors' R package `vimp` (version 2.1.0). Since W20b gives

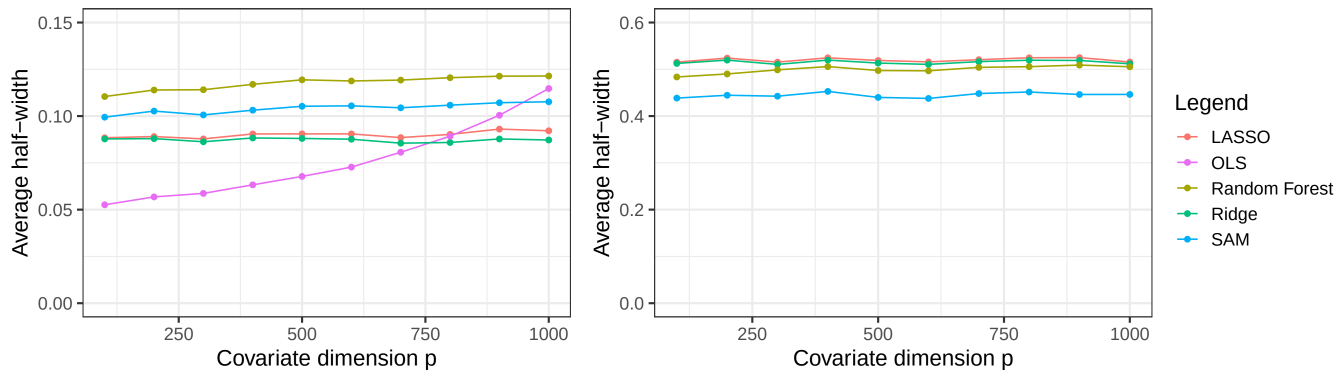

Figure 2: Average half-widths for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 4.3. OLS is run on the full sample. p is varied on the x-axis; see Section 4.1 for remaining details. Standard errors are below 0.002 (left) and 0.008 (right).

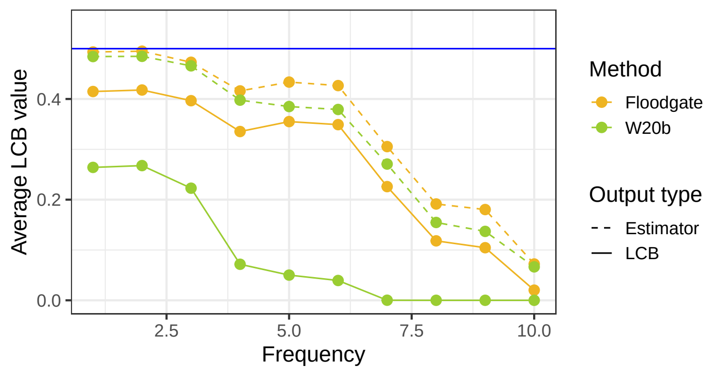

Figure 3: Average LCB values (solid lines) for floodgate and W20b in the sine function simulation of Section 4.4. The frequency λ is varied on the x-axis, and the solid blue line in the plot shows the true value of $\mathcal{I}$. The dashed lines correspond to the average estimator values of $\mathcal{I}$. The results are averaged over 640 independent replicates, and the standard errors are below 0.01.

confidence intervals for $\mathcal{I}^2/\mathrm{Var}(Y)$, we transform its inference into a $1-\alpha$ coverage LCB for $\mathcal{I}$ by taking the lower bound from its $1-2\alpha$ confidence interval, multiplying it by $\mathrm{Var}(Y)$, and then taking the square root. Our simulation example uses a sine function of varying frequency for $\mu^\star$. In particular, $p = 2$, the covariates $(X, Z) \in \mathbb{R}^2$ are i.i.d. uniformly distributed on $(-1, 1)$, and Y equals $A(\lambda)\sin(\lambda X)$ plus standard Gaussian noise, where $\lambda > 0$ controls the frequency and $A(\lambda)$ is chosen so that $\mathcal{I} = 0.5$ regardless of λ (thus ensuring the group mMSE gap of (X, Z) is always bounded away from zero, as required by W20b). Both floodgate and W20b must internally fit an estimate of $\mu^\star$, and for both methods we use locally-constant loess smoothing with tuning parameters selected by 5-fold cross-validation, following a different two-dimensional simulation example from Williamson et al. (2019).

The solid curves in Figure 3 show the average LCBs of the two methods applied to the non-null variable X as λ varies. Larger λ corresponds to less-smooth $\mathbb{E}[Y \mid X, Z]$ and hence a more challenging estimation problem (for both methods), and both methods become generally more conservative and less accurate as λ grows (both methods achieve at or above nominal coverage throughout this simulation; see Appendix J.4 for the coverage plot). Yet floodgate's LCB provides consistently and considerably more accurate inference over the entire range of λ. To better understand this performance difference, we additionally plot as dashed curves the average of the asymptotically normal estimators of $\mathcal{I}$ each method uses for inference. We see from the plot that the two estimators have similar bias, but the gap between the LCB and the estimator is

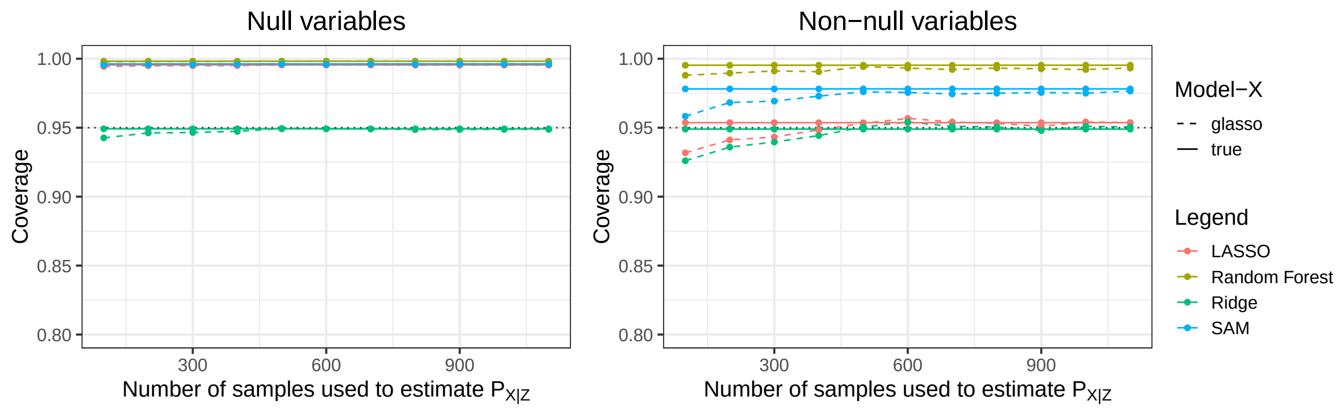

Figure 4: Coverage of null (left) and non-null (right) covariates when the covariate distribution is estimated in-sample for the linear-$\mu^\star$ simulations of Section 4.5. See Section 4.1 for remaining details. Standard errors are below 0.001 (left) and 0.008 (right).

much smaller for floodgate, reflecting a smaller variance. This is likely due to the form of W20b's estimator, which is the difference of two asymptotically normal test statistics, one computed on each half of the split data. Heuristically, one would expect this to lead to higher variance than an estimator computed on (and hence whose variance comes only from) one half of the data, like floodgate's. This general picture is reinforced by a higher-dimensional simulation given in Appendix J.4.

4.5 Robustness

In order to study the robustness of floodgate to misspecification of $P_{X|Z}$, we consider a scenario we expect to arise in practice: a data analyst does not know $P_{X|Z}$ exactly, so instead they estimate it using the data they have, and then treat the estimate as the "known" $P_{X|Z}$ and proceed with floodgate. Note that if the analyst splits the data and uses the same subset for estimating μ and for estimating $P_{X|Z}$, then Theorem E.1 applies, but if they use *all* of their data to estimate $P_{X|Z}$, then our theory does not apply. Also note we are not studying the performance of co-sufficient floodgate in this subsection.

Note that if the analyst splits the data and uses the same subset for estimating μ and for estimating $P_{X|Z}$, then Theorem E.1 applies, but if they use *all* of their data to estimate $P_{X|Z}$, then our theory does not apply. Also note we are not studying the performance of co-sufficient floodgate in this subsection.

Figure 4 varies how much in-sample data is used in $P_{X|Z}$-estimation and shows the coverage of floodgate for null and non-null variables in a linear setting. The estimation procedure is to fit the graphical LASSO (GLASSO) with 3-fold cross-validation to a subset of the in-sample data and treat $P_{X|Z}$ as conditionally Gaussian with covariance matrix given by the GLASSO estimate. Since $n = 1100$ in all these simulations and the sample splitting proportion is 0.5, when the x-axis value passes 550 is when the $P_{X|Z}$-estimation and inference sets start to overlap, and at the value 1100, all of the data is being used to estimate $P_{X|Z}$, including the half used for inference (violating Theorem E.1's assumptions). Nevertheless, we see the coverage is consistently quite high, only dropping slightly from that with true $P_{X|Z}$ for very low estimation sample sizes (i.e., very bad estimates of the covariance matrix). Note that some μ-fitting algorithms in Figure 4 have higher-than-nominal coverage; this is largely because the floodgate procedure will deterministically output a zero LCB (and hence have 100% coverage) when $\mu(x, z)$ does not depend on x. This happens for many covariates when μ is fitted via a sparse regression such as the LASSO and SAM (short for Sparse Additive Models), but also for our version of Random Forests which we effectively sparsify for computational reasons (see Appendix J.2 for details). Figures 5 and 6 show similar overcoverage for the same reason.

Average half-width plots corresponding to Figure 4 can be found in Appendix J.4. In additional to the linear setting in Figure 4, we also observe robust empirical coverage of floodgate when the conditional model of Y is nonlinear; see Appendix J.4 for details.

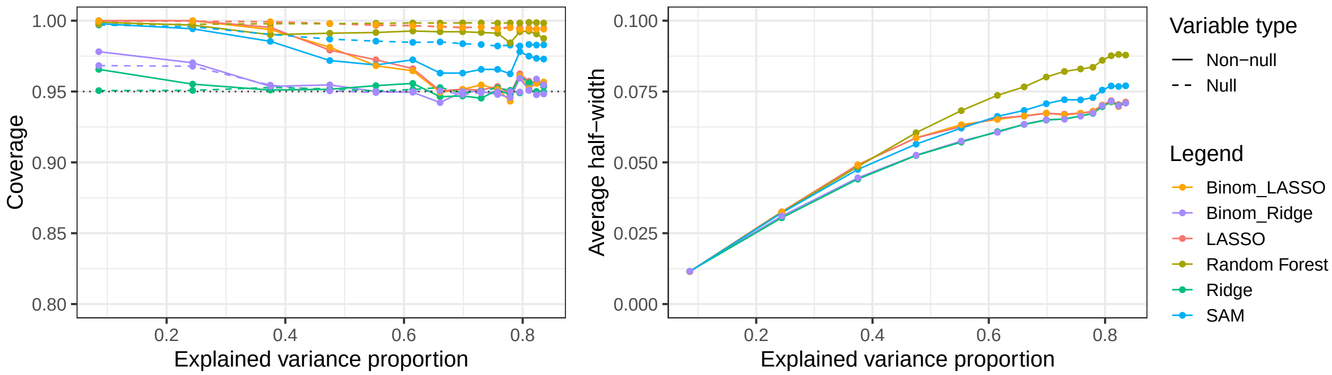

Figure 5: Coverage (left) and average half-widths (right) for the binary response simulations of Section 4.6. The explained variance proportion is varied over the x-axis. See Section 4.1 and 4.6 for remaining details. Standard errors are below 0.006 (left) and 0.001 (right).

4.6 Floodgate for the MACM gap

Here we study the empirical performance of floodgate applied to the MACM gap as described in Section 3.1. Conditional on the covariates, the binary response is generated from a logistic regression with $\frac{\log(\mathbb{P}(Y=1 \mid X,Z))}{\log(\mathbb{P}(Y=-1 \mid X,Z))}$ given by the linear $\mu^\star(X, Z)$ in Section 4.1. We set the sample size $n = 1000$, and the remaining simulation parameters to be the values described in Section 4.1. Figure 5 shows that floodgate has consistent coverage over a range of algorithms for fitting μ, and we see the dynamics of the average half-width as the explained variance proportion in $P_{Y|X,Z}$ increases. Note that R_i in Algorithm 3 needs to in general be estimated by Monte Carlo samples (see Appendix H for details) and in Figure 5, we set $K = 100$ and $M = 400$ whenever the Monte Carlo version is used.

4.7 Co-sufficient floodgate

Finally, we study the empirical performance of co-sufficient floodgate as described in Section 3.2 as compared to the original floodgate method which is given full knowledge of $P_{X|Z}$. We set the covariate dimension $p = 50$, the number of Monte Carlo samples $K = 100$, and the amplitude value for nonlinear-$\mu^\star$ to 30. The remaining simulation parameters are set to the values described in Section 4.1. Co-sufficient floodgate and the original floodgate procedure use the same working regression function, fitted from $n_e = 500$ samples, and use the same number of samples $n - n_e$ for inference. The batch size n_2 for co-sufficient floodgate is 300 and we vary the number of batches $n_1 = (n - n_e)/n_2$ on the x-axes. Co-sufficient floodgate is given the conditional variance of the Gaussian distribution of $X \mid Z$, but not its conditional mean, parameterized by a $(p-1)$-dimensional coefficient vector multiplying Z. Figure 6 shows that co-sufficient floodgate has satisfying coverage even when the number of batches is small, and has average half-width quite close to the original floodgate procedure which is given the conditional mean of $X \mid Z$ exactly. In additional to the nonlinear setting in Figure 6, simulations for a linear $\mu^\star$ lead to similar conclusions; see Appendix J.4.

5 Application to genomic study of platelet count

The study of genetic *heritability* is the study of how much variance in a trait can be explained by genetics. Precise definitions vary based on modeling assumptions (Zuk et al., 2012), but the fundamental concept is intuitive and central to genomics; indeed the goal of genome-wide association studies (GWAS) is often precisely to identify single nucleotide polymorphisms (SNPs) or loci that explain the most variance in a trait. To connect heritability with the present paper, suppose Y denotes a trait, X denotes a SNP or group of SNPs, and Z denotes all the remaining SNPs not included in X. Then as can be seen in Equation (2.2),

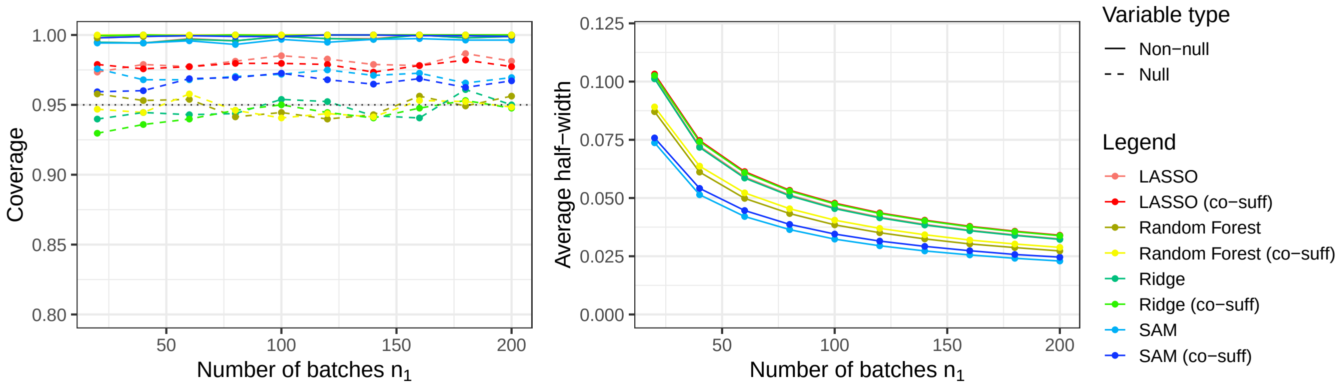

Figure 6: Coverage (left) and average half-widths (right) for co-sufficient floodgate and original floodgate in the nonlinear-$\mu^\star$ simulations. The number of batches n_1 is varied over the x-axis. See Section 4.1 and 4.7 for remaining details. Standard errors are below 0.009 (left) and 0.002 (right).

the mMSE gap $\mathcal{I}^2$ *exactly* measures the variance in Y that is attributable to X. Thinking of $\mathcal{I}^2$ as a sort of *conditional* heritability also makes it easy to include non-genetic factors such as age in Z, since such factors may influence Y but not be of direct interest to geneticists. Thus $\mathcal{I}^2$ can capture both gene-gene and gene-environment interactions.

Having established $\mathcal{I}^2$ as a quantity of interest, we proceed to infer it for blocks of SNPs at various resolutions of the human genome by applying floodgate to a platelet GWAS from the UK Biobank. Our analysis builds on the work of Sesia et al. (2020b), which carefully applied model-X knockoffs to the same data to perform multi-resolution *selection* of important SNPs, and in doing so require, like floodgate, a model for the SNPs X, Z and a working regression function, both of which we reuse in our own analysis. In particular, we follow the literature on genotype/haplotype modeling (Stephens et al., 2001; Zhang et al., 2002; Li and Stephens, 2003; Scheet and Stephens, 2006; Sesia et al., 2019, 2020a,b) and model the SNPs as following a hidden Markov model, and use the cross-validated Lasso as the algorithm to fit our working regression function μ. Although we use a linear μ to match the existing analysis in Sesia et al. (2020b), we remind the reader that one is in general free to use any μ with floodgate, and we hope that domain experts applying floodgate in the future to GWAS data can tailor μ to be even more powerful. The output of the analysis in Sesia et al. (2020b) is a so-called "Chicago plot", which plots stacked blocks of selected SNPs at a range of block resolutions. The height of the Chicago plot at a given location on the genome reflects the resolution at which the SNP at that location was rejected, with a greater height corresponding to a smaller block of SNPs being rejected. However, since the Chicago plot is derived from a pure selection method, it contains no information about the *strength* of the relationship between the trait and any of the blocks of SNPs. Floodgate enables us to construct a *colored* Chicago plot by computing an LCB for each selected block of SNPs and reporting an LCB of zero (without computation) for all unselected blocks of SNPs; see Appendix K for implementation details.

In particular, Figure 7 is a colored version of Figure 1a of Sesia et al. (2020b), which displayed the genomic regions on chromosome 12 that those authors found to be related to platelet count in the UK Biobank data. Our colored figure shows how informative floodgate LCBs can be over and beyond a pure selection method, as it shows the signal is far from being spread evenly over the SNPs selected by Sesia et al. (2020b). This information is crucial for the *prioritization* of selected regions, as without color the Chicago plot does not give any indication which of the selected SNPs the data indicates are most important (we note that the height of the tallest selected block at a SNP need *not* correspond to its importance, and indeed there are many pairs of locations in the figure such that one has a taller block in the original Chicago plot but the other has a brighter color in Figure 7).

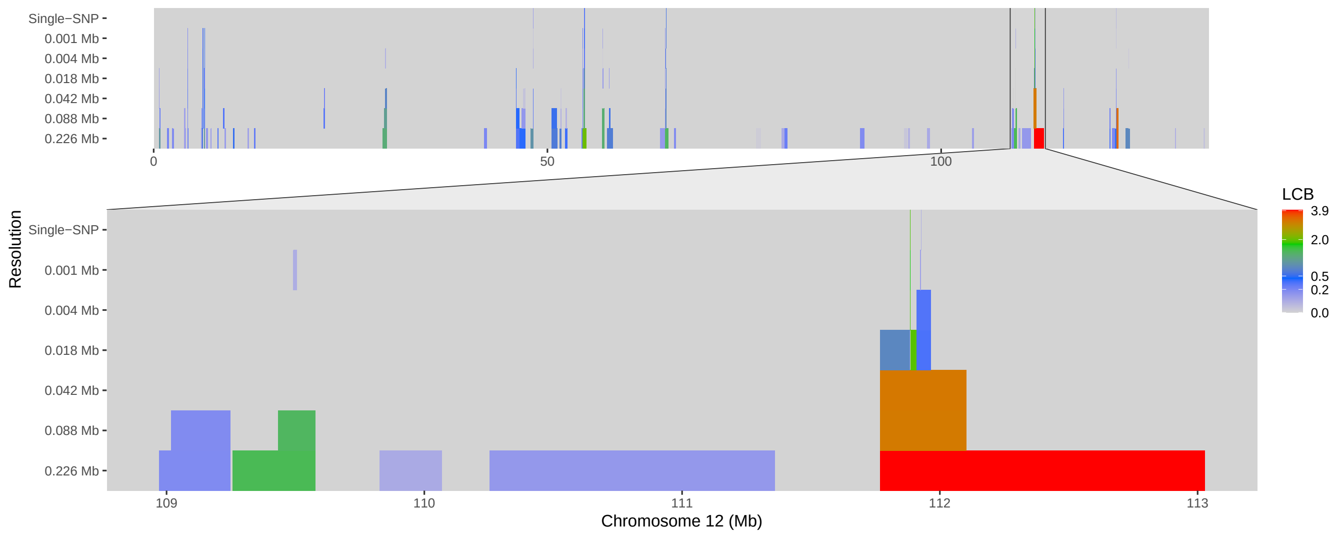

Figure 7: Colored Chicago plot analogous to Figure 1a of Sesia et al. (2020b). The color of each point represents the floodgate LCB for the block that contains the SNP at the location indicated on the x-axis at the resolution (measured by average block width) indicated on the y-axis (note some blocks appearing in the original Chicago plot have an LCB of zero and hence are colored grey). The second panel zooms into the region of the first panel containing the largest floodgate LCB.

6 Discussion

Floodgate is a powerful and flexible framework for rigorously inferring the strength of the conditional relationship between Y and X. We prove results about floodgate's validity, accuracy, and robustness and address a number of extensions/generalizations, but a number of questions remain for future work and we highlight two here:

- Floodgate relies on a working regression function that is not estimated from the same data used for inference, which usually will require data splitting. It would be desirable, both from an accuracy standpoint and a derandomization standpoint, to remove the need for data splitting or at least find a way for samples in one or both splits to be recycled between regression estimation and inference.
- The floodgate framework is applied here to the mMSE gap and the MACM gap, but more generally it constitutes a new tool for flexible inference of nonparametric functionals, and we expect it can find use for inferring other MOVIs. The main challenge for its application is the identification of an appropriate floodgate functional, and it would be of interest to better understand principles or even heuristics for finding such functionals for a given MOVI. Indeed we make no claim that the functionals proposed in this paper are unique for their respective MOVIs, and there may be others that lead to better floodgate procedures.

Acknowledgements

L.Z. is partially supported by the NSF-Simons Center for Mathematical and Statistical Analysis of Biology at Harvard, award number #1764269 and the Harvard Quantitative Biology Initiative. L.J. is partially supported by the William F. Milton Fund. We would like to thank the Neale Lab at the Broad Institute of MIT and Harvard for including us in their application to the UK Biobank Resource (application 31063), Sam Bryant for the access to on-premise data files, Matteo Sesia, Eugene Katsevich, Asher Spector, Benjamin Spector, Masahiro Kanai and Nikolas Baya for the help with the genomics application, and Dongming Huang for helpful discussions.

References

Azadkia, M. and Chatterjee, S. (2019). A simple measure of conditional dependence. *arXiv preprint arXiv:1910.12327*.

Barber, R. F. and Janson, L. (2020). Testing goodness-of-fit and conditional independence with approximate co-sufficient sampling. *arXiv preprint arXiv:2007.09851*.

Bates, S., Sesia, M., Sabatti, C., and Candès, E. J. (2020). Causal inference in genetic trio studies. *Proceedings of the National Academy of Sciences*, 117(39):24117–24126.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testingunder dependency. *The Annals of Statistics*, 29(4):1165 – 1188.

Benjamini, Y. and Yekutieli, D. (2005). False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81.

Bentkus, V., Götze, F., et al. (1996). The berry-esseen bound for student's statistic. *The Annals of Probability*, 24(1):491–503.

Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L., et al. (2013). Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837.

Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. (2020). The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):175–197.

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

Bühlmann, P. et al. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242.

Bühlmann, P., van de Geer, S., et al. (2015). High-dimensional inference in misspecified linear models. *Electronic Journal of Statistics*, 9(1):1449–1473.

Buja, A., Berk, R. A., Brown, L. D., George, E. I., Pitkin, E., Traskin, M., Zhao, L., and Zhang, K. (2015). Models as Approximations–A conspiracy of random regressors and model deviations against classical inference in regression. *Statistical Science*, page 1.

Buja, A. and Brown, L. (2014). Discussion:"a significance test for the lasso". *The Annals of Statistics*, 42(2):509–517.

Buja, A., Brown, L., Berk, R., George, E., Pitkin, E., Traskin, M., Zhang, K., and Zhao, L. (2019a). Models as Approximations I: Consequences illustrated with linear regression. *Statistical Science*, 34(4):523 – 544.

Buja, A., Brown, L., Kuchibhotla, A. K., Berk, R., George, E., Zhao, L., et al. (2019b). Models as Approximations II: A model-free theory of parametric regression. *Statistical Science*, 34(4):545–565.

Candès, E. J., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold: Model-X knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577.

Castro, J., Gómez, D., and Tejada, J. (2009). Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730.

Charnes, A., Golany, B., Keane, M., and Rousseau, J. (1988). Extremal principle solutions of games in characteristic function form: core, Chebychev and Shapley value generalizations. In *Econometrics of Planning and Efficiency*, pages 123–133. Springer.

Chatterjee, S. (2021). A new coefficient of correlation. *Journal of the American Statistical Association*, 116(536):2009–2022.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018a). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018b). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chia, C., Sesia, M., Ho, C.-S., Jeffrey, S. S., Dionne, J., Candès, E. J., and Howe, R. T. (2020). Interpretable signal analysis with knockoffs enhances classification of bacterial raman spectra. *arXiv preprint arXiv:2006.04937*.

Covert, I., Lundberg, S. M., and Lee, S.-I. (2020). Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33.

Deb, N., Ghosal, P., and Sen, B. (2020). Measuring association on topological spaces using kernels and geometric graphs. *arXiv preprint arXiv:2010.01768*.

Deb, N. and Sen, B. (2021). Multivariate rank-based distribution-free nonparametric testing using measure transportation. *Journal of the American Statistical Association*, pages 1–16.

Dezeure, R., Bühlmann, P., and Zhang, C.-H. (2017). High-dimensional simultaneous inference with the bootstrap. *Test*, 26(4):685–719.

Dharmadhikari, S., Jogdeo, K., et al. (1969). Bounds on moments of certain random variables. *The Annals of Mathematical Statistics*, 40(4):1506–1509.

Feng, J., Williamson, B., Simon, N., and Carone, M. (2018). Nonparametric variable importance using an augmented neural network with multi-task learning. In *International Conference on Machine Learning*, pages 1496–1505.

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2008). Kernel measures of conditional dependence. In *Advances in neural information processing systems*, pages 489–496.

Gil, M. (2011). *On Rényi divergence measures for continuous alphabet sources*. PhD thesis, Citeseer.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. In *International Conference on Algorithmic Learning Theory*, pages 63–77. Springer.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. (2007). A kernel statistical test of independence. *Advances in Neural Information Processing Systems*, 20.

Hainmueller, J. and Hopkins, D. J. (2014). Public attitudes toward immigration. *Annual Review of Political Science*, 17(1):225–249.

Heller, R., Heller, Y., and Gorfine, M. (2013). A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2):503–510.

Hirschfeld, H. O. (1935). A connection between correlation and contingency. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 31, pages 520–524. Cambridge University Press.

Huang, D. and Janson, L. (2020). Relaxing the assumptions of knockoffs by conditioning. *The Annals of Statistics*, 48(5):3021 – 3042.

Huang, Z., Deb, N., and Sen, B. (2020). Kernel partial correlation coefficient–a measure of conditional dependence. *arXiv preprint arXiv:2012.14804*.

Janson, L. (2017). *A Model-Free Approach to High-Dimensional Inference*. PhD thesis, Stanford University.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909.

Katsevich, E. and Ramdas, A. (2020). A theoretical treatment of conditional independence testing under model-X. *arXiv preprint arXiv:2005.05506*.

Katsevich, E. and Roeder, K. (2020). Conditional resampling improves sensitivity and specificity of single cell crispr regulatory screens. *bioRxiv*.

Katsevich, E. and Sabatti, C. (2019). Multilayer knockoff filter: Controlled variable selection at multiple resolutions. *The Annals of Applied Statistics*, 13(1):1 – 33.

Lee, J. D., Sun, D. L., Sun, Y., Taylor, J. E., et al. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Li, L., Tchetgen Tchetgen, E., van der Vaart, A., and Robins, J. M. (2011). Higher order inference on a treatment effect under low regularity conditions. *Statistics & Probability Letters*, 81(7):821–828. Statistics in Biological and Medical Sciences.

Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233.

Liu, L., Mukherjee, R., and Robins, J. M. (2019). On assumption-free tests and confidence intervals for causal effects estimated by machine learning. *arXiv preprint arXiv:1904.04276*.

Liu, M., Katsevich, E., Janson, L., and Ramdas, A. (2021). Fast and powerful conditional randomization testing via distillation. *Biometrika*. asab039.

Luce, R. and Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, 1(1):1–27.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):2522–5839.

Newey, W. K. and Robins, J. R. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*.

Nickl, R., Van De Geer, S., et al. (2013). Confidence sets in sparse regression. *The Annals of Statistics*, 41(6):2852–2876.

Owen, A. B. and Prieur, C. (2017). On shapley value for measuring importance of dependent inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):986–1002.

Park, J. and Muandet, K. (2020). A measure-theoretic approach to kernel conditional mean embeddings. *Advances in Neural Information Processing Systems*, 33:21247–21259.

Pinelis, I., Molzon, R., et al. (2016). Optimal-order bounds on the rate of convergence to normality in the multivariate delta method. *Electronic Journal of Statistics*, 10(1):1001–1063.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030.

Ren, B., Schwager, E., Tickle, T., and Huttenhower, C. (2016). sparsDOSSA sparse data observations for simulating synthetic abundance. *R package version 1.12. 0*.

Rinaldo, A., Wasserman, L., and G'Sell, M. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *The Annals of Statistics*, 47(6):3438 – 3469.

Robins, J., Li, L., Tchetgen, E., van der Vaart, A., et al. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics.

Robins, J., Tchetgen, E. T., Li, L., and van der Vaart, A. (2009). Semiparametric minimax rates. *Electronic Journal of Statistics*, 3:1305.

Robins, J. M., Li, L., Mukherjee, R., Tchetgen, E. T., van der Vaart, A., et al. (2017). Minimax estimation of a functional on a structured high-dimensional model. *The Annals of Statistics*, 45(5):1951–1987.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. (2008). *Global sensitivity analysis: the primer*. John Wiley & Sons.

Scheet, P. and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78(4):629–644.

Sesia, M., Bates, S., Candès, E. J., Marchini, J., and Sabatti, C. (2020a). Controlling the false discovery rate in gwas with population structure. *bioRxiv*.

Sesia, M., Katsevich, E., Bates, S., Candès, E. J., and Sabatti, C. (2020b). Multi-resolution localization of causal variants across the genome. *Nature communications*, 11(1):1–10.

Sesia, M., Sabatti, C., and Candès, E. J. (2019). Gene hunting with hidden markov model knockoffs. *Biometrika*, 106(1):1–18.

Shah, R. D. and Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514 – 1538.

Shao, X. and Zhang, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association*, 109(507):1302–1318.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317.

Stephens, M., Smith, N. J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, 68(4):978–989.

Stephens, M. A. (2012). Goodness-of-fit and sufficiency: Exact and approximate tests. *Methodology and Computing in Applied Probability*, 14(3):785–791.

Štrumbelj, E. and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665.

Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794.

Tansey, W., Veitch, V., Zhang, H., Rabadan, R., and Blei, D. M. (2022). The holdout randomization test for feature selection in black box models. *Journal of Computational and Graphical Statistics*, 31(1):151–162.

Taylor, J., Lockhart, R., Tibshirani, R. J., and Tibshirani, R. (2014). Exact post-selection inference for forward stepwise and least angle regression. *arXiv preprint arXiv:1401.3889*, 7:10–1.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., et al. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

Wang, X., Jiang, B., and Liu, J. S. (2017). Generalized r-squared for detecting dependence. *Biometrika*, 104(1):129–139.

Wang, X., Pan, W., Hu, W., Tian, Y., and Zhang, H. (2015). Conditional distance correlation. *Journal of the American Statistical Association*, 110(512):1726–1734.

Watson, D. S. and Wright, M. N. (2019). Testing conditional predictive independence in supervised learning algorithms. *arXiv preprint arXiv:1901.09917*.

Williamson, B. and Feng, J. (2020). Efficient nonparametric statistical inference on population feature importance using shapley values. In *International Conference on Machine Learning*, pages 10282–10291.

Williamson, B. D., Gilbert, P. B., Carone, M., and Simon, N. (2019). Nonparametric variable importance assessment using machine learning techniques. *Biometrics*.

Williamson, B. D., Gilbert, P. B., Simon, N. R., and Carone, M. (2020). A unified approach for inference on algorithm-agnostic variable importance. *arXiv preprint arXiv:2004.03683*.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.

Zhang, K., Deng, M., Chen, T., Waterman, M. S., and Sun, F. (2002). A dynamic programming algorithm for haplotype block partitioning. *Proceedings of the National Academy of Sciences*, 99(11):7335–7339.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 804–813. AUAI Press.

Zhang, X. and Cheng, G. (2017). Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association*, 112(518):757–768.

Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198.

A Proofs for main text

Throughout the proofs, we will abbreviate $(X, Z) = W$, $(\tilde{X}, Z) = \tilde{W}$ for simplicity and write $w = (x, z)$. And $g^\star, g : \mathbb{R}^{p-1} \to \mathbb{R}$; $h^\star, h : \mathbb{R}^p \to \mathbb{R}$ are defined as below:

$$g^\star(z) = \mathbb{E}[\mu^\star(W) \mid Z = z], \ g(z) = \mathbb{E}[\mu(W) \mid Z = z], \tag{A.1}$$

$$h^\star(w) = \mu^\star(w) - g^\star(z), \ h(w) = \mu(w) - g(z). \tag{A.2}$$

And we can further decompose Y:

$$Y = \mathbb{E}[Y \mid X, Z] + \epsilon(Y, X, Z) = \mu^\star(W) + \epsilon(Y, W) = g^\star(Z) + h^\star(W) + \epsilon(Y, W). \tag{A.3}$$

Let $L_2(\Omega, \mathcal{F}, P)$ denote the vector space of real-valued random variables with finite second moments, which is a Hilbert space, and define its subspace $L_2(W) := L_2(\Omega, \mathscr{A}(W), P)$, where $\mathscr{A}(W)$ is the sub σ-algebra generated by $W = (X, Z)$. ($L_2(Z) := L_2(\Omega, \mathscr{A}(Z), P)$ is defined analogously). Then $\mu^\star(W)$ and $g^\star(Z)$ can be interpreted as the projections of Y onto the subspaces $L_2(W)$ and $L_2(Z)$, respectively. Y and $\mu^\star(W)$ admit the orthogonal decompositions $Y = \mu^\star(W) + \epsilon(Y, W)$ and $\mu^\star(W) = g^\star(Z) + h^\star(W)$, respectively. Similarly note the projection of $\mu(W)$ onto $L_2(Z)$ and the decomposition $\mu(W) = g(Z) + h(W)$. We remark these imply the following facts:

$$\begin{gathered} \mathbb{E}[\epsilon(Y, W) \mid W] = 0, \ \mathbb{E}[\epsilon(Y, W)\lambda(W)] = 0, \\ \mathbb{E}[h^\star(W) \mid Z] = 0, \ \mathbb{E}[h^\star(W)\gamma(Z)] = 0, \mathbb{E}[h(W) \mid Z] = 0, \ \mathbb{E}[h(W)\gamma(Z)] = 0. \end{gathered} \tag{A.4}$$

for any function $\lambda(w)$ and any function $\gamma(z)$. Thus we can rewrite the denominator of $f(\mu)$ by noticing the equivalence below:

$$\mathbb{E}[\mathrm{Var}(\mu(X, Z) \mid Z)] = \mathbb{E}[\mathrm{Var}(h(W) \mid Z)] = \mathbb{E}\left[\mathbb{E}\left[h^2(W) \mid Z\right]\right] = \mathbb{E}\left[h^2(W)\right]. \tag{A.5}$$

As for the numerator of $f(\mu)$, (2.7) mentions the rewritten expression. Here we formally derive the following equivalent expressions of $f(\mu)$,

$$\begin{aligned} f(\mu) \ &:= \ \frac{\mathbb{E}[\mathrm{Cov}(\mu^\star(X, Z), \mu(X, Z) \mid Z)]}{\sqrt{\mathbb{E}[\mathrm{Var}(\mu(X, Z) \mid Z)]}} \\ &= \ \frac{\mathbb{E}[\mathrm{Cov}(h^\star(W), h(W) \mid Z)]}{\sqrt{\mathbb{E}[h^2(W)]}} \\ &= \ \frac{\mathbb{E}[h^\star(W)h(W)]}{\sqrt{\mathbb{E}[h^2(W)]}} \qquad \text{(A.6)} \\ &= \ \frac{\mathbb{E}[Yh(W)]}{\sqrt{\mathbb{E}[h^2(W)]}} - \frac{\mathbb{E}[\epsilon(Y, W)h(W)]}{\sqrt{\mathbb{E}[h^2(W)]}} - \frac{\mathbb{E}[g^\star(Z)h(W)]}{\sqrt{\mathbb{E}[h^2(W)]}} \\ &= \ \frac{\mathbb{E}[Yh(W)]}{\sqrt{\mathbb{E}[h^2(W)]}} \qquad \text{(A.7)} \end{aligned}$$

where the second equality is by (A.5) and the definitions of $h^\star(W), h(W)$, the third equality holds by the total law of conditional expectation and (A.4), the fourth equality comes from (A.3), and the last equality holds due to (A.4) and the total law of conditional expectation. As (A.7) is very concise, we will work with this expression of $f(\mu)$ throughout the following proof. Also note we have an equivalent expression of $\mathcal{I}$.

$$\sqrt{\mathbb{E}[(h^\star)^2(W)]} = \sqrt{\mathbb{E}\left[\mathbb{E}\left[(\mu^\star(W) - \mathbb{E}[\mu^\star(W) \mid Z])^2 \,\middle|\, Z\right]\right]} = \sqrt{\mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid X, Z] \mid Z)]} = \mathcal{I}. \tag{A.8}$$

Note that the proofs of Theorems 2.3 and 2.5 only require moment conditions on $h(W)$, which will hold under the corresponding moment conditions on $\mu(X,Z)$. This can be seen from the following example where the finiteness of $\mathbb{E}[\mu^r(W)]$ implies that of $\mathbb{E}[h^r(W)]$ for some positive integer r:

$$\begin{aligned}\mathbb{E}\left[h^r(W)\right] &= \mathbb{E}\left[(\mu(W) - \mathbb{E}\left[\mu(W) \mid Z\right])^r\right] \\ &\leq 2^{r-1}(\mathbb{E}\left[\mu^r(W)\right] + \mathbb{E}\left[(\mathbb{E}\left[\mu(W) \mid Z\right])^r\right]) \\ &\leq 2^{r-1}(\mathbb{E}\left[\mu^r(W)\right] + \mathbb{E}\left[\mathbb{E}\left[\mu^r(W) \mid Z\right]\right]) = 2^r \mathbb{E}\left[\mu^r(X,Z)\right],\end{aligned} \tag{A.9}$$

where the first inequality holds due to the C_r inequality (which states that $\mathbb{E}[|X+Y|^r] \leq C_r(\mathbb{E}[|X|^r] + \mathbb{E}[|Y|^r])$ with $C_r = 1$ for $0 < r \leq 1$ and $C_r = 2^{r-1}$ for $r \geq 1$), the second inequality holds by Jensen's inequality, and the last equality holds due to the tower property of conditional expectation.

In the proofs of Theorems 2.3 and 2.5, we will use a key fact to simplify exposition: when $\mathbb{E}\left[h^2(W)\right] > 0$, $\mathbb{E}\left[h^2(W)\right] = 1$ can be assumed without loss of generality. This is because (A.7) says $f(\mu) = \frac{\mathbb{E}[Y h(W)]}{\sqrt{\mathbb{E}[h^2(W)]}}$ and R_i, V_i in Algorithm 1 can be rewritten as

$$\begin{aligned} R_i &= Y_i(\mu(X_i, Z_i) - \mathbb{E}\left[\mu(X_i, Z_i \mid Z_i\right]) = Y_i h(W_i), \\ V_i &= \mathrm{Var}\left(\mu(X_i, Z_i) \mid Z_i\right) = \mathrm{Var}\left(h(X_i, Z_i) \mid Z_i\right) \end{aligned}$$

by definition of h. Regarding Theorem 2.5, R_i^K, V_i^K can be rewritten as

$$\begin{aligned} R_i^K &= Y_i\left(h(W_i) - \frac{1}{K}\sum_{k=1}^K h(X_i^{(k)}, Z_i)\right), \\ V_i^K &= \frac{1}{K-1}\sum_{k=1}^K \left(h(X_i^{(k)}, Z_i) - \frac{1}{K}\sum_{k=1}^K h(X_i^{(k)}, Z_i)\right)^2 \end{aligned}$$

due to (A.51), (A.53). It is immediate that the floodgate procedure is invariant to positive scaling thus we assume $\mathbb{E}\left[h^2(W)\right] = 1$ without loss of generality.

A.1 Proofs in Section 2.2

A.1.1 Lemma 2.2

Proof of Lemma 2.2. When $\mathbb{E}\left[\mathrm{Var}(\mu(X,Z) \mid Z)\right] = 0$, the numerator must also be zero, and hence the ratio is 0 by convention and $f(\mu) \leq \mathcal{I}$. Now assuming $\mathbb{E}\left[\mathrm{Var}(\mu(X,Z) \mid Z)\right] > 0$,

$$\begin{aligned} f(\mu) &= \frac{\mathbb{E}\left[\mathrm{Cov}(\mu(X,Z), \mu^\star(X,Z) \mid Z)\right]}{\sqrt{\mathbb{E}\left[\mathrm{Var}(\mu(X,Z) \mid Z)\right]}} \\ &= \frac{\mathbb{E}\left[\sqrt{\mathrm{Var}(\mu(X,Z) \mid Z)}\sqrt{\mathrm{Var}(\mu^\star(X,Z) \mid Z)}\,\mathrm{Cor}\left(\mu(X,Z), \mu^\star(X,Z) \mid Z\right)\right]}{\sqrt{\mathbb{E}\left[\mathrm{Var}(\mu(X,Z) \mid Z)\right]}} \\ &\leq \frac{\mathbb{E}\left[\sqrt{\mathrm{Var}(\mu(X,Z) \mid Z)}\sqrt{\mathrm{Var}(\mu^\star(X,Z) \mid Z)}\right]}{\sqrt{\mathbb{E}\left[\mathrm{Var}(\mu(X,Z) \mid Z)\right]}} \\ &\leq \frac{\sqrt{\mathbb{E}\left[\mathrm{Var}(\mu(X,Z) \mid Z)\right]}\sqrt{\mathbb{E}\left[\mathrm{Var}(\mu^\star(X,Z) \mid Z)\right]}}{\sqrt{\mathbb{E}\left[\mathrm{Var}(\mu(X,Z) \mid Z)\right]}} = \mathcal{I}, \end{aligned}$$

where the first inequality uses the fact that correlation is bounded by 1, and the second inequality uses Cauchy–Schwarz. Finally, it is immediate that $f(\mu^\star) = \mathcal{I}$. □

A.1.2 Theorem 2.3

Proof of Theorem 2.3. Due to (A.9), $\mathbb{E}\left[\mu^4(X,Z)\right]<\infty$ implies $\mathbb{E}\left[h^4(W)\right]<\infty$. In the following proof, we will only assume the weaker moment conditions $\mathbb{E}\left[Y^4\right], \mathbb{E}\left[h^4(W)\right]<\infty$. Under such moment conditions, we also have $\mathbb{E}\left[Yh(W)\right]\le\sqrt{\mathbb{E}\left[Y^2\right]}\sqrt{\mathbb{E}\left[h^2(W)\right]}$ and $\mathbb{E}\left[h^2(W)\right]<\infty$ since the finiteness of higher moments implies that of lower moments.

When $\mu(X,Z)\in\mathscr{A}(Z)$, i.e., $\mathbb{E}\left[\mathrm{Var}(\mu(X,Z)\,|\,Z)\right]=0$, we immediately have coverage since $L_n^\alpha(\mu)=0$ by construction and $\mathcal{I}\ge 0$ by its definition. Regarding the case where $\mathbb{E}\left[\mathrm{Var}(\mu(X,Z)\,|\,Z)\right]\neq 0$, we have $\mathbb{E}\left[h^2(W)\right]=\mathbb{E}\left[\mathrm{Var}(\mu(X,Z)\,|\,Z)\right]>0$ due to (A.5). Based on the discussions in the part after (A.9), we can assume $\mathbb{E}\left[h^2(W)\right]=1$ without loss of generality.

Recall in Algorithm 1, we denote $R_i=Y_i\big(\mu(X_i,Z_i)-\mathbb{E}\left[\mu(X_i,Z_i)\,|\,Z_i\right]\big)$ and $V_i=\mathrm{Var}\left(\mu(X_i,Z_i)\,|\,Z_i\right)$ for each $i\in[n]$, and compute their sample mean $(\bar{R},\bar{V})$ and sample covariance matrix $\hat{\Sigma}$. The LCB is constructed as

$$L_n^\alpha(\mu)=\max\left\{\frac{\bar{R}}{\sqrt{\bar{V}}}-\frac{z_\alpha s}{\sqrt{n}},0\right\},\quad\text{where}\quad s^2=\frac{1}{\bar{V}}\left[\left(\frac{\bar{R}}{2\bar{V}}\right)^2\hat{\Sigma}_{22}+\hat{\Sigma}_{11}-\frac{\bar{R}}{\bar{V}}\hat{\Sigma}_{12}\right].$$

And we have

$$\{L_n^\alpha(\mu)\le\mathcal{I}\}=\left\{\frac{\bar{R}}{\sqrt{\bar{V}}}-\frac{z_\alpha s}{\sqrt{n}}\le\mathcal{I}\right\}\supset\left\{\frac{\bar{R}}{\sqrt{\bar{V}}}-\frac{z_\alpha s}{\sqrt{n}}\le f(\mu)\right\},$$

where the first equality holds since $\mathcal{I}\ge 0$ and the subset relation holds due to Lemma 2.2. Hence it suffices to show that

$$\mathbb{P}\left(\frac{\bar{R}}{\sqrt{\bar{V}}}-\frac{z_\alpha s}{\sqrt{n}}\le f(\mu)\right)\ge 1-\alpha-o(1). \tag{A.10}$$

We will utilize the central limit theorem (CLT) and the delta method to prove the above result. Now we consider four different cases.

(I) $\mathrm{Var}\left(Yh(W)\right)=0$ and $\mathrm{Var}\left(\mathrm{Var}\left(h(W)\,|\,Z\right)\right)=0$.

(II) $\mathrm{Var}\left(Yh(W)\right)>0$ and $\mathrm{Var}\left(\mathrm{Var}\left(h(W)\,|\,Z\right)\right)=0$.

(III) $\mathrm{Var}\left(Yh(W)\right)=0$ and $\mathrm{Var}\left(\mathrm{Var}\left(h(W)\,|\,Z\right)\right)>0$.

(IV) $\mathrm{Var}\left(Yh(W)\right)>0$ and $\mathrm{Var}\left(\mathrm{Var}\left(h(W)\,|\,Z\right)\right)>0$.

Note that assuming $\mathbb{E}\left[Y^4\right]$ and $\mathbb{E}\left[h^4(W)\right]<\infty$ ensures all the above variances exist; the bounding strategy is the same as (A.9), thus we omit the proof. When $\mathrm{Var}\left(Yh(W)\right)=0$ and $\mathrm{Var}\left(\mathrm{Var}\left(h(W)\,|\,Z\right)\right)=0$, respectively, we have the following facts.

$$\mathrm{Var}\left(Yh(W)\right)=0\ \Rightarrow\ R_i=\mathbb{E}\left[Yh(W)\right],\forall\, i\in[n],\bar{R}=\mathbb{E}\left[Yh(W)\right],\ \hat{\Sigma}_{11}=\hat{\Sigma}_{12}=0, \tag{A.11}$$

$$\mathrm{Var}\left(\mathrm{Var}\left(h(W)\,|\,Z\right)\right)=0\ \Rightarrow\ V_i=\mathbb{E}\left[h^2(W)\right],\forall\, i\in[n],\bar{V}=\mathbb{E}\left[h^2(W)\right],\ \hat{\Sigma}_{22}=\hat{\Sigma}_{12}=0. \tag{A.12}$$

Case (I): due to (A.11) and (A.12), we simply have $\frac{\bar{R}}{\sqrt{\bar{V}}}=\mathbb{E}\left[Yh(W)\right]/\sqrt{\mathbb{E}\left[h^2(W)\right]}=f(\mu)$ and $s=0$, thus (A.10) holds.

Case (II): due to (A.12), $s^2=\hat{\Sigma}_{11}/\bar{V}=\hat{\Sigma}_{11}/\mathbb{E}\left[h^2(W)\right]$, hence we have the following equivalence

$$\left\{\frac{\bar{R}}{\sqrt{\bar{V}}}-\frac{z_\alpha s}{\sqrt{n}}\le f(\mu)\right\}=\left\{\bar{R}-\frac{z_\alpha(\hat{\Sigma}_{11})^{1/2}}{\sqrt{n}}\le\mathbb{E}\left[Yh(W)\right]\right\}.$$

Thus the problem is reduced to showing that

$$\mathbb{P}\left(\bar{R}-\frac{z_\alpha(\hat{\Sigma}_{11})^{1/2}}{\sqrt{n}}\le\mathbb{E}\left[Yh(W)\right]\right)\ge 1-\alpha-o(1). \tag{A.13}$$

Notice $\bar{R}$ is simply the sample mean estimator of the quantity $\mathbb{E}[Yh(W)]$ and $\hat{\Sigma}_{11}$ is the corresponding sample variance. (A.13) is an immediate result of the central limit theorem and Slutsky's theorem.

Case (III): due to (A.11), we have

$$\frac{\bar{R}}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n}} = \frac{\mathbb{E}[Yh(W)]}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n}}, \quad \text{where} \quad s^2 = \frac{1}{\bar{V}}\left(\frac{\mathbb{E}[Yh(W)]}{2\bar{V}}\right)^2 \hat{\Sigma}_{22}.$$

$\frac{\mathbb{E}[Yh(W)]}{\sqrt{\bar{V}}}$ is a nonlinear function of the moment estimators. We will use the delta method to establish the asymptotic normality result. In case (IV), $\mathbb{E}[Yh(W)]$ is further replaced by its moment estimator, and we are dealing with a bit more complicated nonlinear statistic than $1/\sqrt{\bar{V}}$. Hence we focus on case (IV) and omit the very similar proof for case (III).

Case (IV): since $\mathrm{Var}(Yh(W)) > 0$ and $\mathrm{Var}(\mathrm{Var}(h(W) \mid Z)) > 0$, we have as $n \to \infty$,

$$\sqrt{n}\begin{pmatrix} \bar{R} - \mathbb{E}[Yh(W)] \\ \bar{V} - \mathbb{E}\left[h^2(W)\right] \end{pmatrix} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma) \tag{A.14}$$

by the multivariate central limit theorem, where the covariance matrix of the random vector $(R_i, V_i) \in \mathbb{R}^2$ is denoted by Σ with

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} \mathrm{Var}(Yh(W)) & \mathrm{Cov}(Yh(W), \mathrm{Var}(h(W) \mid Z)) \\ \mathrm{Cov}(Yh(W), \mathrm{Var}(h(W) \mid Z)) & \mathrm{Var}(\mathrm{Var}(h(W) \mid Z)) \end{pmatrix}.$$

$\mathbb{E}\left[Y^4\right], \mathbb{E}\left[h^4(W)\right] < \infty$ ensures the finiteness of $\Sigma_{11}, \Sigma_{12}, \Sigma_{22}$. Denote

$$\tilde{\sigma}_0^2 = \frac{1}{\mathbb{E}[h^2(W)]}\left[\left(\frac{\mathbb{E}[Yh(W)]}{2\mathbb{E}[h^2(W)]}\right)^2 \Sigma_{22} + \Sigma_{11} - \frac{\mathbb{E}[Yh(W)]}{\mathbb{E}[h^2(W)]}\Sigma_{12}\right], \tag{A.15}$$

and we will show $\tilde{\sigma}_0 > 0$ over the course of derivations from (A.20) to the end of the proof. Now consider

$$\left(\frac{\bar{R}}{\sqrt{\bar{V}}} - f(\mu)\right)/s = \left(\frac{\bar{R}}{\sqrt{\bar{V}}} - f(\mu)\right)/\tilde{\sigma}_0 \cdot \frac{\tilde{\sigma}_0}{s} := \frac{H(\bar{R}, \bar{V}) - f(\mu)}{\tilde{\sigma}_0} \cdot \left(\frac{s}{\tilde{\sigma}_0}\right)^{-1}, \tag{A.16}$$

where $H(x_1, x_2) : \mathbb{R}^2 \to \mathbb{R}$ is defined as $H(x_1, x_2) = x_1/\sqrt{x_2}$ for $x_2 > 0$ and its gradient equals $\nabla H(x_1, x_2) = (\frac{\partial H}{\partial x_1}, \frac{\partial H}{\partial x_2}) = \frac{1}{\sqrt{x_2}}(1, -\frac{x_1}{2x_2})$. Let $\theta = (\mathbb{E}[Yh(W)], \mathbb{E}\left[h^2(W)\right])$, then

$$\nabla H(\theta) = \frac{1}{\sqrt{\mathbb{E}[h^2(W)]}}\left(1, -\frac{\mathbb{E}[Yh(W)]}{2\mathbb{E}[h^2(W)]}\right), \tag{A.17}$$

and we obtain

$$\mathrm{Var}\left(\nabla H(\theta)^\top \begin{pmatrix} \sqrt{n}\bar{R} \\ \sqrt{n}\bar{V} \end{pmatrix}\right) = \mathrm{Var}\left(\nabla H(\theta)^\top \begin{pmatrix} R_i \\ V_i \end{pmatrix}\right) = \nabla H(\theta)^\top \, \Sigma \, \nabla H(\theta) = \tilde{\sigma}_0^2 \tag{A.18}$$

where the second equality holds by the definition of Σ and the last equality holds by elementary calculation. Therefore, by applying the multivariate delta method to (A.14), we have $\sqrt{n}(H(\bar{R}, \bar{V}) - H(\theta)) \xrightarrow{d} \mathcal{N}\left(0, \nabla H(\theta)^\top \, \Sigma \, \nabla H(\theta)\right)$, i.e.,

$$\sqrt{n}(H(\bar{R}, \bar{V}) - f(\mu))/\tilde{\sigma}_0 \xrightarrow{d} \mathcal{N}(0, 1). \tag{A.19}$$

Replacing the means, variances and covariances in $\tilde{\sigma}_0^2$ by their moment estimators, we obtain

$$\frac{1}{\bar{V}}\left[\left(\frac{\bar{R}}{2\bar{V}}\right)^2 \hat{\Sigma}_{22} + \hat{\Sigma}_{11} - \frac{\bar{R}}{\bar{V}}\hat{\Sigma}_{12}\right],$$

which equals s^2 by its definition. Due to the finiteness of $\mathbb{E}[Yh(W)], \mathbb{E}[h^2(W)], \Sigma_{11}, \Sigma_{12}, \Sigma_{22}$, we have

$$(\bar{R}, \bar{V}, \hat{\Sigma}_{11}, \hat{\Sigma}_{12}, \hat{\Sigma}_{22}) \xrightarrow{p} (\mathbb{E}[Yh(W)], \mathbb{E}[h^2(W)], \Sigma_{11}, \Sigma_{12}, \Sigma_{22})$$

by the law of large numbers. Then by the continuous mapping theorem, we have $s \xrightarrow{p} \tilde{\sigma}_0$ as $n \to \infty$. Combining this with (A.16) and (A.19), we have

$$\sqrt{n}\left(\frac{\bar{R}}{\sqrt{\bar{V}}} - f(\mu)\right)/s \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$, which establishes (A.10).

Now we will verify the positiveness of $\tilde{\sigma}_0$. Recall $\mathbb{E}[h^2(W)] = 1$ as assumed without loss of generality; we rewrite $\tilde{\sigma}_0^2$

$$\begin{aligned} \tilde{\sigma}_0^2 &= \operatorname{Var}\left(\nabla H(\theta)^\top \begin{pmatrix} R_i - \mathbb{E}[Yh(W)] \\ V_i - \mathbb{E}[h^2(W)] \end{pmatrix}\right) && \text{(A.20)} \\ &= \operatorname{Var}\left(\left(1, -\frac{\mathbb{E}[Yh(W)]}{2}\right)^\top \begin{pmatrix} R_i - \mathbb{E}[Yh(W)] \\ V_i - \mathbb{E}[h^2(W)] \end{pmatrix}\right) \\ &= \mathbb{E}\left[(R_i - \mathbb{E}[Yh(W)] - 0.5\,\mathbb{E}[Yh(W)](V_i - 1))^2\right] && \text{(A.21)} \\ &:= \mathbb{E}\left[(A+B)^2\right] \end{aligned}$$

where the first equality holds due to (A.18) and the basic property of variance, the second equality holds due to (A.17), and the last equality is by rearranging and the terms A, B are defined as below:

$$\begin{aligned} A &:= R_i - \mathbb{E}[Y_i h(W_i) \mid Z_i] = Y_i h(W_i) - \mathbb{E}[Y_i h(W_i) \mid Z_i], && \text{(A.22)} \\ B &:= \mathbb{E}[Y_i h(W_i) \mid Z_i] - \mathbb{E}[Yh(W)] - 0.5\,\mathbb{E}[Yh(W)]\,(\operatorname{Var}(h(W_i) \mid Z_i) - 1). && \text{(A.23)} \end{aligned}$$

Now we can expand (A.20) as

$$\begin{aligned} \tilde{\sigma}_0^2 = \mathbb{E}\left[(A+B)^2\right] &= \mathbb{E}\left[\mathbb{E}\left[(A+B)^2 \mid Z_i\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[A^2 \mid Z_i\right] - 2B\,\mathbb{E}[A \mid Z_i] + B^2\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[A^2 \mid Z_i\right] + B^2\right] \\ &\geq \mathbb{E}[\operatorname{Var}(Yh(W) \mid Z)], && \text{(A.24)} \end{aligned}$$

where the first equality comes from the tower property of conditional expectation, the second equality holds since $B \in \mathscr{A}(Z_i)$ and the third equality holds due to $\mathbb{E}[A \mid Z_i] = 0$.

Since (A.24) gives a lower bound for $\tilde{\sigma}_0^2$, we are done when $\mathbb{E}[\operatorname{Var}(Yh(W) \mid Z)] > 0$. Otherwise, we assume $\mathbb{E}[\operatorname{Var}(Yh(W) \mid Z)] = 0$, then $\tilde{\sigma}_0^2 = \nabla H(\theta)^\top\ \Sigma\ \nabla H(\theta) = 0$ implies the degeneracy of Σ since the vector $\nabla H(\theta) = (1, -0.5\,\mathbb{E}[Yh(W)])$ is nonzero. It suffices to show it is impossible to have Σ degenerate when $\mathbb{E}[\operatorname{Var}(Yh(W) \mid Z)] = 0$. According to the definition of Σ, we have that $Yh(W)$ is a linear function of $\operatorname{Var}(h(W) \mid Z)$ in the degenerate case. This means $Yh(W) = c\operatorname{Var}(h(W) \mid Z) + d$ for some constants c, d. Then we obtain

$$\operatorname{Var}(Yh(W) \mid Z) = \operatorname{Var}(c\operatorname{Var}(h(W) \mid Z) + d \mid Z) = c^2 \operatorname{Var}(\operatorname{Var}(h(W) \mid Z)) > 0,$$

since we are dealing with case (IV) where $\operatorname{Var}(\operatorname{Var}(h(W) \mid Z)) > 0$ and $\operatorname{Var}(Yh(W)) > 0$ (thus $c^2 > 0$). The above result contradicts the assumption $\mathbb{E}[\operatorname{Var}(Yh(W) \mid Z)] = 0$. This finishes showing the positiveness of $\tilde{\sigma}_0$, □

A.1.3 Lemma 2.3

Proof of Lemma 2.3. Recall the notations $g(z) = \mathbb{E}[\mu(X,Z) \mid Z = z]$ and $h(w) = h(x,z) = \mu(x,z) - g(z)$ introduced in (A.1) and (A.2). When $Q_x = P_{X|Z}$, we immediately have $\mu(X,Z) - \mathbb{E}_{Q_x}[\mu(X,Z) \mid Z] = \mu(X,Z) - \mathbb{E}[\mu(X,Z) \mid Z] = h(W)$, thus

$$f_{Q_y,Q_x}(\mu) = \frac{\mathbb{E}\left[(Y - \mathbb{E}_{Q_y}[Y \mid Z])h(W)\right]}{\sqrt{\mathbb{E}[h^2(W)]}} = \frac{\mathbb{E}[Yh(W)]}{\sqrt{\mathbb{E}[h^2(W)]}} = f(\mu)$$

where the second equality holds since $\mathbb{E}\left[\mathbb{E}_{Q_y}[Y \mid Z]\, h(W)\right] = 0$ by (A.4) and the last equality holds by (A.7). Hence $f_{Q_y,P_{X|Z}}(\mu) = f(\mu)$ is proved. For convenience, we also use the following notations throughout this proof: $P_x := P_{X|Z}$, $g_y(Z) := \mathbb{E}_{Q_y}[Y \mid Z]$, $g_x(Z) := \mathbb{E}_{Q_x}[\mu(X,Z) \mid Z]$. Thus we rewrite $f_{Q_y,Q_x}(\mu)$ in (2.8) as

$$f_{Q_y,Q_x}(\mu) = \frac{\mathbb{E}[(Y - g_y(Z))(\mu(X,Z) - g_x(Z))]}{\sqrt{\mathbb{E}[(\mu(X,Z) - g_x(Z))^2]}} \le \frac{\sqrt{\mathbb{E}[(Y - g_y(Z))^2]}\sqrt{\mathbb{E}[(\mu(X,Z) - g_x(Z))^2]}}{\sqrt{\mathbb{E}[(\mu(X,Z) - g_x(Z))^2]}}, \quad \text{(A.25)}$$

where the inequality holds by the Cauchy–Schwarz inequality. If $\mathbb{E}\left[(\mu(X,Z) - g_x(Z))^2\right] = 0$, $f_{Q_y,Q_x}(\mu)$ is $0/0 = 0$ by convention and thus $f_{g_y,g_x}(\mu) \le \mathcal{I} + \Delta$ automatically holds due to the non-negativeness of Δ and $\mathcal{I}$. Otherwise, we notice that

$$\begin{aligned}\mathbb{E}\left[(\mu(X,Z) - g_x(Z))^2\right] &= \mathbb{E}\left[(\mu(X,Z) - \mathbb{E}[\mu(X,Z) \mid Z] + \mathbb{E}[\mu(X,Z) \mid Z] - g_x(Z))^2\right] \\ &= \mathbb{E}\left[(\mu(X,Z) - g(Z))^2\right] + \mathbb{E}\left[(g(Z) - g_x(Z))^2\right] \\ &\ge \mathbb{E}\left[h^2(W)\right], \end{aligned} \quad \text{(A.26)}$$

where the first equality holds due to rearranging, the second equality holds since

$$\mathbb{E}[(\mu(X,Z) - \mathbb{E}[\mu(X,Z) \mid Z])(\mathbb{E}[\mu(X,Z) \mid Z] - g_x(Z))] = \mathbb{E}[h(W)(g(Z) - g_x(Z))] = 0$$

by (A.4), and the last inequality holds due to the definition of $h(w)$ and the non-negativeness of $\mathbb{E}\left[(g(Z) - g_x(Z))^2\right]$. We note that $f_{Q_y,Q_x}(\mu) \le 0 \le \mathcal{I} + \Delta$ when the numerator $\mathbb{E}[(Y - g_y(Z))(\mu(X,Z) - g_x(Z))] \le 0$. Thus it remains to deal with the case where $\mathbb{E}[(Y - g_y(Z))(\mu(X,Z) - g_x(Z))] > 0$. Now we expand $f_{Q_y,Q_x}(\mu)$ and bound it as below:

$$\begin{aligned} f_{Q_y,Q_x}(\mu) =&\ \frac{\mathbb{E}[(Y - g_y(Z))(\mu(X,Z) - g_x(Z))]}{\sqrt{\mathbb{E}[(\mu(X,Z) - g_x(Z))^2]}} \\ =&\ \frac{\mathbb{E}[(Y - g_y(Z))(\mu(X,Z) - g(Z))]}{\sqrt{\mathbb{E}[(\mu(X,Z) - g_x(Z))^2]}} + \frac{\mathbb{E}[(Y - g_y(Z))(g(Z) - g_x(Z))]}{\sqrt{\mathbb{E}[(\mu(X,Z) - g_x(Z))^2]}} \\ \le&\ \frac{\mathbb{E}[(Y - g_y(Z))(\mu(X,Z) - g(Z))]}{\sqrt{\mathbb{E}[h^2(W)]}} + \frac{\mathbb{E}[(Y - g_y(Z))(g(Z) - g_x(Z))]}{\sqrt{\mathbb{E}[(\mu(X,Z) - g_x(Z))^2]}} \\ =&\ \frac{\mathbb{E}[Yh(W))]}{\sqrt{\mathbb{E}[h^2(W)]}} + \frac{\mathbb{E}[(\epsilon(Y,W) + h^\star(W) + g^\star(Z) - g_y(Z))(g(Z) - g_x(Z))]}{\sqrt{\mathbb{E}[(\mu(X,Z) - g_x(Z))^2]}} \\ =&\ f(\mu) + \frac{\mathbb{E}[(g^\star(Z) - g_y(Z))(g(Z) - g_x(Z))]}{\sqrt{\mathbb{E}[(\mu(X,Z) - g_x(Z))^2]}}, \\ \le&\ \mathcal{I} + \frac{\mathbb{E}[|g^\star(Z) - g_y(Z)| \cdot |g(Z) - g_x(Z)|]}{\sqrt{\mathbb{E}[h^2(W)]}} \end{aligned} \quad \text{(A.27)}$$

where the first equality comes from (A.25), the second equality is by rearranging, the first inequality holds due to $\mathbb{E}[(Y - g_y(Z))(\mu(X,Z) - g_x(Z))] > 0$ and (A.26), the third equality holds since $\mathbb{E}[g_y(Z)(\mu(X,Z) - g(Z))] = \mathbb{E}[g_y(Z)h(W)] = 0$ by (A.4) and we expand Y as in (A.3), the last equality holds by (A.4), (A.5) and (A.7), and the last inequality holds due to Lemma 2.2, $\mathbb{E}[|g^\star(Z) - g_y(Z)| \cdot |g(Z) - g_x(Z)|] > 0$ and (A.26). In the following, we bound $\mathbb{E}\left[|g^\star(Z) - g_y(Z)| \cdot |g(Z) - g_x(Z)|\right]$. Since we denote $g_x(z) = \mathbb{E}_{Q_x}\left[\mu(X,Z) \mid Z = z\right]$ with Q_x being the estimate of the true conditional distribution of X given Z (i.e., $P_{X|Z}$, abbreviated as P_x), we can rewrite $|g(Z) - g_x(Z)|$ then bound it as:

$$
\begin{aligned}
|g(Z) - g_x(Z)| &= \left|\mathbb{E}_{P_x}\left[\mu(X,Z) \mid Z\right] - \mathbb{E}_{Q_x}\left[\mu(X,Z) \mid Z\right]\right| \\
&= \left|\mathbb{E}_{P_x}\left[h(W) + g(Z) \mid Z\right] - \mathbb{E}_{Q_x}\left[h(W) + g(Z) \mid Z\right]\right| \\
&= \left|\mathbb{E}_{P_x}\left[h(W) \mid Z\right] - \mathbb{E}_{Q_x}\left[h(W) \mid Z\right]\right| \\
&= \left|\int h(x,Z)(1-\delta(x,Z))dP_{X|Z}(x \mid Z)\right| \\
&= \left|\mathbb{E}_{P_x}\left[h(W)(1-\delta(W)) \mid Z\right]\right| \le \sqrt{\mathbb{E}_{P_x}\left[h^2(W) \mid Z\right]}\sqrt{\chi^2\left(Q_x \| P_{X|Z}\right)}, \qquad \text{(A.28)}
\end{aligned}
$$

where the second equality holds due to (A.2), the third equality holds since $g(Z) \in \mathscr{A}(Z)$, the fourth equality holds since Q_x is absolutely continuous with respect to $P_{X|Z}$ and we denote $\delta(x,Z) := \frac{dQ_x(x|Z)}{dP_{X|Z}(x|Z)}$ and rewrite the third line in the form of integral, and the last inequality holds by the Cauchy–Schwarz inequality and the definition of the χ^2 divergence. Hence replacing the term $|g(Z) - g_x(Z)|$ in (A.27) by its upper bound in (A.28) produces the following

$$
f_{Q_y,Q_x}(\mu) \le \mathcal{I} + \frac{\mathbb{E}\left[|g^\star(Z) - g_y(Z)|\sqrt{\mathbb{E}_{P_x}\left[h^2(W) \mid Z\right]}\sqrt{\chi^2\left(Q_x \| P_{X|Z}\right)}\right]}{\sqrt{\mathbb{E}\left[h^2(W)\right]}}. \qquad \text{(A.29)}
$$

Now we will bound $\mathrm{III} := \mathbb{E}\left[|g^\star(Z) - g_y(Z)|\sqrt{\mathbb{E}_{P_x}\left[h^2(W) \mid Z\right]}\sqrt{\chi^2\left(Q_x \| P_{X|Z}\right)}\right]$ in three different versions.

Firstly, we apply the Cauchy–Schwarz inequality to $\sqrt{\mathbb{E}_{P_x}\left[h^2(W) \mid Z\right]}\sqrt{\chi^2\left(Q_x \| P_{X|Z}\right)}$ and $|g^\star(Z) - g_y(Z)|$, producing

$$
\begin{aligned}
\mathrm{III} &= \mathbb{E}\left[|g^\star(Z) - g_y(Z)|\sqrt{\mathbb{E}_{P_x}\left[h^2(W) \mid Z\right]}\sqrt{\chi^2\left(Q_x \| P_{X|Z}\right)}\right] \\
&\le \sqrt{\mathbb{E}\left[(g^\star(Z) - g_y(Z))^2\right]}\sqrt{\mathbb{E}\left[\mathbb{E}_{P_x}\left[h^2(W) \mid Z\right]\chi^2\left(Q_x \| P_{X|Z}\right)\right]} \\
&= \sqrt{\mathbb{E}\left[(g^\star(Z) - g_y(Z))^2\right]}\sqrt{\mathbb{E}\left[\mathbb{E}_{P_x}\left[h^2(W)\chi^2\left(Q_x \| P_{X|Z}\right) \mid Z\right]\right]} \\
&= \sqrt{\mathbb{E}\left[(g^\star(Z) - g_y(Z))^2\right]}\sqrt{\mathbb{E}\left[h^2(W)\chi^2\left(Q_x \| P_{X|Z}\right)\right]}, \qquad \text{(A.30)}
\end{aligned}
$$

where the second equality holds since $\chi^2\left(Q_x \| P_{X|Z}\right) \in \mathscr{A}(Z)$, and the last equality holds due to the notation $P_x = P_{X|Z}$ and the law of total expectation. Noting the definition of III and combining (A.29) with (A.30) yields

$$
f_{Q_y,Q_x}(\mu) \le \mathcal{I} + \sqrt{\mathbb{E}\left[(g^\star(Z) - g_y(Z))^2\right]}\sqrt{\mathbb{E}\left[\left(\frac{h(W)}{\sqrt{\mathbb{E}\left[h^2(W)\right]}}\right)^2\chi^2\left(Q_x \| P_{X|Z}\right)\right]}.
$$

Recalling the notations:

$$
g^\star(Z) = \mathbb{E}\left[Y \mid Z\right], \quad g_y(Z) = \mathbb{E}_{Q_y}\left[Y \mid Z\right], \quad h(W) = \mu(X,Z) - \mathbb{E}\left[\mu(X,Z) \mid Z\right], \qquad \text{(A.31)}
$$

and $w_\mu(X,Z) = \frac{(\mu(X,Z) - \mathbb{E}[\mu(X,Z) \mid Z])^2}{\mathbb{E}[(\mu(X,Z) - \mathbb{E}[\mu(X,Z) \mid Z])^2]}$, (2.9) is thus established.

Secondly, we apply the Cauchy–Schwarz inequality to $\sqrt{\chi^2\left(Q_x \| P_{X|Z}\right)}$ and $|g^\star(Z) - g_y(Z)|\sqrt{\mathbb{E}_{P_x}\left[h^2(W) \mid Z\right]}$

in III, producing

$$
\begin{aligned}
\text{III} &= \mathbb{E}\left[|g^\star(Z) - g_y(Z)|\sqrt{\mathbb{E}_{P_x}\left[h^2(W) \mid Z\right]}\sqrt{\chi^2\left(Q_x \| P_{X|Z}\right)}\right] \\
&\leq \sqrt{\mathbb{E}\left[\chi^2\left(Q_x \| P_{X|Z}\right)\right]}\sqrt{\mathbb{E}\left[\mathbb{E}_{P_x}\left[h^2(W) \mid Z\right](g^\star(Z) - g_y(Z))^2\right]} \\
&= \sqrt{\mathbb{E}\left[\chi^2\left(Q_x \| P_{X|Z}\right)\right]}\sqrt{\mathbb{E}\left[\mathbb{E}_{P_x}\left[h^2(W)(g^\star(Z) - g_y(Z))^2 \mid Z\right]\right]} \\
&= \sqrt{\mathbb{E}\left[\chi^2\left(Q_x \| P_{X|Z}\right)\right]}\sqrt{\mathbb{E}\left[h^2(W)(g^\star(Z) - g_y(Z))^2\right]}, \qquad \text{(A.32)}
\end{aligned}
$$

where the second equality holds since $(g^\star(Z) - g_y(Z))^2 \in \mathscr{A}(Z)$, and the last equality holds due to the notation $P_x = P_{X|Z}$ and the law of total expectation. Combining (A.29) and (A.31) with (A.32) and recalling $w_\mu(X, Z) = \frac{(\mu(X,Z) - \mathbb{E}[\mu(X,Z) \mid Z])^2}{\mathbb{E}[(\mu(X,Z) - \mathbb{E}[\mu(X,Z) \mid Z])^2]} = \frac{h^2(W)}{\mathbb{E}[h^2(W)]}$ yields a different bound on $f_{Q_y,Q_x}(\mu)$, namely,

$$
\begin{aligned}
&f_{Q_y,Q_x}(\mu) \leq f(\mu) + \Delta', \text{ where} \\
\Delta' &= \sqrt{\mathbb{E}\left[\chi^2\left(Q_x \| P_{X|Z}\right)\right]}\sqrt{\mathbb{E}\left[w_\mu(X, Z)(\mathbb{E}\left[Y \mid Z\right] - \mathbb{E}_{Q_y}\left[Y \mid Z\right])^2\right]}.
\end{aligned} \qquad \text{(A.33)}
$$

Lastly, we apply the Cauchy–Schwarz inequality to $\sqrt{\mathbb{E}_{P_x}\left[h^2(W) \mid Z\right]}$ and $|g^\star(Z) - g_y(Z)|\sqrt{\chi^2\left(Q_x \| P_{X|Z}\right)}$ in III, producing

$$
\begin{aligned}
\text{III} &= \mathbb{E}\left[|g^\star(Z) - g_y(Z)|\sqrt{\mathbb{E}_{P_x}\left[h^2(W) \mid Z\right]}\sqrt{\chi^2\left(Q_x \| P_{X|Z}\right)}\right] \\
&\leq \sqrt{\mathbb{E}\left[\mathbb{E}_{P_x}\left[h^2(W) \mid Z\right]\right]}\sqrt{\mathbb{E}\left[\chi^2\left(Q_x \| P_{X|Z}\right)(g^\star(Z) - g_y(Z))^2\right]} \\
&= \sqrt{\mathbb{E}\left[h^2(W)\right]}\sqrt{\mathbb{E}\left[\chi^2\left(Q_x \| P_{X|Z}\right)(g^\star(Z) - g_y(Z))^2\right]} \\
&= \sqrt{\mathbb{E}\left[h^2(W)\right]}\left(\mathbb{E}\left[(g^\star(Z) - g_y(Z))^4\right]\right)^{1/4}\left(\mathbb{E}\left[\left(\chi^2\left(Q_x \| P_{X|Z}\right)\right)^2\right]\right)^{1/4}, \qquad \text{(A.34)}
\end{aligned}
$$

where the second equality holds due to the notation $P_x = P_{X|Z}$ and the law of total expectation, and the last inequality holds by applying the Cauchy–Schwarz inequality again. Combining (A.29) and (A.31) with (A.34) yields a final different bound on $f_{Q_y,Q_x}(\mu)$, namely,

$$
\begin{aligned}
&f_{Q_y,Q_x}(\mu) \leq f(\mu) + \Delta'', \text{ where} \\
\Delta'' &= \left(\mathbb{E}\left[(\mathbb{E}\left[Y \mid Z\right] - \mathbb{E}_{Q_y}\left[Y \mid Z\right])^4\right]\right)^{1/4}\left(\mathbb{E}\left[\left(\chi^2\left(Q_x \| P_{X|Z}\right)\right)^2\right]\right)^{1/4}.
\end{aligned} \qquad \text{(A.35)}
$$

□

A.2 Proofs in Section 2.3

Proof of Theorem 2.4. We prove by contradiction. Suppose there exists an upper confidence bound procedure ensuring asymptotic coverage such that (2.10) holds, that is, there exists a joint law over (Y, X, Z), denoted by $F_\infty \in \mathcal{F}$ such that

$$
\limsup_{n\to\infty} \mathbb{P}_\infty\left(U(D_n) - \mathcal{I}^2_{F_\infty} < \mathbb{E}_\infty\left[\mathrm{Var}_\infty\left(Y \mid X, Z\right)\right]\right) > \alpha. \qquad \text{(A.36)}
$$

where $\mathbb{P}_\infty$, $\mathbb{E}_\infty$, Var_∞ denote that the data generating distribution for i.i.d. sample D_n is F_∞. Note that $\mathbb{P}_\infty\left(U(D_n) - \mathcal{I}^2_{F_\infty} < \mathbb{E}_\infty\left[\mathrm{Var}_\infty\left(Y \mid X, Z\right)\right]\right) = \mathbb{P}_\infty\left(U(D_n) < \mathbb{E}_\infty\left[\mathrm{Var}_\infty\left(Y \mid Z\right)\right]\right)$ by the definition of $\mathcal{I}^2_{F_\infty}$.

Let $\lambda_1 = \mathbb{E}_\infty\left[\mathrm{Var}_\infty\left(Y \mid Z\right)\right]$. When $\lambda_1 = 0$, we have $\mathbb{E}_\infty\left[\mathrm{Var}_\infty\left(Y \mid Z\right)\right] = \mathbb{E}_\infty\left[\mathrm{Var}_\infty\left(Y \mid X, Z\right)\right] = \mathcal{I}^2_{F_\infty} = 0$ and immediately show

$$\limsup_{n\to\infty} \mathbb{P}_\infty\left(U(D_n) - \mathcal{I}^2_{F_\infty} < \mathbb{E}_\infty\left[\mathrm{Var}_\infty\left(Y \mid X, Z\right)\right]\right) = \limsup_{n\to\infty} \mathbb{P}_\infty\left(U(D_n) < \mathcal{I}^2_{F_\infty}\right) \leq \alpha,$$

which contradicts (A.36). In the following we consider the case where $\lambda_1 > 0$. Now we construct a sequence of joint laws over (Y, X, Z), denoted by $\{F_k\}_{k=1}^\infty$, $F_k \in \mathcal{F}$, such that the conditional distribution of $\epsilon \mid X, Z$ is the same as that under F_∞, where $\epsilon = Y - \mathbb{E}\left[Y \mid X, Z\right]$, that is,

$$\mathbb{P}_k\left(\epsilon \mid X, Z\right) = \mathbb{P}_\infty\left(\epsilon \mid X, Z\right), \qquad \forall\ k \geq 1 \tag{A.37}$$

and there exist Borel sets $A_k \in \mathbb{R}^{p-1}$ satisfying the following:

(a) $\mathbb{P}_k\left(Z \in A_k\right) = 1/k$;

(b) $\mathbb{P}_k\left(Y \mid X, Z\right) = \mathbb{P}_\infty\left(Y \mid X, Z\right)$ when $Z \notin A_k$;

(c) $\mathbb{E}_k\left[\mu^\star_k(X, Z) \mid Z\right] = \mathbb{E}_\infty\left[\mu^\star_\infty(X, Z) \mid Z\right]$ when $Z \in A_k$;

(d) $\mathrm{Var}_k\left(\mu^\star_k(X, Z) \mid Z\right) = \mathrm{Var}_\infty\left(\mu^\star_\infty(X, Z) \mid Z\right) + k\left(2\lambda_1 - \mathcal{I}^2_{F_\infty}\right)$ when $Z \in A_k$;

where $\mathbb{P}_k$, $\mathbb{E}_k$, Var_k denote that the data generating distribution for i.i.d. sample D_n is F_k, and $\mu^\star_k(X, Z) := \mathbb{E}_k\left[Y \mid X, Z\right], \mu^\star_\infty(X, Z) := \mathbb{E}_\infty\left[Y \mid X, Z\right]$. According to the statement of Theorem 2.4, the covariate distribution $P_{X,Z}$ is continuous and fixed. Therefore we have (a) is possible and immediately know

$$\mathbb{P}_k\left(X, Z\right) = \mathbb{P}_\infty\left(X, Z\right), \qquad \forall\ k \geq 1. \tag{A.38}$$

Note here $\mathbb{E}_k\left[\cdot \mid Z\right], \mathrm{Var}_k\left(\cdot \mid Z\right)$ are the same as $\mathbb{E}_\infty\left[\cdot \mid Z\right], \mathrm{Var}_\infty\left(\cdot \mid Z\right)$ due to (A.38). Hence we can calculate $\mathcal{I}_{F_k}$ through the following

$$\begin{aligned} \mathcal{I}^2_{F_k} - \mathcal{I}^2_{F_\infty} &= \mathbb{E}_\infty\left[\mathbb{1}_{\{A_k\}}\left(\mathrm{Var}_\infty\left(\mu^\star_k(X, Z) \mid Z\right) - \mathrm{Var}_\infty\left(\mu^\star_\infty(X, Z) \mid Z\right)\right)\right] \\ &= \mathbb{E}_\infty\left[\mathbb{1}_{\{A_k\}} k\left(2\lambda_1 - \mathcal{I}^2_{F_\infty}\right)\right] \\ &= 2\lambda_1 - \mathcal{I}^2_{F_\infty} =: \lambda_2, \end{aligned} \tag{A.39}$$

where the first equality comes from the definition of $\mathcal{I}^2_F$, (A.38) and (b), the second equality holds due to (d) and the third equality holds due to (a). Therefore $\mathcal{I}^2_{F_k} = 2\lambda_1$. We should also check whether F_k belongs to $\mathcal{F}$. Indeed, we consider the following

$$\begin{aligned} \mathrm{Var}_k\left(Y\right) &= \mathbb{E}_k\left[\mathrm{Var}_k\left(Y \mid X, Z\right)\right] + \mathrm{Var}_k\left(\mathbb{E}_k\left[Y \mid X, Z\right]\right) \\ &= \mathbb{E}_k\left[\mathrm{Var}_k\left(\epsilon \mid X, Z\right)\right] + \mathrm{Var}_k\left(\mathbb{E}_k\left[Y \mid Z\right]\right) + \mathcal{I}^2_{F_k} \\ &= \mathbb{E}_\infty\left[\mathrm{Var}_k\left(\epsilon \mid X, Z\right)\right] + \mathrm{Var}_\infty\left(\mathbb{E}_k\left[Y \mid Z\right]\right) + \mathcal{I}^2_{F_k} \\ &= \mathbb{E}_\infty\left[\mathrm{Var}_\infty\left(\epsilon \mid X, Z\right)\right] + \mathrm{Var}_\infty\left(\mathbb{E}_\infty\left[Y \mid Z\right]\right) + \mathcal{I}^2_{F_k} \\ &= \mathbb{E}_\infty\left[\mathrm{Var}_\infty\left(\epsilon \mid X, Z\right)\right] + \mathrm{Var}_\infty\left(\mathbb{E}_\infty\left[Y \mid Z\right]\right) + \mathcal{I}^2_{F_\infty} + \lambda_2 \\ &= \mathrm{Var}_\infty\left(Y\right) + \lambda_2 < \infty, \end{aligned}$$

where the first equality comes from the law of total variance, the second equality holds as a result of the decomposition $Y = \mu^\star(X, Z) + \epsilon$ and the equivalent expression of the mMSE gap (2.2), the third equality holds due to (A.38), the fourth equality holds due to (A.37), (b) and (c), the fifth equality comes from (A.39). Thus we verify $F_k \in \mathcal{F},\ \forall\ k \geq 1$. As the upper confidence bound procedure U ensures asymptotic coverage validity and $\mathcal{I}^2_{F_k} = 2\lambda_1$, we have

$$\mathbb{P}_k\left(U(D_n) \geq 2\lambda_1\right) \geq 1 - \alpha + o_k(1) \tag{A.40}$$

where the subscript in $o_k(1)$ emphasizes that the convergence is with respect to data generating function F_k. Remark we only require for fixed k, $o_k(1) \to 0$ as $n \to \infty$. Also notice the following

$$|\mathbb{P}_\infty (U(D_n) \geq 2\lambda_1) - \mathbb{P}_k (U(D_n) \geq 2\lambda_1)| \leq d_{TV}(F_k, F_\infty) \leq \frac{1}{k}, \quad \forall\, k \geq 1, \tag{A.41}$$

where the first inequality comes from the property of total variation distance and the second equality holds as a result of (a), according to the construction of F_k. Combining (A.40) and (A.41) yields the following

$$\mathbb{P}_\infty (U(D_n) \geq 2\lambda_1) \geq 1 - \alpha - 1/k + o_k(1), \quad \forall\, k \geq 1.$$

First let $n \to \infty$ then send k to infinity, we obtain

$$\liminf_{n\to\infty} \mathbb{P}_\infty (U(D_n) \geq 2\lambda_1) \geq 1 - \alpha,$$

which contradicts

$$\limsup_{n\to\infty} \mathbb{P}_\infty (U(D_n) < \mathbb{E}_\infty [\mathrm{Var}_\infty (Y \mid Z)] = \lambda_1) > \alpha.$$

□

A.3 Proofs in Section 2.4

Proof of Theorem 2.5. As in the proof of Theorem 2.3, we immediately have coverage validity when $\mu(X, Z) \in \mathscr{A}(Z)$. Otherwise, it suffices to show

$$\mathbb{P}\left(\frac{\bar{R}}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n}} \leq f(\mu)\right) \geq 1 - \alpha - o(1). \tag{A.42}$$

for any given $K > 1$, where the sample mean $(\bar{R}, \bar{V})$ and sample covariance matrix $\hat{\Sigma}$ are defined the same way as in Algorithm 1 except that R_i, V_i are replaced by their Monte Carlo estimators R_i^K, V_i^K as defined below.

$$\begin{aligned} R_i^K &:= Y_i \left(\mu(X_i, Z_i) - \frac{1}{K}\sum_{k=1}^K \mu(X_i^{(k)}, Z_i)\right), \\ V_i^K &:= \frac{1}{K-1}\sum_{k=1}^K \left(\mu(X_i^{(k)}, Z_i) - \frac{1}{K}\sum_{k=1}^K \mu(X_i^{(k)}, Z_i)\right)^2, \end{aligned} \tag{A.43}$$

for any fixed $K > 1$.

First we verify

$$\mathbb{E}\left[R_i^K\right] = \mathbb{E}\left[Y h(W)\right], \quad \mathbb{E}\left[V_i^K\right] = \mathbb{E}\left[h^2(W)\right]. \tag{A.44}$$

By the construction of the null samples, $X_i^{(k)}$ satisfy the following properties:

$$\{X_i^{(k)}\}_{k=1}^K \perp\!\!\!\perp (X_i, Y_i) \mid Z_i, \tag{A.45}$$

$$\{X_i^{(k)}\}_{k=1}^K \mid Z_i \overset{i.i.d.}{\sim} X_i \mid Z_i, \tag{A.46}$$

thus we have

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^K \mu(\tilde{X}_i^{(k)}, Z_i) \,\middle|\, Z_i\right] = \mathbb{E}\left[\mu(X_i, Z_i) \mid Z_i\right], \tag{A.47}$$

$$\mathbb{E}\left[\frac{1}{K-1}\sum_{k=1}^K \left(\mu(\tilde{X}_i^{(k)}, Z_i) - \frac{1}{K}\sum_{k=1}^K \mu(\tilde{X}_i^{(k)}, Z_i)\right)^2 \,\middle|\, Z_i\right] = \mathrm{Var}\left(\mu(X_i, Z_i) \mid Z_i\right), \tag{A.48}$$

and further obtain

$$
\begin{aligned}
\mathbb{E}\left[R_i^K\right] &= \mathbb{E}\left[Y_i\left(\mu(X_i, Z_i) - \frac{1}{K}\sum_{k=1}^{K}\mu(\tilde{X}_i^{(k)}, Z_i)\right)\right] \\
&= \mathbb{E}\left[Y_i\mu(W_i)\right] - \mathbb{E}\left[\mathbb{E}\left[Y_i \mid Z_i\right]\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\mu(\tilde{X}_i^{(k)}, Z_i)\,\middle|\, Z_i\right]\right] \\
&= \mathbb{E}\left[Y_i\mu(W_i)\right] - \mathbb{E}\left[\mathbb{E}\left[Y_i \mid Z_i\right]\mathbb{E}\left[\mu(X_i, Z_i) \mid Z_i\right]\right] \\
&= \mathbb{E}\left[Y_i\mu(W_i)\right] - \mathbb{E}\left[Y_i\mathbb{E}\left[\mu(X_i, Z_i) \mid Z_i\right]\right] = \mathbb{E}\left[Yh(W)\right],
\end{aligned}
$$

where the first equality holds due to (A.43), the second equality holds due to (A.45), the third equality holds due to (A.47), the fourth equality comes from the tower property of total expectation and the last one is by the definition of $h(W)$. Regarding the term $\mathbb{E}\left[V_i^K\right]$, (A.48) and (A.5) immediately imply $\mathbb{E}\left[V_i^K\right] = \mathbb{E}\left[h^2(W)\right]$.

To prove (A.42), we can follow a similar strategy as in the proof of Theorem 2.3. Note Appendix A.1.2 considers 4 different cases then deals with them separately. Essentially we can conduct similar analysis, but to avoid lengthy proof, we focus on the most complicated case where $\mathrm{Var}\left(Yh(W)\right) > 0$ and $\mathrm{Var}\left(\mathrm{Var}\left(h(X) \mid Z\right)\right) > 0$ and omit the derivations for the other three cases. Under the moment conditions $\mathbb{E}\left[Y^4\right], \mathbb{E}\left[h^4(W)\right] < \infty$, we have $\mathbb{E}\left[R_i^K\right] = \mathbb{E}\left[Yh(W)\right] < \infty$ and $\mathbb{E}\left[V_i^K\right] = \mathbb{E}\left[h^2(W)\right] < \infty$.

By applying the multivariate central limit theorem and the delta method, we obtain the following asymptotic normality result as in the proof of Theorem 2.3: as $n \to \infty$,

$$
\sqrt{n}\left(\frac{\frac{1}{n}\sum_{i=1}^{n} R_i^K}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} V_i^K}} - f(\mu)\right) \xrightarrow{d} \mathcal{N}\left(0, \tilde{\sigma}_0^2\right), \tag{A.49}
$$

where $\tilde{\sigma}_0^2$ is similarly defined as in (A.15) and its positiveness will be proved over the course of derivations from (A.59) toward the end of this proof. Due to the law of large numbers and the continuous mapping theorem, we can prove $s \xrightarrow{p} \tilde{\sigma}_0$ as in Appendix A.1.2. The asymptotic normality and the consistency result only require us to verify the finiteness of $\Sigma_{11} = \mathrm{Var}\left(R_i^K\right), \Sigma_{12} = \mathrm{Cov}\left(R_i^K, V_i^K\right), \Sigma_{22} = \mathrm{Var}\left(V_i^K\right)$. Since $\mathbb{E}\left[R_i^K\right], \mathbb{E}\left[V_i^K\right] < \infty$ under the stated moment conditions and $\mathrm{Cov}\left(R_i^K, V_i^K\right) \le \sqrt{\mathrm{Var}\left(R_i^K\right)\mathrm{Var}\left(V_i^K\right)}$ by the Cauchy–Schwarz inequality, it suffices to prove

$$
\mathbb{E}\left[|R_i^K|^2\right] < \infty, \mathbb{E}\left[|V_i^K|^2\right] < \infty. \tag{A.50}
$$

Denote $\bar{h}_i^K = \frac{1}{K}\sum_{k=1}^{K} h(\tilde{X}_i^{(k)}, Z_i)$ and we rewrite R_i^K and V_i^K.

$$
\begin{aligned}
R_i^K &= Y_i\left(\mu(X_i, Z_i) - \frac{1}{K}\sum_{k=1}^{K}\mu(\tilde{X}_i^{(k)}, Z_i)\right) \\
&= Y_i\left(\mu(X_i, Z_i) - \mathbb{E}\left[\mu(X_i, Z_i) \mid Z_i\right] - \frac{1}{K}\sum_{k=1}^{K}(\mu(\tilde{X}_i^{(k)}, Z_i) - \mathbb{E}\left[\mu(\tilde{X}_i^{(k)}, Z_i) \mid Z_i\right])\right) \\
&= Y_i\left(h(X_i, Z_i) - \frac{1}{K}\sum_{k=1}^{K} h(\tilde{X}_i^{(k)}, Z_i)\right) \qquad (A.51) \\
&= Y_i(h(X_i, Z_i) - \bar{h}_i^K) \qquad (A.52)
\end{aligned}
$$

where the first equality holds by (A.43), the second equality holds by (A.47) and the third equality holds by the definition of $h(w)$.

$$
\begin{aligned}
V_i^K &= \frac{1}{K-1}\sum_{k=1}^K\left(\mu(X_i^{(k)},Z_i)-\frac{1}{K}\sum_{k=1}^K\mu(X_i^{(k)},Z_i)\right)^2 \\
&= \frac{1}{K-1}\sum_{k=1}^K\left(h(X_i^{(k)},Z_i)-\frac{1}{K}\sum_{k=1}^K h(X_i^{(k)},Z_i)\right)^2 \\
&= \frac{1}{K-1}\sum_{k=1}^K h^2(X_i^{(k)},Z_i)-\frac{K}{K-1}\left(\frac{1}{K}\sum_{k=1}^K h(X_i^{(k)},Z_i)\right)^2 \qquad \text{(A.53)} \\
&= \frac{K}{K-1}\left(\frac{1}{K}\sum_{k=1}^K h^2(X_i^{(k)},Z_i)-(\bar{h}_i^K)^2\right) \qquad \text{(A.54)}
\end{aligned}
$$

where the first equality holds by (A.43), the second equality holds due to similar derivations as (A.51) and the last two equalities are simply by expanding and rearranging. Now we bound

$$
\begin{aligned}
\left(\mathbb{E}\left[|R_i^K|^2\right]\right)^2 &= \left(\mathbb{E}\left[Y_i^2(h(X_i,Z_i)-\bar{h}_i^K)^2\right]\right)^2 \\
&\leq \mathbb{E}\left[Y^4\right]\mathbb{E}\left[(h(X_i,Z_i)-\bar{h}_i^K)^4\right] \\
&\leq \mathbb{E}\left[Y^4\right]\cdot 2^{4-1}\left(\mathbb{E}\left[h^4(X_i,Z_i)+\mathbb{E}\left[\left(\bar{h}_i^K\right)^4\right]\right]\right) \qquad \text{(A.55)}
\end{aligned}
$$

where the first equality holds due to (A.52), the first inequality holds by the Cauchy–Schwarz inequality, the second inequality comes from the C_r inequality. Regarding $\mathbb{E}\left[|V_i^K|^2\right]$, we have

$$
\begin{aligned}
\mathbb{E}\left[|V_i^K|^2\right] =\; & \mathbb{E}\left[\left|\frac{K}{K-1}\left(\frac{1}{K}\sum_{k=1}^K h^2(X_i^{(k)},Z_i)-(\bar{h}_i^K)^2\right)\right|^2\right] \\
\leq\; & \frac{2^{2-1}K^2}{(K-1)^2}\mathbb{E}\left[\left(\frac{1}{K}\sum_{k=1}^K h^2(X_i^{(k)},Z_i)\right)^2\right]+\frac{2^{2-1}K^2}{(K-1)^2}\mathbb{E}\left[(\bar{h}_i^K)^4\right] \\
\leq\; & 2^3\left(\mathbb{E}\left[\left(\frac{1}{K}\sum_{k=1}^K h^2(X_i^{(k)},Z_i)\right)^2\right]+\mathbb{E}\left[(\bar{h}_i^K)^4\right]\right) := 2^3(\mathrm{II}+\mathbb{E}\left[(\bar{h}_i^K)^4\right]), \qquad \text{(A.56)}
\end{aligned}
$$

where the first equality holds by (A.54), the first inequality holds due to the C_r inequality, and the second inequality comes from rearranging and the fact that $K \leq 2(K-1)$ (since $K>1$). The term II and $\mathbb{E}\left[(\bar{h}_i^K)^4\right]$ can be bounded using the same strategy. Below we give the bounding details of $\mathbb{E}\left[(\bar{h}_i^K)^4\right]$ and omit that of II. By the tower property of conditional expectation, we have

$$
\begin{aligned}
\mathbb{E}\left[(\bar{h}_i^K)^4\right] &= \mathbb{E}\left[\left(\frac{1}{K}\sum_{k=1}^K h(\tilde{X}_i^{(k)},Z_i)\right)^4\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{K}\sum_{k=1}^K h(\tilde{X}_i^{(k)},Z_i)\right)^4\middle| Z_i\right]\right]. \qquad \text{(A.57)}
\end{aligned}
$$

To bound $\mathbb{E}\left[\left(\frac{1}{K}\sum_{k=1}^K h(\tilde{X}_i^{(k)},Z_i)\right)^4\middle| Z_i\right]$, we notice that, conditional on Z_i, $\{h(\tilde{X}_i^{(k)},Z_i)\}_{k=1}^K$ are i.i.d. mean zero random variables, hence we can apply the extension of the Bahr–Esseen inequality in Dharmadhikari et al. (1969) to obtain

$$
\mathbb{E}\left[\left(\sum_{k=1}^K h(\tilde{X}_i^{(k)},Z_i)\right)^4\middle| Z_i\right] \leq c_{4,K}\sum_{k=1}^K\mathbb{E}\left[h^4(\tilde{X}_i^{(k)},Z_i)\middle| Z_i\right], \qquad \text{(A.58)}
$$

Note for generic $d \geq 2$ and n, the term $c_{d,n}$ is defined as

$$c_{d,n} = n^{d/2-1}\frac{d(d-1)}{2}\max\{1, 2^{d-3}\}\left[1 + 2d^{-1}D_{2m}^{(d-2)/2m}\right]$$

where the integer m satisfies $2m \leq d < 2m+2$, and

$$D_{2m} = \sum_{t=1}^{m}\frac{t^{2m-1}}{(t-1)!}.$$

We then can simply bound $c_{4,K}$ by $C_4 K$ for some universal constant C_4 which do not depend on K. Therefore, combining (A.57) and (A.58) gives us

$$\begin{aligned}\mathbb{E}\left[(\bar{h}_i^K)^4\right] &\leq \mathbb{E}\left[\frac{C_4 K}{K^4}\sum_{k=1}^{K}\mathbb{E}\left[h^4(\tilde{X}_i^{(k)}, Z_i)\,\middle|\, Z_i\right]\right] \\ &= \frac{C_4}{K^2}\mathbb{E}\left[\mathbb{E}\left[h^4(X_i, Z_i)\,|\, Z_i\right]\right] = \frac{C_4}{K^2}\mathbb{E}\left[h^4(W)\right]\end{aligned}$$

where the equality holds by (A.46) and the second equality holds by the tower property of conditional expectation. Since $\mathbb{E}\left[h^4(W)\right] < \infty$, we have $\mathbb{E}\left[(\bar{h}_i^K)^4\right] < \infty$. The finiteness of II is similarly proved. Due to (A.55) and (A.56), we thus establish (A.50) under the stated moment conditions $\mathbb{E}\left[Y^4\right], \mathbb{E}\left[h^4(W)\right] < \infty$. Applying Slutsky's theorem to (A.49) and the consistency result that $s \xrightarrow{p} \tilde{\sigma}_0$, we have

$$\frac{\sqrt{n}}{s}\left(\frac{\frac{1}{n}\sum_{i=1}^{n} R_i^K}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} V_i^K}} - f(\mu)\right) \xrightarrow{d} \mathcal{N}(0,1),$$

which establishes (A.42).

Now we will verify the positiveness of $\tilde{\sigma}_0$ as promised. Recall in the proof of Theorem 2.3, the variance term in the asymptotic normality result is also denoted as $\tilde{\sigma}_0^2$ and admits the following expression

$$\mathbb{E}\left[(R_i - \mathbb{E}\left[Yh(W)\right] - 0.5\,\mathbb{E}\left[Yh(W)\right](V_i - 1))^2\right] = \mathbb{E}\left[(A+B)^2\right] > 0 \tag{A.59}$$

according to (A.21), where A and B are defined in (A.22) and (A.23) and $\mathbb{E}\left[(A+B)^2\right] > 0$ as proved over the course of derivations from (A.21) to the end of the proof of Theorem 2.3. In this proof, it is not hard to see $\tilde{\sigma}_0^2$ has a similar form except that R_i, V_i in the above expression are replaced by their Monte Carlo estimators R_i^K, V_i^K, thus giving

$$\begin{aligned}\tilde{\sigma}_0^2 &= \mathbb{E}\left[(R_i^K - \mathbb{E}\left[Yh(W)\right] - 0.5\,\mathbb{E}\left[Yh(W)\right](V_i^K - 1))^2\right] \\ &= \mathbb{E}\left[(Y_i(h(X_i, Z_i) - \bar{h}_i^K) - \mathbb{E}\left[Yh(W)\right] - 0.5\,\mathbb{E}\left[Yh(W)\right](V_i^K - 1))^2\right] \\ &= \mathbb{E}\left[(\mathrm{III}_1 - \mathrm{III}_2)^2\right],\end{aligned} \tag{A.60}$$

where the second equality holds by (A.52) and rearranging, the terms $\mathrm{III}_1, \mathrm{III}_2$ in the last equality are defined as:

$$\begin{aligned}\mathrm{III}_1 &:= Y_i h(W_i) - \mathbb{E}\left[Yh(W)\right] - 0.5\,\mathbb{E}\left[Yh(W)\right](\mathrm{Var}\left(h(W_i)\,|\,Z_i\right) - 1) \\ \mathrm{III}_2 &:= Y_i\bar{h}_i^K + 0.5\,\mathbb{E}\left[Yh(W)\right](V_i^K - \mathrm{Var}\left(h(W_i)\,|\,Z_i\right)).\end{aligned}$$

To bound $\mathbb{E}\left[(\mathrm{III}_1 - \mathrm{III}_2)^2\right]$, we will show $\mathbb{E}\left[\mathrm{III}_2\,|\,Y_i, W_i\right] = 0$. Recall the definition that $\bar{h}_i^K = \frac{1}{K}\sum_{k=1}^{K} h(\tilde{X}_i^{(k)}, Z_i)$, we obtain

$$\mathbb{E}\left[\bar{h}_i^K\,|\,Y_i, W_i\right] = \mathbb{E}\left[\bar{h}_i^K\,|\,Z_i\right] = \mathbb{E}\left[h(W_i)\,|\,Z_i\right] = 0,$$

where the first equality holds due to $W_i = (X_1, Z_i)$ and (A.45), the second equality holds by (A.46), and the last equality holds due to (A.4). Similarly we have

$$\mathbb{E}\left[V_i^K \mid Y_i, W_i\right] = \mathbb{E}\left[V_i^K \mid Z_i\right] = \mathrm{Var}\left(h(W_i) \mid Z_i\right),$$

due to (A.43), (A.45), and (A.47). Thus we have shown

$$\mathbb{E}\left[\mathrm{III}_2 \mid Y_i, W_i\right] = 0. \tag{A.61}$$

Applying the tower property of conditional expectation to (A.60) then expanding yields the following expression:

$$\begin{aligned} \tilde{\sigma}_0^2 &= \mathbb{E}\left[\mathbb{E}\left[\left(\mathrm{III}_1^2 + \mathrm{III}_2^2 - 2\mathrm{III}_1\mathrm{III}_2\right) \mid Y_i, W_i\right]\right] \\ &= \mathbb{E}\left[\mathrm{III}_1^2 + \mathbb{E}\left[\mathrm{III}_2^2 \mid Y_i, W_i\right] - 2\mathrm{III}_1\mathbb{E}\left[\mathrm{III}_2 \mid Y_i, W_i\right]\right] \\ &= \mathbb{E}\left[\mathrm{III}_1^2 + \mathbb{E}\left[\mathrm{III}_2^2 \mid Y_i, W_i\right]\right] \\ &\geq \mathbb{E}\left[\mathrm{III}_1^2\right] = \mathbb{E}\left[(A+B)^2\right], \end{aligned} \tag{A.62}$$

where the second equality holds since $\mathrm{III}_1 \in \mathscr{A}(Y_i, W_i)$, and the third equality comes from (A.61). Note in the last line we have $\mathrm{III}_1 = A + B$ due to the definitions of A, B in (A.22) and (A.23) and $\mathbb{E}\left[(A+B)^2\right] > 0$ due to (A.59). Note $\mathbb{E}\left[(A+B)^2\right]$ does not depend on K, therefore we establish the positiveness of $\tilde{\sigma}_0$ for any $K > 1$. □

A.4 Proofs in Section 2.5

Proof of Theorem 2.6. First we write

$$\mathcal{I} - L_\alpha^n(\mu_n) = \mathcal{I} - f(\mu_n) + f(\mu_n) - L_\alpha^n(\mu_n),$$

where $f(\mu_n)$ is defined as

$$f(\mu_n) =: \frac{\mathbb{E}\left[\mathrm{Cov}(\mu^\star(X,Z), \mu_n(X,Z) \mid Z)\right]}{\sqrt{\mathbb{E}\left[\mathrm{Var}(\mu_n(X,Z) \mid Z)\right]}}.$$

Then it suffices to separately show

$$\mathcal{I} - f(\mu_n) = O_p\left(\inf_{\mu' \in S_{\mu_n}} \mathbb{E}\left[(\mu_n'(X,Z) - \mu^\star(X,Z))^2\right]\right), \tag{A.63}$$

$$f(\mu_n) - L_\alpha^n(\mu_n) = O_p\left(n^{-1/2}\right). \tag{A.64}$$

In the following, we first show (A.64). Recall the definitions in Algorithm 1, when $\mu(X,Z) \in \mathscr{A}(Z)$, we have $f(\mu_n) = L_\alpha^n(\mu_n) = 0$, hence in the following we focus on the case where $\mu(X,Z) \notin \mathscr{A}(Z)$. Note we have

$$L_\alpha^n(\mu_n) \geq \frac{\bar{R}}{\sqrt{V}} - \frac{z_\alpha s}{\sqrt{n}},$$

then since $f(\mu_n) - L_\alpha^n(\mu_n) \leq s\left(\left|\left(\frac{\bar{R}}{\sqrt{V}} - f(\mu_n)\right)/s\right| + \frac{z_\alpha}{\sqrt{n}}\right)$, it suffices to show

$$T := \frac{\bar{R}/\sqrt{V} - f(\mu_n)}{s} = O_p\left(n^{-1/2}\right), \quad s = O_p(1).$$

For given μ_n, showing the above is quite straightforward: in the proof of Theorem 2.3, we establish the asymptotic normality of T; we also show s converges in probability to $\tilde{\sigma}_0$ (which is the variance of the asymptotic normal distribution, as defined in (A.15)). For a sequence of working regression functions μ_n, we need more work and the stated uniform moment conditions. The proof proceeds through verifying the following: note that by definition of bounded in probability, $T = O_p\left(n^{-1/2}\right)$ says for any $\epsilon > 0$, there exists M for which

$$\sup_n P(\sqrt{n}|T| > M) \leq \epsilon.$$

The case that $\mu(X,Z) \in \mathscr{A}(Z)$, i.e., $\mathbb{E}\left[\mathrm{Var}\left(\mu_n(X,Z) \mid Z\right)\right] = 0$, was dealt with in the first sentence after (A.64). Now it suffices to show for any μ_n in the function class $\mathcal{U} := \{\mu : \mathbb{E}\left[\mu^{12}(X,Z)\right] / (\mathbb{E}\left[\mathrm{Var}\left(\mu(X,Z) \mid Z\right)\right])^6 \leq C\}$,

$$\sup_n \mathbb{P}\left(\sqrt{n}|T| > M\right) \leq \epsilon, \tag{A.65}$$

and the choice of M (when fixing ϵ) is uniform over $\mu_n \in \mathcal{U}$. Define the standard Gaussian random variable by G. Then we have

$$\mathbb{P}\left(\sqrt{n}|T| > M\right) \leq \mathbb{P}\left(|G| > M\right) + \Delta, \tag{A.66}$$

where Δ is defined as

$$\Delta := \sup_{\mu_n \in \mathcal{U}} \sup_{M>0} \left|\mathbb{P}\left(\sqrt{n}|T| > M\right) - \mathbb{P}\left(|G| > M\right)\right|. \tag{A.67}$$

Due to (A.9), $\mathbb{E}\left[\mu^{12}(X,Z)\right] < \infty$ implies $\mathbb{E}\left[h^{12}(W)\right] < \infty$, where h is defined in (A.2). In the following proof, we will only assume weaker moment conditions, i.e., $\mathbb{E}\left[\mathrm{Var}\left(\mu_n(X,Z) \mid Z\right)\right] = 0$ or $\frac{\mathbb{E}\left[\mu_n^{12}(X,Z)\right]}{\mathbb{E}\left[\mathrm{Var}\left(\mu_n(X,Z) \mid Z\right)\right]^6} \leq C$ stated in Theorem 2.6 is replaced by $\mathbb{E}\left[\mathrm{Var}\left(\mu_n(X,Z) \mid Z\right)\right] = 0$ or $\frac{\mathbb{E}\left[h_n^{12}(X,Z)\right]}{\mathbb{E}\left[\mathrm{Var}\left(\mu_n(X,Z) \mid Z\right)\right]^6} \leq C$, where h_n is defined accordingly.

In the proof of Theorem C.1, we assume $\mathbb{E}\left[h^2(W)\right] = 1$ without loss of generality. This is because we can always scale h by dividing by $\sqrt{\mathbb{E}\left[h^2(W)\right]}$ when the given working regression function satisfies $\mu(X,Z) \notin \mathscr{A}(Z)$. The floodgate inference procedure and results are the same with the corresponding scaled version $\tilde{h}(W)$. And the scaled version still satisfies the finite moment condition $\mathbb{E}\left[\tilde{h}^{12}(W)\right] < \infty$. Now we are dealing with a sequence of working regression functions μ_n. If we scale h_n analogously by dividing it by $\sqrt{\mathbb{E}\left[h_n^2(W)\right]}$, the corresponding function sequence $\{\tilde{h}_n\}$ does not necessarily satisfy the uniform moment condition, i.e., for all n, $\mathbb{E}\left[\tilde{h}_n^{12}(W)\right] < C$ for some constant C. But the moment conditions $\mathbb{E}[Y^{12}] < \infty$ and $\mathbb{E}\left[h_n^{12}(X,Z)\right] / (\mathbb{E}\left[\mathrm{Var}\left(\mu_n(X,Z) \mid Z\right)\right])^6 = \mathbb{E}\left[h_n^{12}(W)\right] / (\sqrt{\mathbb{E}\left[h_n^2(W)\right]})^{12} \leq C$ for all n ensure the uniform moment bound after scaling, hence for the following we can assume $\mathbb{E}\left[h_n^2(W)\right] = 1$.

According to the proof of Theorem C.1, we have the following Berry–Esseen bound

$$\sup_{M>0} \left|\mathbb{P}\left(\sqrt{n}|T| > M\right) - \mathbb{P}\left(|G| > M\right)\right| = O\left(\frac{1}{\sqrt{n}}\right),$$

which relies on verifying the following:

(i) $\mathbb{E}\left[|U_{01}|^3\right], \mathbb{E}\left[|U_{02}|^3\right], \mathbb{E}\left[|U_{03}|^3\right], \mathbb{E}\left[|U_{04}|^3\right], \mathbb{E}\left[|U_{05}|^3\right] < \infty$,

(ii) $\tilde{\sigma}_0^2(\mu_n) = H_2(\mathbf{0}) > 0$,

(iii) $\tilde{\sigma}^2(\mu_n) = \|L(U_0)\|_2 > 0$.

Note the above terms are defined similarly as in the proof of Theorem C.1 except the dependence on μ_n (but we abbreviate the notation dependence on μ_n for the random variables). We have $\tilde{\sigma}^2(\mu_n) = 1$ due to the derivations after (C.21) in the proof of Theorem C.1. To show the constant in the above rate of $\frac{1}{\sqrt{n}}$ is uniformly bounded, we need to prove $\inf_{\mu_n \in \mathcal{U}} \tilde{\sigma}^2(\mu_n) > 0$ and uniformly control the the 3rd moments in the condition (i). First notice that

$$\begin{aligned}
\inf_{\mu_n \in \mathcal{U}} \tilde{\sigma}^2(\mu_n) &\geq \inf_{\mu_n \in \mathcal{U}} \mathbb{E}\left[\mathrm{Var}\left(Y h_n(W) \mid Z\right)\right] \\
&\geq \inf_{\mu_n \in \mathcal{U}} \mathbb{E}\left[\mathrm{Var}\left(Y h_n(W) \mid X, Z\right)\right] \\
&= \inf_{\mu_n \in \mathcal{U}} \mathbb{E}\left[h_n^2(W) \mathrm{Var}\left(Y \mid X, Z\right)\right] \\
&\geq \tau > 0
\end{aligned}$$

where the first inequality holds due to (A.24), the second inequality holds as a result of the law of total conditional variance, the last equality holds by the assumption that $\mathbb{E}\left[h_n^2(W)\right] = 1$ and the moment lower bound condition $\mathrm{Var}\left(Y|X,Z\right) \geq \tau > 0$. Assuming $\mathbb{E}[Y^{12}] < \infty$ and $\mathbb{E}\left[\mu_n^{12}(X,Z)\right] / (\mathbb{E}\left[\mathrm{Var}\left(\mu_n(X,Z) \mid Z\right)\right])^6 \leq C$, we can uniformly control the moments $\mathbb{E}\left[|U_{01}|^3\right]$, $\mathbb{E}\left[|U_{02}|^3\right]$, $\mathbb{E}\left[|U_{03}|^3\right]$, $\mathbb{E}\left[|U_{04}|^3\right]$, $\mathbb{E}\left[|U_{05}|^3\right]$, therefore establish the rate of $\frac{1}{\sqrt{n}}$ in (A.67):

$$\Delta = O\left(\frac{1}{\sqrt{n}}\right).$$

Combining this with (A.66), we have

$$\sup_{\mu_n \in \mathcal{U}} \mathbb{P}\left(\sqrt{n}|T| > M\right) \leq \mathbb{P}\left(|G| > M\right) + \frac{C'}{\sqrt{n}}$$

for some constant C' depending on C, τ and $\mathbb{E}\left[Y^{12}\right]$. Therefore we obtain (A.65) and the choice of M can be universally chosen over $\mu_n \in \mathcal{U}$, which finally establishes $T = O_p\left(n^{-1/2}\right)$. Using similar strategies, we can prove $s = O_p\left(1\right)$. Hence we have shown (A.64).

Now we proceed to prove (A.63), first it can be simplified into the following form due to (A.6) and (A.8),

$$\mathcal{I} - f(\mu_n) = \sqrt{\mathbb{E}[(h^\star)^2(W)]} - \frac{\mathbb{E}\left[h_n(W) h^\star(W)\right]}{\sqrt{\mathbb{E}\left[h_n^2(W)\right]}} \tag{A.68}$$

where $h_n(W) = \mu_n(W) - \mathbb{E}\left[\mu_n(W) \mid Z\right]$ and $h^\star$ are defined the same way. Remark we have $0/0 = 0$ by convention for (A.68). We also find it is more convenient to work with $f(\bar{\mu}_n)$ (note $f(\mu_n) = f(\bar{\mu}_n)$), recall that the definition of $\bar{\mu}_n$:

$$\bar{\mu}_n(x,z) := \sqrt{\frac{\mathcal{I}}{\mathbb{E}\left[h_n^2(W)\right]}} \left(\mu_n(x,z) - \mathbb{E}\left[\mu_n(X,Z) \mid Z = z\right]\right) + \mathbb{E}\left[\mu^\star(X,Z) \mid Z = z\right],$$

and similarly denote $\bar{h}_n(w) = \bar{\mu}_n(x,z) - \mathbb{E}\left[\bar{\mu}_n(X,Z) \mid Z = z\right]$. When $\mu(X,Z) \in \mathscr{A}(Z)$, we have $\bar{\mu}_n(x,z) = \mathbb{E}\left[\mu^\star(X,Z) \mid Z = z\right], \bar{h}_n(w) = 0$, thus

$$\mathcal{I} - f(\mu_n) = \mathcal{I} = \frac{\mathbb{E}\left[(\bar{h}_n(W) - h^\star(W))^2\right]}{\sqrt{\mathbb{E}[(h^\star)^2(W)]}} \tag{A.69}$$

Otherwise when $\mathbb{E}\left[h_n^2(W)\right] > 0$, we have $\sqrt{\mathbb{E}\left[\bar{\mu}_n^2(W)\right]} = \mathcal{I}$. In this case, we rewrite the right hand side of (A.68) in terms of $\bar{\mu}_n$ and further simplify it as below,

$$\frac{\mathbb{E}\left[(\bar{h}_n(W) - h^\star(W))^2\right] - \left(\sqrt{\mathbb{E}\left[\bar{h}_n^2(W)\right]} - \sqrt{\mathbb{E}[(h^\star)^2(W)]}\right)^2}{2\sqrt{\mathbb{E}\left[\bar{h}_n^2(W)\right]}} = \frac{\mathbb{E}\left[(\bar{h}_n(W) - h^\star(W))^2\right]}{2\sqrt{\mathbb{E}[(h^\star)^2(W)]}}$$

which says that

$$\mathcal{I} - f(\mu_n) = \frac{\mathbb{E}\left[(\bar{h}_n(W) - h^\star(W))^2\right]}{2\sqrt{\mathbb{E}[(h^\star)^2(W)]}} \tag{A.70}$$

Note that $\sqrt{\mathbb{E}[(h^\star)^2(W)]} = \mathcal{I}$ which does not depend on μ, hence it suffices to show

$$\mathbb{E}\left[(\bar{h}_n(W) - h^\star(W))^2\right] = O_p\left(\inf_{\mu' \in S_{\mu_n}} \mathbb{E}\left[(\mu'(X,Z) - \mu^\star(X,Z))^2\right]\right). \tag{A.71}$$

We prove it by considering two cases:

(a) $\mathbb{E}\left[h_n(W)h^\star(W)\right] \leq 0$,

(b) $\mathbb{E}\left[h_n(W)h^\star(W)\right] > 0$.

Regarding case (a), we have

$$\begin{aligned}
\inf_{\mu' \in S_{\mu_n}} \mathbb{E}\left[(\mu'(X,Z) - \mu^\star(X,Z))^2\right] &= \inf_{c>0, \forall g(z)} \left(\mathbb{E}\left[(ch_n(W) - h^\star(W))^2\right] + \mathbb{E}\left[(g(Z) - \mathbb{E}\left[\mu^\star(W) \mid Z\right])^2\right]\right) \\
&= \inf_{c>0} \mathbb{E}\left[(ch_n(W) - h^\star(W))^2\right] \\
&= \mathbb{E}\left[(h^\star)^2(W)\right] + \inf_{c>0} c^2 \mathbb{E}\left[h_n^2(W)\right] - 2c\mathbb{E}\left[h_n(W)h^\star(W)\right] \\
&= \mathbb{E}\left[(h^\star)^2(W)\right]
\end{aligned}$$

where the first equality holds by the definition of S_{μ_n} and the fact that, for any $g(Z)$,

$$\mathbb{E}\left[h^\star(W)g(Z)\right] = \mathbb{E}\left[g(Z)\mathbb{E}\left[h^\star(W) \mid Z\right]\right] = 0$$

and similarly $\mathbb{E}\left[h_n(W)g(Z)\right] = 0$. The second equality holds by choosing $g(z)$ to be $\mathbb{E}\left[h^\star(W) \mid Z = z\right]$. The third equality is simply from expanding and the last equality holds in case (a). Noticing

$$\mathbb{E}\left[(\bar{h}_n(W) - h^\star(W))^2\right] \leq 2\left(\mathbb{E}\left[\bar{h}_n^2(W)\right] + \mathbb{E}\left[(h^\star)^2(W)\right]\right) = 4\mathbb{E}\left[(h^\star)^2(W)\right]$$

we thus establish (A.71). Regarding case (b), we have

$$\begin{aligned}
\inf_{\mu' \in S_{\mu_n}} \mathbb{E}\left[(\mu'(X,Z) - \mu^\star(X,Z))^2\right] &= \inf_{c>0} \mathbb{E}\left[(ch_n(W) - h^\star(W))^2\right] \\
&= \inf_{c>0} \mathbb{E}\left[(ch_n(W) - h_0(W) + h_0(W) - h^\star(W))^2\right] \\
&= \mathbb{E}\left[(h_0(W) - h^\star(W))^2\right] + \inf_{c>0} \mathbb{E}\left[(ch_n(W) - h_0(W))^2\right] \\
&= \mathbb{E}\left[(h_0(W) - h^\star(W))^2\right] \\
&= \mathbb{E}\left[(h^\star)^2(W)\right] - \mathbb{E}\left[(h_0(W))^2\right]
\end{aligned} \tag{A.72}$$

where in the second equality, h_0 is defined to be

$$h_0(w) := \frac{\mathbb{E}\left[h_n(W)h^\star(W)\right]}{\mathbb{E}\left[h_n^2(W)\right]} h_n(w).$$

It satisfies the property $\mathbb{E}\left[h_n(W)\left(h^\star(W) - h_0(W)\right)\right] = 0$ thus the third equality holds. The fourth equality comes from choosing c to be $\frac{\mathbb{E}[h_n(W)h^\star(W)]}{\mathbb{E}[h_n^2(W)]}$, which is positive in case (b). The last equality holds again due to $\mathbb{E}\left[h_n(W)\left(h^\star(W) - h_0(W)\right)\right] = 0$. And we have

$$\begin{aligned}
\mathbb{E}\left[(\bar{h}_n(W) - h^\star(W))^2\right] &= 2\mathbb{E}\left[(h^\star)^2(W)\right] - 2\mathbb{E}\left[\bar{h}_n(W)h^\star(W)\right] \\
&= 2\mathbb{E}\left[(h^\star)^2(W)\right] - 2\mathbb{E}\left[(h_0(W))^2\right]\rho
\end{aligned} \tag{A.73}$$

where ρ denotes the following term and can be further simplified based on the definition of $\bar{h}_n(W)$ and $h_0(W)$.

$$\begin{aligned}
\rho &:= \frac{\mathbb{E}\left[\bar{h}_n(W)h^\star(W)\right]}{\mathbb{E}\left[(h_0(W))^2\right]} \\
&= \frac{\mathcal{I}\sqrt{\mathbb{E}\left[h_n^2(W)\right]}}{\mathbb{E}\left[h_n(W)h^\star(W)\right]}
\end{aligned}$$

thus we have $\rho > 0$ in case (b) and $\rho \geq 1$ by the Cauchy–Schwarz inequality. Combining this with (A.72) and (A.73) yields (A.71). Finally we establish the bound in (2.11).

□

A.5 Proofs in Section 3.1

Proof of Lemma 3.2. We prove this lemma by a small trick, taking advantage of the idea of symmetry. Remember as in (A.45), X's null copy $\tilde{X}$ is constructed such that

$$\tilde{X} \perp\!\!\!\perp (X, Y) \mid Z, \quad \text{and} \quad \tilde{X} \mid Z \overset{d}{=} X \mid Z. \tag{A.74}$$

We can define the null copy of $\tilde{Y}$ by drawing from the conditional distribution of of Y given Z, without looking at (X, Y). Remark that introducing $\tilde{Y}$ is just for the convenience of proof and does not necessarily mean we need to be able to sample it. Formally it satisfy

$$\tilde{Y} \perp\!\!\!\perp (X, Y) \mid Z, \quad \tilde{Y} \mid Z \overset{d}{=} Y \mid Z \tag{A.75}$$

More specifically, we "generate" $\tilde{Y}$ conditioning on $(\tilde{X}, Z)$, following the same conditional distribution as $Y|X, Z$ (It can be verified this will satisfy (A.75)). Now by the symmetry argument, we have

$$\mathbb{E}\left[\mathbb{1}_{\{Y\cdot[\mu(\tilde{X},Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]]<0\}}\right] = \mathbb{E}\left[\mathbb{1}_{\{\tilde{Y}\cdot[\mu(X,Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]]<0\}}\right]. \tag{A.76}$$

Let $W = (X, Z)$ and define $g(Z) := \mathbb{E}[\mu(W) \mid Z]$, $h(W) := \mu(W) - g(Z)$ with the associated functions denoted by $g(z)$, $h(w)$, we can rewrite $f_{\ell_1}(\mu)/2$ as

$$\begin{aligned}
f_{\ell_1}(\mu)/2 &= \mathbb{P}\big(Y(\mu(\tilde{X}, Z) - \mathbb{E}[\mu(X, Z) \mid Z]) < 0\big) - \mathbb{P}\big(Y(\mu(X, Z) - \mathbb{E}[\mu(X, Z) \mid Z]) < 0\big) \\
&= \mathbb{E}\left[\mathbb{1}_{\{\tilde{Y}\cdot[\mu(W)-\mathbb{E}[\mu(W)\,|\,Z]]<0\}}\right] - \mathbb{E}\left[\mathbb{1}_{\{Y\cdot[\mu(W)-\mathbb{E}[\mu(W)\,|\,Z]]<0\}}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\mathbb{1}_{\{\tilde{Y}\cdot[\mu(W)-\mathbb{E}[\mu(W)\,|\,Z]]<0\}} - \mathbb{1}_{\{Y\cdot[\mu(W)-\mathbb{E}[\mu(W)\,|\,Z]]<0\}}\right) \,\middle|\, W\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\mathbb{1}_{\{\tilde{Y}\cdot h(W)<0\}} - \mathbb{1}_{\{Y\cdot h(W)<0\}}\right) \,\middle|\, W\right]\right]
\end{aligned}$$

where the second equality is by (A.76), the third one comes from the law of total expectation and the fourth one is by the definition of $h(W)$. Now it suffices to consider maximizing the following quantity

$$\mathbb{E}\left[\left(\mathbb{1}_{\{\tilde{Y}\cdot h(W)<0\}} - \mathbb{1}_{\{Y\cdot h(W)<0\}}\right) \,\middle|\, W = w\right] \tag{A.77}$$

for each $w = (x, z)$. Due to the property (A.75), we have

$$\mathbb{P}\left(\tilde{Y} = y \,|\, W\right) = \mathbb{P}\left(\tilde{Y} = y \,|\, Z\right) = \mathbb{P}(Y = y \,|\, Z) \quad y \in \{-1, 1\}$$

hence we can simplify the conditional expectation of the first indicator function in (A.77) into the following

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}_{\{\tilde{Y}\cdot h(W)<0\}} \,|\, W = w\right] &= \mathbb{P}\left(\tilde{Y} = 1,\ h(W) < 0 \,|\, W = w\right) + \mathbb{P}\left(\tilde{Y} = -1,\ h(W) > 0 \,|\, W = w\right) \\
&= \mathbb{P}(Y = 1 \,|\, Z = z)\, \mathbb{1}_{\{h(w)<0\}} + \mathbb{P}(Y = -1 \,|\, Z = z)\, \mathbb{1}_{\{h(w)>0\}}
\end{aligned} \tag{A.78}$$

Similarly we have

$$\mathbb{E}\left[\mathbb{1}_{\{Y\cdot h(W)<0\}} \,|\, W = w\right] = \mathbb{P}(Y = 1 \,|\, W = w)\, \mathbb{1}_{\{h(w)<0\}} + \mathbb{P}(Y = -1 \,|\, W = w)\, \mathbb{1}_{\{h(w)>0\}} \tag{A.79}$$

when $\mathbb{E}[Y \,|\, W = w] > \mathbb{E}[Y \,|\, Z = z]$, we have

$$\mathbb{P}(Y = 1 \,|\, W = w) > \mathbb{P}(Y = 1 \,|\, Z = z), \ \mathbb{P}(Y = -1 \,|\, W = w) < \mathbb{P}(Y = -1 \,|\, Z = z),$$

hence in this case, by comparing (A.78) and (A.79) we know $h(w) > 0$ will maximize (A.77) with maximum value

$$\begin{aligned} \mathbb{P}(Y=-1 \mid Z=z) - \mathbb{P}(Y=-1 \mid W=w) &= (1-\mathbb{E}[Y \mid Z=z])/2 - (1-\mathbb{E}[Y \mid W=w])/2 \\ &= (\mathbb{E}[Y \mid W=w] - \mathbb{E}[Y \mid Z=z])/2 \end{aligned} \tag{A.80}$$

Similarly we can figure out the maximizer of $h(w)$, when $\mathbb{E}[Y \mid W=w] < \mathbb{E}[Y \mid Z=z]$. Finally we have

$$h(w) \begin{cases} > 0, & \text{when } \mathbb{E}[Y \mid W=w] > \mathbb{E}[Y \mid Z=z] \\ < 0, & \text{when } \mathbb{E}[Y \mid W=w] < \mathbb{E}[Y \mid Z=z] \\ \text{can be any choice}, & \text{when } \mathbb{E}[Y \mid W=w] = \mathbb{E}[Y \mid Z=z] \end{cases} \tag{A.81}$$

will maximize (A.77) with the maximum value $|\mathbb{E}[Y \mid W=w] - \mathbb{E}[Y \mid Z=z]|/2$. Remark the definition of $h(w) = \mu(w) - g(z)$, we can restate (A.81) as

$$\begin{cases} \mu(x,z) = \mu(w) > g(z), & \text{when } \mathbb{E}[Y \mid W=w] > \mathbb{E}[Y \mid Z=z] \\ \mu(x,z) = \mu(w) < g(z), & \text{when } \mathbb{E}[Y \mid W=w] < \mathbb{E}[Y \mid Z=z] \\ \text{can be any choice}, & \text{when } \mathbb{E}[Y \mid W=w] = \mathbb{E}[Y \mid Z=z] \end{cases} \tag{A.82}$$

where again $g(z) = \mathbb{E}[\mu(X,Z) \mid Z=z]$. Apparently, choosing $\mu(x,z)$ to be the true regression function $\mu^\star(x,z)$ will satisfy (A.82). Hence we show $f_{\ell_1}(\mu)$ is maximized at $\mu^\star$ with maximum value

$$\mathbb{E}\,|\mathbb{E}[Y \mid Z] - \mathbb{E}[Y \mid X, Z]|$$

which equals $\mathcal{I}_{\ell_1}$. Clearly from (A.82), $\mu^\star(x,z)$ is not the unique maximizer and any function in the set described in the following set can attain the maximum.

$$\{\mu : \mathbb{R}^p \to \mathbb{R} \mid \operatorname{sign}(\mu(x,z) - \mathbb{E}[\mu(X,Z) \mid Z=z]) = \operatorname{sign}(\mathbb{E}[Y \mid X=x] - \mathbb{E}[Y \mid Z=z])\}. \tag{A.83}$$

□

Proof of Theorem 3.3. According to Algorithm 3, we first denote

$$\begin{aligned} &U := \mu(X,Z), \quad g(z) := \mathbb{E}[\mu(X,Z) \mid Z=z], \\ &G_z(u) := \mathbb{P}(U < u \mid Z=z), \quad F_z(u) := \mathbb{P}(U \le u \mid Z=z). \end{aligned} \tag{A.84}$$

thus have the following expression of R_i:

$$R_i = G_{Z_i}(g(Z_i))\mathbb{1}_{\{Y_i=1\}} + (1 - F_{Z_i}(g(Z_i)))\mathbb{1}_{\{Y_i=-1\}} - \mathbb{1}_{\{Y_i(\mu(W_i)-g(Z_i))<0\}}$$

First we prove that $\mathbb{E}[R_i] = f_{\ell_1}(\mu)/2$. Recall the definition of $f_{\ell_1}(\mu)$ in (3.2),

$$f_{\ell_1}(\mu)/2 = \mathbb{E}\left[\mathbb{1}_{\{Y\cdot[\mu(\tilde{X},Z)-\mathbb{E}[\mu(X,Z) \mid Z]]<0\}}\right] - \mathbb{E}\left[\mathbb{1}_{\{Y\cdot[\mu(X,Z)-\mathbb{E}[\mu(X,Z) \mid Z]]<0\}}\right],$$

let $W = (X,Z)$, then it suffices to show the following

$$\mathbb{E}\left[G_Z(g(Z))\mathbb{1}_{\{Y=1\}} + (1-F_Z(g(Z)))\mathbb{1}_{\{Y=-1\}}\right] = \mathbb{E}\left[\mathbb{1}_{\{Y\cdot[\mu(\tilde{X},Z)-\mathbb{E}[\mu(X,Z) \mid Z]]<0\}}\right]. \tag{A.85}$$

By the law of total expectation we can rewrite the right hand side as

$$\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{Y\cdot[\mu(\tilde{X},Z)-\mathbb{E}[\mu(X,Z) \mid Z]]<0\}} \mid Z, Y\right]\right].$$

Due to the property (A.74), we have $\tilde{X} \perp\!\!\!\perp (Y, Z) \mid Z$ and $\tilde{X} \mid Z \sim X \mid Z$, which yields

$$\mathbb{E}\left[\mathbb{1}_{\{Y\cdot[\mu(\tilde{X},Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]]<0\}} \,|\, Z=z, Y=1\right] = G_Z(g(Z))\mathbb{1}_{\{Y=1\}}.$$

And we can do similar derivations when $Y=-1$. Thus we can prove $\mathbb{E}[R_i] = f_{\ell_1}(\mu)/2$ by showing (A.85). In light of the deterministic relationship in Lemma 3.2, we have $\{L_n^\alpha(\mu) \le f_{\ell_1}(\mu)\} \subset \{L_n^\alpha(\mu) \le \mathcal{I}_{\ell_1}\}$, hence it suffices to prove

$$\mathbb{P}\left(L_n^\alpha(\mu) \le f_{\ell_1}(\mu)\right) \ge 1-\alpha-O(n^{-1/2}). \tag{A.86}$$

Note that $\mathrm{Var}(R_i)$ always exist due to the boundedness. When $\mathrm{Var}(R_i)=0$, we have $R_i = f_{\ell_1}(\mu)/2 = \bar{R}$ and $s=0$, thus $L_n^\alpha(\mu) = f_{\ell_1}(\mu)$, hence (A.86) trivially holds. Remark this includes the case when $\mu(X,Z) \in \mathscr{A}(Z)$. Otherwise, applying Lemma C.4 to i.i.d. bounded random variables R_i will yield (A.86), where the constant will depend on $\mathrm{Var}(R_i)$. □

A.6 Proofs in Section 3.2

Proof of Theorem 3.4. When $\mathcal{T}$ is degenerate or $\mu(X) \in \mathscr{A}(Z)$, we immediately have $L_n^{\alpha,\mathcal{T}}(\mu)=0$ according to Algorithm 4, which implies the coverage validity. Below we focus on the non-trivial case. Due to the deterministic relationship

$$f_n^{\mathcal{T}}(\mu) \le f_n^{\mathcal{T}}(\mu^\star) \le f(\mu^\star) = \mathcal{I},$$

it suffices to prove

$$\mathbb{P}_P\left(L_n^{\alpha,\mathcal{T}}(\mu) \le f_n^{\mathcal{T}}(\mu)\right) \ge 1-\alpha-o(1). \tag{A.87}$$

which can be reduced to establishing certain asymptotic normality based on i.i.d. random variables $R_m, V_m, m \in [n_1]$ whenever the variance of the asymptotic distribution is nonzero. First, we verify that under the stated conditions, all the involving moments are finite, which can be reduced to show

$$\mathrm{Var}(R_m), \mathrm{Var}(V_m) < \infty.$$

For a given n_2, it can be further reduced to the following

$$\begin{aligned}\mathrm{Var}\left(Y_i\left(\mu(X_i,Z_i) - \mathbb{E}\left[\mu(X_i,Z_i)\,|\,\boldsymbol{Z}_m, \boldsymbol{T}_m\right]\right)\right.\\ \mathrm{Var}\left(\mathrm{Var}\left(\mu(X_i,Z_i)\,|\,\boldsymbol{Z}_m,\boldsymbol{T}_m\right)\right) < \infty.\end{aligned}$$

Using similar strategies in the proof of Theorem 2.3, we can show the above holds under the moment conditions $\mathbb{E}\left[Y^4\right], \mathbb{E}\left[\mu^4(X)\right] < \infty$ by the Cauchy–Schwarz inequality and the tower property of conditional expectation.

Note that in the proof of the main result, i.e. Theorem 2.3, we consider four different cases based on whether some variances are zero or not. Here we only pursue the asymptotic coverage validity, then the discussion on those four different cases becomes very straighforward. When both the variances of R_m, V_m are zero, we have $\bar{R}/\bar{V} = f_n^{\mathcal{T}}(\mu)$, $s^2=0$, then (A.87) holds immediately. When $\mathrm{Var}(V_m)=0$, we can simply establish the asymptotic normality by the central limit theorem. Otherwise, delta method can be applied. Here we give the derivation for the most non-trivial case where $\mathrm{Var}(R_m), \mathrm{Var}(V_m) > 0$. Denote random vectors $\{U_m\}_{m=1}^{n_1} = \{(U_{m1}, U_{m2})\}_{m=1}^{n_1} \overset{i.i.d.}{\sim} U = (U_1, U_2)$ to be

$$U_{m1} = R_m - \mathbb{E}\left[Y_i\left(\mu(X_i,Z_i) - \mathbb{E}\left[\mu(X_i,Z_i)\,|\,\boldsymbol{Z}_m,\boldsymbol{T}_m\right]\right]\right], \tag{A.88}$$

$$U_{m2} = V_m - \mathbb{E}\left[\mathrm{Var}\left(\mu(X_i,Z_i)\,|\,\boldsymbol{Z}_m,\boldsymbol{T}_m\right)\right] \tag{A.89}$$

hence we have $\mathbb{E}[U]=0$. Denote $h^{\mathcal{T}}(W_i)=\mu(X_i,Z_i)-\mathbb{E}[\mu(X_i,Z_i)\,|\,\boldsymbol{Z}_m,\boldsymbol{T}_m]$, we have the following holds

$$
\begin{aligned}
f_n^{\mathcal{T}}(\mu) &= \frac{\mathbb{E}[\mathrm{Cov}(\mu^\star(X_i,Z_i),\mu(X_i,Z_i)\,|\,\boldsymbol{Z},\boldsymbol{T})]}{\sqrt{\mathbb{E}[\mathrm{Var}(\mu(X_i,Z_i)\,|\,\boldsymbol{Z},\boldsymbol{T})]}} \\
&= \frac{\mathbb{E}\left[\mathrm{Cov}(\mu^\star(X_i,Z_i),h^{\mathcal{T}}(W_i)\,|\,\boldsymbol{Z},\boldsymbol{T})\right]}{\sqrt{\mathbb{E}\left[\mathbb{E}\left[(h^{\mathcal{T}}(W_i)^2)\right]\right]}} \\
&= \frac{\mathbb{E}\left[\mu^\star(X_i,Z_i)h^{\mathcal{T}}(W_i)\right]}{\sqrt{\mathbb{E}\left[\mathbb{E}\left[(h^{\mathcal{T}}(W_i)^2)\right]\right]}} \\
&= \frac{\mathbb{E}\left[Y_i h^{\mathcal{T}}(W_i)\right]}{\sqrt{\mathbb{E}\left[(h^{\mathcal{T}}(W_i)^2)\right]}},
\end{aligned}
$$

where the first equality holds by the definition of $f_n^{\mathcal{T}}(\mu)$, the second inequality holds by the definition of $h^{\mathcal{T}}(W_i)$. Regarding the third equality, we make use of the fact $\mathbb{E}\left[h^{\mathcal{T}}(W_i)\,|\,\boldsymbol{Z}_m,\boldsymbol{T}_m\right]=0$ and the tower property of conditional expectation. The last inequality holds by the tower property of conditional expectation and the fact that $h^{\mathcal{T}}(W_i)\in\mathscr{A}(\boldsymbol{X}_m,\boldsymbol{Z}_m)$. Let $T=\bar{R}/\bar{V}$, then $T-f_n^{\mathcal{T}}(\mu)$ can be rewritten as

$$
T-f_n^{\mathcal{T}}(\mu)=\frac{\bar{U}_1+\mathbb{E}\left[Y_i h^{\mathcal{T}}(W_i)\right]}{\sqrt{\bar{U}_2+\mathbb{E}\left[(h^{\mathcal{T}}(W_i)^2)\right]}}-\frac{\mathbb{E}\left[Y_i h^{\mathcal{T}}(W_i)\right]}{\sqrt{\mathbb{E}\left[(h^{\mathcal{T}}(W_i)^2)\right]}}:=H(\bar{U})
$$

where $\bar{U}=(\bar{U}_1,\bar{U}_2)=\frac{1}{n_1}\sum_{i=1}^n U_m$ and $H:\mathbb{R}^2\to\mathbb{R}$ is defined through the following:

$$
H(x)=H(x_1,x_2):=\frac{x_1+\mathbb{E}\left[Y_i h^{\mathcal{T}}(W_i)\right]}{\sqrt{x_2+\mathbb{E}\left[(h^{\mathcal{T}}(W_i)^2)\right]}}-\frac{\mathbb{E}\left[Y_i h^{\mathcal{T}}(W_i)\right]}{\sqrt{\mathbb{E}\left[(h^{\mathcal{T}}(W_i)^2)\right]}}:=H(\bar{U})
$$

when $x_2>-\mathbb{E}\left[(h^{\mathcal{T}}(W_i)^2)\right]$ and is set to be $\frac{\mathbb{E}\left[Y_i h^{\mathcal{T}}(W_i)\right]}{\sqrt{\mathbb{E}\left[(h^{\mathcal{T}}(W_i)^2)\right]}}$ otherwise. Note that the first order derivatives of $H(x)$ exists, by applying the multivariate Delta method to mean zero random vectors $\{(U_{m1},U_{m2})\}_{m=1}^{n_1}$ with the nonlinear function chosen as H, we have

$$
\sqrt{n_1}(T-f_n^{\mathcal{T}}(\mu))\xrightarrow{d}\mathcal{N}\left(0,\tilde{\sigma}^2\right)
$$

whenever the variance term $\tilde{\sigma}^2$ is nonzero. Exactly following the strategy in the proof of Theorem 2.3, we have $\tilde{\sigma}^2>0$ under the case where $\mathrm{Var}(R_m),\mathrm{Var}(V_m)>0$. Also notice s^2 is a consistent estimator of $\tilde{\sigma}^2$, then by the argument of Slutsky's Theorem, (A.87) is established. □

B An example for projection methods

Consider covariates $W=(W_1,W_2)$ distributed as $W_1\sim\mathcal{N}(0,1)$ and $W_2=W_1^2+\mathcal{N}(0,1)$. Let $Y=W_1^2+\mathcal{N}(0,1)$, with all the Gaussian random variables independent. Then W_1 is the only important variable; formally: $W_1\not\perp\!\!\!\perp Y\mid W_2$ and $W_2\perp\!\!\!\perp Y\mid W_1$. But the projection parameters are $(\mathbb{E}\left[W^\top W\right])^{-1}\mathbb{E}[WY]=(0,\frac{3}{4})^\top$, i.e., zero for the non-null covariate and non-zero for the null covariate.

C Rate results

Theorem C.1 (Floodgate validity)**.** *For any given working regression function $\mu:\mathbb{R}^p\to\mathbb{R}$ and i.i.d. data $\{(Y_i,X_i,Z_i)\}_{i=1}^n$, if $\mathbb{E}[Y^{12}]$, $\mathbb{E}[\mu^{12}(X,Z)]<\infty$, then $L_n^\alpha(\mu)$ from Algorithm 1 satisfies*

$$
\mathbb{P}\left(L_n^\alpha(\mu)\le\mathcal{I}\right)\ge 1-\alpha-Cn^{-1/2}
$$

for some constant C depending only on the moments of Y and $\mu(X,Z)$.

The proof can be found in Appendix C.1.1. Establishing the $n^{-1/2}$ rate requires relatively recent Berry–Esseen-type results for the delta method (Pinelis et al., 2016) and also necessitates the existence of 12th moments.

Theorem C.2. *Assume the conditions of Theorem C.1 and* $\mathbb{E}\left[\mathrm{Var}\left(Y\left(\mu(X,Z)-\mathbb{E}\left[\mu(X,Z)\,|\,Z\right]\right)\,|\,Z\right)\right]>0$. $L^{\alpha}_{n,K}(\mu)$ *computed by replacing* R_i *and* V_i *with* R_i^K *and* V_i^K*, respectively, in Algorithm 1 satisfies*

$$\inf_{K>1}\mathbb{P}\left(L^{\alpha}_{n,K}(\mu)\le\mathcal{I}\right)\ge 1-\alpha-Cn^{-1/2}$$

for some constant C *depending only on the moments of* Y *and* $\mu(X,Z)$.

The proof can be found in Appendix C.1.2. Note that the additional assumption beyond Theorem C.1 of $\mathbb{E}\left[\mathrm{Var}\left(Y\left(\mu(X,Z)-\mathbb{E}\left[\mu(X,Z)\,|\,Z\right]\right)\,|\,Z\right)\right]>0$ is only needed for $n^{-1/2}$-rate coverage validity *uniformly* over $K>1$, and could be removed for the same result for any fixed $K>1$.

C.1 Proofs in Appendix C

C.1.1 Theorem C.1

Proof of Theorem C.1. Recall in Algorithm 1, we denote $R_i=Y_i\big(\mu(X_i,Z_i)-\mathbb{E}\left[\mu(X_i,Z_i)\,|\,Z_i\right]\big)$ and $V_i=\mathrm{Var}\left(\mu(X_i,Z_i)\,|\,Z_i\right)$ for each $i\in[n]$, and compute their sample mean $(\bar{R},\bar{V})$ and sample covariance matrix $\hat{\Sigma}$. The LCB is constructed as

$$L^{\alpha}_{n}(\mu)=\max\left\{\frac{\bar{R}}{\sqrt{\bar{V}}}-\frac{z_{\alpha}s}{\sqrt{n}},0\right\},\quad\text{where }\; s^2=\frac{1}{\bar{V}}\left[\left(\frac{\bar{R}}{2\bar{V}}\right)^2\hat{\Sigma}_{22}+\hat{\Sigma}_{11}-\frac{\bar{R}}{\bar{V}}\hat{\Sigma}_{12}\right].$$

Following exactly the same discussions as those from the beginning to (A.10) in the proof of Theorem 2.3, we have

- Theorem C.1 can be proved under the weaker moment conditions that $\mathbb{E}\left[Y^{12}\right],\mathbb{E}\left[h^{12}(W)\right]<\infty$, which is assumed for the following proof;
- it suffices to prove
$$\mathbb{P}\left(\frac{\bar{R}}{\sqrt{\bar{V}}}-\frac{z_{\alpha}s}{\sqrt{n}}\le f(\mu)\right)\ge 1-\alpha-C/\sqrt{n}\tag{C.1}$$
for some constant C when $\mathbb{E}\left[\mathrm{Var}(\mu(X,Z)\,|\,Z)\right]\neq 0$;
- we can assume $\mathbb{E}\left[h^2(W)\right]=1$ without loss of generality.

We will utilize Berry–Esseen-type bounds to prove (C.1). Now we still consider the following four cases.

(I) $\mathrm{Var}\left(Yh(W)\right)=0$ and $\mathrm{Var}\left(\mathrm{Var}\left(h(W)\,|\,Z\right)\right)=0$.

(II) $\mathrm{Var}\left(Yh(W)\right)>0$ and $\mathrm{Var}\left(\mathrm{Var}\left(h(W)\,|\,Z\right)\right)=0$.

(III) $\mathrm{Var}\left(Yh(W)\right)=0$ and $\mathrm{Var}\left(\mathrm{Var}\left(h(W)\,|\,Z\right)\right)>0$.

(IV) $\mathrm{Var}\left(Yh(W)\right)>0$ and $\mathrm{Var}\left(\mathrm{Var}\left(h(X)\,|\,Z\right)\right)>0$.

Note that assuming $\mathbb{E}\left[Y^{12}\right]$ and $\mathbb{E}\left[h^{12}(W)\right]<\infty$ ensures all the above variances exist due to the same bounding strategy as (A.9).

Case (I): (C.1) holds by the discussion for Case (I) in the proof of Theorem 2.3.

Case (II): due to the derivations for Case (II) in the proof of Theorem 2.3, the problem is reduced to showing

$$\mathbb{P}\left(\bar{R}-\frac{z_{\alpha}(\hat{\Sigma}_{11})^{1/2}}{\sqrt{n}}\le\mathbb{E}\left[Yh(W)\right]\right)\ge 1-\alpha-C/\sqrt{n}.\tag{C.2}$$

As mentioned in the proof of Theorem 2.3, $\bar{R}$ is simply the sample mean estimator of the quantity $\mathbb{E}[Yh(W)]$ and $\hat{\Sigma}_{11}$ is the corresponding sample variance. Therefore, the CLT and Slutsky's theorem immediately establish the asymptotic coverage validity. To prove the $1/\sqrt{n}$ rate in (C.2), stronger results are needed. The classical Berry–Esseen bound serves as the main ingredient, which states that

Lemma C.3 (Berry–Esseen bound)**.** *There exists a positive constant C, such that for i.i.d. mean zero random variables $X_1, \ldots, X_n$ satisfying*

(1) $\mathbb{E}[X_1^2] = \sigma^2 > 0$

(2) $\mathbb{E}[|X_1|^3] = \rho < \infty$

if we define $F_n(x)$ to be the cumulative distribution function (CDF) of the scaled average $\sqrt{n}\bar{X}/\sigma$ and denote the CDF of the standard normal distribution by $\Phi(x)$, then we have

$$\sup_{x\in\mathbb{R}} |F_n(x) - \Phi(x)| \leq \frac{C\rho}{\sigma^3\sqrt{n}}. \tag{C.3}$$

Since σ in the above result is generally unknown and usually replaced by the sample variance $s_\sigma^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$, we need the following lemma, which is proved in Bentkus et al. (1996).

Lemma C.4 (Berry–Esseen bound for Student's statistic)**.** *Under the same conditions as in Lemma C.3, if we redefine $F_n(x)$ to be the cumulative distribution function (CDF) of the Student t-statistic $\sqrt{n}\bar{X}/s_\sigma$, then we have the following Berry–Esseen bound*

$$\sup_{x\in\mathbb{R}} |F_n(x) - \Phi(x)| \leq \frac{C'\rho}{\sigma^3\sqrt{n}}. \tag{C.4}$$

To apply Lemma C.4, since we are in Case (II) where $\mathrm{Var}\left(\mathrm{Var}\left(h(W) \mid Z\right)\right) = 0$ and $\mathrm{Var}\left(Yh(W)\right) > 0$, it suffices to verify the finiteness of the term "ρ" in our context:

$$\begin{aligned} \rho &= \mathbb{E}\left[\left|Yh(W) - \mathbb{E}\left[Yh(W)\right]\right|^3\right] \\ &\leq 2^{3-1}\left(\mathbb{E}\left[Y^3 h^3(W)\right] + \left|\mathbb{E}\left[Yh(W)\right]\right|^3\right) < \infty \end{aligned}$$

where the equality holds since we assume $\mathbb{E}\left[h^2(W)\right] = 1$ and the inequality comes from the C_r inequality. For the last inequality, using the Cauchy–Schwarz inequality and the fact that higher moments dominate lower moments, we obtain the finiteness when assuming $\mathbb{E}\left[Y^6\right], \mathbb{E}\left[h^6(W)\right] < \infty$, which holds under the assumed moment conditions. Now by applying the Berry–Esseen bound in Lemma C.4 with $\bar{X} = \bar{R} - \mathbb{E}\left[Yh(W)\right]$ and $s_\sigma^2 = \hat{\Sigma}_{11}$, we obtain (C.2).

Case (III): due to (A.11), we have

$$\frac{\bar{R}}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n}} = \frac{\mathbb{E}\left[Yh(W)\right]}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n}}, \quad \text{where} \quad s^2 = \frac{1}{\bar{V}}\left(\frac{\mathbb{E}\left[Yh(W)\right]}{2\bar{V}}\right)^2 \hat{\Sigma}_{22}.$$

Note $\frac{\mathbb{E}[Yh(W)]}{\sqrt{\bar{V}}}$ is a nonlinear function of the moment estimators, so the following asymptotic normality result is a direct consequence of the multivariate delta method,

$$\sqrt{n}\left(\frac{\mathbb{E}\left[Yh(W)\right]}{\sqrt{\bar{V}}} - f(\mu)\right) \xrightarrow{d} \mathcal{N}\left(0, \tilde{\sigma}_0^2\right),$$

where $\tilde{\sigma}_0^2 = H_2(\mathbf{0})$ will be specified later (see the definition of $H_2(x)$ in (C.10)) and s^2 in $L_n^\alpha(\mu)$ is a consistent estimator of it. To establish the rate $1/\sqrt{n}$, the classical Berry–Esseen result needs to be extended for nonlinear statistics. Note that Case (IV) involves a nonlinear statistic too, and is a bit more complicated. Hence we focus on Case (IV) and omit the very similar proof for Case (III).

Case (IV): Denote $T := \left(\frac{\bar{R}}{\sqrt{\bar{V}}} - f(\mu)\right)/s$. Under specific moment conditions, we will establish the Berry–Esseen-type bound below:

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}\left(\sqrt{n}T \le t\right) - \Phi(t)\right| = O\left(\frac{1}{\sqrt{n}}\right) \tag{C.5}$$

where $\Phi(t)$ denotes the CDF of the standard normal distribution.

The proof relies on a careful analysis of nonlinear statistics. We take advantage of the results in a recent paper (Pinelis et al., 2016) that establishes Berry–Esseen bounds with rate $1/\sqrt{n}$ for the multivariate delta method when the function applied to the sample mean estimator satisfies certain smoothness conditions. And the constants in the rate depend on the distribution only through several moments. Specifically, consider $U, U_1, \ldots, U_n$ to be i.i.d. random vectors on a set $\mathcal{X}$ and a functional $H : \mathcal{X} \to \mathbb{R}$ which satisfies the following smoothness condition:

Condition C.5. *There exists $\varepsilon, M_\varepsilon > 0$ and a continuous linear functional $L : \mathcal{X} \to \mathbb{R}$ such that*

$$|H(x) - L(x)| \le M_\varepsilon \|x\|^2 \quad \text{for all } x \in \mathcal{X} \text{ with } \|x\| \le \varepsilon \tag{C.6}$$

We can think of L as the first-order Taylor expansion of H. This smoothness condition basically requires H to be nearly linear around the origin and can be satisfied if its second derivatives are bounded in the small neighbourhood $\{x : \|x\| \le \varepsilon\}$. Before stating Pinelis et al. (2016)'s result (we change their notation to avoid conflicts with the notation in the main text of this paper), define $\bar{U} := \frac{1}{n}\sum_{i=1}^n U_i$ and

$$\tilde{\sigma} := \|L(U)\|_2, \quad \nu_p := \|U\|_p, \quad \varsigma_p := \frac{\|L(U)\|_p}{\tilde{\sigma}},$$

where for a given random vector $U = (U_1, \cdots, U_d) \in \mathbb{R}^d$, $\|U\|_p$ is defined as $\|U\|_p = \left(\mathbb{E}\left[\|U\|^p\right]\right)^{1/p}$ with $\|u\|^p := \sum_{j=1}^d |u_j|^p$.

Theorem C.6. *(Pinelis et al., 2016, Theorem 2.11) Let $\mathcal{X}$ be a Hilbert space, let H satisfy Condition C.5 for some $\epsilon > 0$, and assume $\mathbb{E}[U] = 0$, $\tilde{\sigma} > 0$ and $\nu_3 < \infty$, then*

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}\left(\frac{\sqrt{n}H(\bar{U})}{\tilde{\sigma}} \le t\right) - \Phi(t)\right| \le \frac{C}{\sqrt{n}} \tag{C.7}$$

where the constant C depends on the distribution of U only through $\tilde{\sigma}, \nu_2, \nu_3, \varsigma_3$ (it also depends on the smoothness of the functional H through ϵ, M_ϵ).

Note that the above result is a generalization of the standard Berry–Esseen bound. $\tilde{\sigma}^2$ is the variance term of the asymptotic normal distribution. ς_3 is closely related to the term ρ/σ^2 in (C.3). The quantities $\tilde{\sigma}, \nu_2, \nu_3, \varsigma_3$ involved in the constant C only involve up to third moments, which is in accordance with the standard Berry–Esseen bound in Lemmas C.3 and C.4. Note the existence of $\tilde{\sigma}, \nu_2, \varsigma_3$ is implied by $\nu_3 < \infty$ due to the fact that lower moments can be controlled by higher moments, together with the linearity of the functional L. To apply Theorem C.6 to our problem, we first let $\mathcal{X} = \mathbb{R}^5$ and random vectors $\{U_i\}_{i=1}^n = \{(U_{i1}, U_{i2}, U_{i3}, U_{i4}, U_{i5})\}_{i=1}^n \overset{i.i.d.}{\sim} U_0 = (U_{01}, U_{02}, U_{03}, U_{04}, U_{05})$ to be

$$\begin{aligned} U_{i1} = R_i - \mathbb{E}\left[Yh(W)\right], \quad U_{i2} = V_i - \mathbb{E}\left[h^2(W)\right], \\ U_{i3} = Y_i^2 h^2(W_i) - \mathbb{E}\left[Y^2 h^2(W)\right], \quad U_{i4} = \left(\mathrm{Var}\left(h(W_i) \mid Z_i\right)\right)^2 - \mathbb{E}\left[\left(\mathrm{Var}\left(h(W) \mid Z\right)\right)^2\right], \\ U_{i5} = R_i \mathrm{Var}\left(h(W_i) \mid Z_i\right) - \mathbb{E}\left[Yh(W)\mathrm{Var}\left(h(W) \mid Z\right)\right]. \end{aligned} \tag{C.8}$$

Recall the definition $R_i = Y_i\big(\mu(X_i, Z_i) - \mathbb{E}\left[\mu(X, Z_i) \mid Z_i\right]\big)$ and $V_i = \mathrm{Var}\left(\mu(X_i, Z_i) \mid Z_i\right)$, hence we have $\mathbb{E}[U_i] = \mathbb{E}[U_0] = \mathbf{0}$. Let $\bar{U} = (\bar{U}_1, \bar{U}_2, \bar{U}_3, \bar{U}_4, \bar{U}_5) = \frac{1}{n}\sum_{i=1}^n U_i \in \mathbb{R}^5$, recall the definition $T = \left(\frac{\bar{R}}{\sqrt{\bar{V}}} - f(\mu)\right)/s$ where $s^2 = \frac{1}{\bar{V}}\left[\left(\frac{\bar{R}}{2\bar{V}}\right)^2 \hat{\Sigma}_{22} + \hat{\Sigma}_{11} - \frac{\bar{R}}{\bar{V}}\hat{\Sigma}_{12}\right]$, then T can be rewritten as

$$T = H(\bar{U}) := \frac{H_1(\bar{U}_1, \bar{U}_2)}{\sqrt{H_2(\bar{U})}},$$

where $H_1(\bar{U}_1, \bar{U}_2)$ and $H_2(\bar{U})$ are defined as

$$H_1(\bar{U}_1, \bar{U}_2) := \frac{\bar{U}_1 + \mathbb{E}[Yh(W)]}{\sqrt{\bar{U}_2 + \mathbb{E}[h^2(W)]}} - \frac{\mathbb{E}[Yh(W)]}{\sqrt{\mathbb{E}[h^2(W)]}}, \tag{C.9}$$

$$\begin{aligned} H_2(\bar{U}) := \frac{1}{\bar{U}_2 + \mathbb{E}[h^2(W)]}\Bigg[& \left(\frac{\bar{U}_1 + \mathbb{E}[Yh(W)]}{2(\bar{U}_2 + \mathbb{E}[h^2(W)])}\right)^2 \left(\bar{U}_4 + \mathbb{E}\left[(\mathrm{Var}(h(W) \mid Z))^2\right] - (\bar{U}_2 + \mathbb{E}\left[h^2(W)\right])^2\right) \\ & + \bar{U}_3 + \mathbb{E}\left[Y^2 h^2(W)\right] - (\bar{U}_1 + \mathbb{E}[Yh(W)])^2 \\ & - \frac{\bar{U}_1 + \mathbb{E}[Yh(W)]}{\bar{U}_2 + \mathbb{E}[h^2(W)]}\left(\bar{U}_5 + \mathbb{E}[Yh(W)\mathrm{Var}(h(W) \mid Z)] - (\bar{U}_1 + \mathbb{E}[Yh(W)])(\bar{U}_2 + \mathbb{E}\left[h^2(W)\right])\right)\Bigg]. \end{aligned} \tag{C.10}$$

Note $H(x) = H(x_1, x_2, x_3, x_4, x_5) : \mathbb{R}^5 \to \mathbb{R}$ is defined by replacing the above $\bar{U} = (\bar{U}_1, \bar{U}_2, \bar{U}_3, \bar{U}_4, \bar{U}_5)$ by $x := (x_1, x_2, x_3, x_4, x_5)$ respectively. When $x_2 > -\mathbb{E}\left[h^2(W)\right]$ or $H_2(x) = 0$, $H(x)$ is set to be 0. If we can verify the conditions for $T = H(\bar{U})$, Theorem C.6 implies

$$\sup_{t \in \mathbb{R}} \left|\mathbb{P}\left(\sqrt{n}T \le t\tilde{\sigma}\right) - \Phi(t)\right| \le \frac{C}{\sqrt{n}},$$

for some constant C, where $\tilde{\sigma} = \|L(U_0)\|_2 > 0$ (we will define $L(x)$ shortly and subsequently show $\tilde{\sigma} = 1$). Theorem C.6 says that the constant C above only depends on some universal constants and $\tilde{\sigma}, \nu_2, \nu_3, \varsigma_3$, which are the moments of U (i.e., the moments of U_i). Since U_i (defined in the three lines around (C.8)) is a function of Y_i and $h(W_i)$, we will apply the Cauchy–Schwarz inequality to further bound the moments of U by the moments of Y and $h(W) = \mu(X, Z) - \mathbb{E}[\mu(X, Z) \mid Z]$. First we need to verify Condition C.5, i.e., there exists $\varepsilon, M_\varepsilon > 0$ and a continuous linear functional $L : \mathbb{R}^5 \to \mathbb{R}$ such that

$$|H(x) - L(x)| \le M_\varepsilon \|x\|^2 \quad \text{for all } x \in \mathbb{R}^5 \text{ with } \|x\| \le \varepsilon. \tag{C.11}$$

Second, we will show $\tilde{\sigma}$, ν_3, and ς_3 are finite under the stated moment conditions.

Regarding the smoothness condition, consider the first order Taylor expansion of H at zero,

$$H(\mathbf{0}) + \frac{\partial H}{\partial x_1}(\mathbf{0})x_1 + \frac{\partial H}{\partial x_2}(\mathbf{0})x_2 + \frac{\partial H}{\partial x_3}(\mathbf{0})x_3 + \frac{\partial H}{\partial x_4}(\mathbf{0})x_4 + \frac{\partial H}{\partial x_5}(\mathbf{0})x_5.$$

Note that for $H(\mathbf{0}) = H_1(\mathbf{0})/\sqrt{H_2(\mathbf{0})}$, we have $H_1(\mathbf{0}) = 0$ and $H_2(\mathbf{0}) > 0$ (denote $\tilde{\sigma}_0^2 := H_2(\mathbf{0})$ and we will show it is positive over the course of derivations from (C.17) to (C.21). After simplifying the expression of $H_2(\mathbf{0})$, we give the explicit form of $\tilde{\sigma}_0^2$ below:

$$\begin{aligned} \tilde{\sigma}_0^2 = \frac{1}{\mathbb{E}[h^2(W)]}\Bigg[& \left(\frac{\mathbb{E}[Yh(W)]}{2(\mathbb{E}[h^2(W)])}\right)^2 \mathrm{Var}(\mathrm{Var}(h(W) \mid Z)) + \mathrm{Var}(Yh(W)) \\ & - \frac{\mathbb{E}[Yh(W)]}{\mathbb{E}[h^2(W)]}\mathrm{Cov}(Yh(W), \mathrm{Var}(h(W) \mid Z))\Bigg]. \end{aligned} \tag{C.12}$$

Using the chain rule of derivatives, we have for $m \in [5]$,

$$\frac{\partial H}{\partial x_m}(\mathbf{0}) = \frac{\partial H_1}{\partial x_m}(\mathbf{0})/\sqrt{H_2(\mathbf{0})} - \frac{H_1(\mathbf{0})}{2H_2(\mathbf{0})^{3/2}} \cdot \frac{\partial H_2}{\partial x_m}(\mathbf{0}) = \frac{\partial H_1}{\partial x_m}(\mathbf{0})/\tilde{\sigma}_0.$$

Since $H_1(x_1, x_2)$ only depends on x_1, x_2, we only need to evaluate two partial derivatives to compute the first order Taylor expansion of H at zero, yielding the following linear function

$$\frac{1}{\tilde{\sigma}_0}\left(\frac{1}{\sqrt{\mathbb{E}[h^2(W)]}}x_1 - \frac{\mathbb{E}[Yh(W)]}{2(\sqrt{\mathbb{E}[h^2(W)]})^3}x_2\right) := L(x), \tag{C.13}$$

which is denoted by $L(x) = L(x_1, x_2)$ and satisfies $L(\mathbf{0}) = 0$. Note that when $\epsilon = \mathbb{E}\left[h^2(W)\right]/2$, we have

$$\min_{\|x\|\le\epsilon}\left(x_2 + \mathbb{E}\left[h^2(W)\right]\right) = \mathbb{E}\left[h^2(W)\right] - \epsilon > 0.$$

Since $H_2(x)$ is continuous around zero and $H_2(\mathbf{0}) > 0$ (which will be shown in the following proof), we can similarly choose ϵ sufficiently small such that $\min_{\|x\|\le\epsilon} H_2(x) > 0$. Recall $H(x) = H_1(x)/\sqrt{H_2(x)}$, where H_1, H_2 are defined in (C.9) and (C.10), so $H(x)$ is continuous on $\{x : \|x\| \le \varepsilon\}$. Furthermore, its second partial derivatives exist and are continuous over the compact set $\{x : \|x\| \le \varepsilon\}$, thus are also bounded, which implies that there exists $M_\epsilon > 0$ such that (C.11) holds.

As for $\tilde{\sigma}$, ν_3, and ς_3, we will now establish the following moment bounds:

$$0 < \tilde{\sigma} := \|L(U_0)\|_2 < \infty, \tag{C.14}$$

$$\nu_2 := \|U_0\|_2, \qquad \nu_3 := \|U_0\|_3 < \infty,$$

$$\varsigma_3 := \frac{\|L(U_0)\|_3}{\tilde{\sigma}} < \infty. \tag{C.15}$$

Note that $\nu_3^3 = \|U_0\|_3^3 = \mathbb{E}\left[|U_{01}|^3\right] + \mathbb{E}\left[|U_{02}|^3\right] + \mathbb{E}\left[|U_{03}|^3\right] + \mathbb{E}\left[|U_{04}|^3\right] + \mathbb{E}\left[|U_{05}|^3\right]$ and

$$\begin{aligned}(\varsigma_3\tilde{\sigma})^3 = \mathbb{E}\left[|L(U_0)|^3\right] &= \frac{1}{\tilde{\sigma}_0^3}\mathbb{E}\left[\left|\frac{1}{\sqrt{\mathbb{E}\left[h^2(W)\right]}}U_{01} - \frac{\mathbb{E}\left[Yh(W)\right]}{2(\sqrt{\mathbb{E}\left[h^2(W)\right]})^3}U_{02}\right|^3\right] \\ &\le \frac{2^{3-1}}{\tilde{\sigma}_0^3}\left(\frac{1}{(\sqrt{\mathbb{E}\left[h^2(W)\right]})^3}\mathbb{E}\left[|U_{01}|^3\right] + \frac{(\mathbb{E}\left[Yh(W)\right])^3}{8(\sqrt{\mathbb{E}\left[h^2(W)\right]})^9}\mathbb{E}\left[|U_{02}|^3\right]\right)\end{aligned} \tag{C.16}$$

where the equalities hold due to the definitions of L and ς_3 in (C.13), (C.15), and the inequality holds as a result of the C_r inequality. Due to the fact that the finiteness of higher moments implies that of lower moments and (C.16), we only need to show

(i) $\mathbb{E}\left[|U_{01}|^3\right]$, $\mathbb{E}\left[|U_{02}|^3\right]$, $\mathbb{E}\left[|U_{03}|^3\right]$, $\mathbb{E}\left[|U_{04}|^3\right]$, $\mathbb{E}\left[|U_{05}|^3\right] < \infty$,

(ii) $\tilde{\sigma}_0^2 = H_2(\mathbf{0}) > 0$,

(iii) $\tilde{\sigma}^2 = \|L(U_0)\|_2 > 0$,

under the stated moment conditions. For (iii), actually we will show $\tilde{\sigma}^2 = 1$.

Starting with (i), we have

$$\begin{aligned}\mathbb{E}\left[|U_{02}|^3\right] = \mathbb{E}\left[|U_{i2}|^3\right] &= \mathbb{E}\left[\left|V_i - \mathbb{E}\left[h^2(W)\right]\right|^3\right] \\ &\le 2^{3-1}\left(\mathbb{E}\left[\left|\mathrm{Var}\left(\mu(W_i)\mid Z_i\right)\right|^3\right] + (\mathbb{E}\left[h^2(W)\right])^3\right) \\ &\le 2^{3-1}\left(\mathbb{E}\left[\mathbb{E}\left[h^6(W_i)\mid Z_i\right]\right] + (\mathbb{E}\left[h^2(W)\right])^3\right) < \infty,\end{aligned}$$

where the first inequality comes from the C_r inequality, the second holds by the definition of h and Jensen's inequality, and the third inequality holds due to the tower property of conditional expectation and $\mathbb{E}\left[h^6(W)\right] < \infty$ under the assumed moment conditions. For the term $\mathbb{E}\left[|U_{01}|^3\right]$, we have

$$\begin{aligned}\mathbb{E}\left[|U_{01}|^3\right] = \mathbb{E}\left[|U_{i1}|^3\right] &= \mathbb{E}\left[|R_i - \mathbb{E}\left[Yh(W)\right]|^3\right] \\ &\le 2^{3-1}\left(\mathbb{E}\left[|Y_i(\mu(W_i) - \mathbb{E}\left[\mu(W_i)\mid Z_i\right])|^3\right] + (\mathbb{E}\left[Yh(W)\right])^3\right) \\ &= 2^{3-1}\left(\mathbb{E}[|Y^3h^3(W)|] + (\mathbb{E}\left[Yh(W)\right])^3\right) < \infty,\end{aligned}$$

where the first inequality holds due to the C_r inequality and the second inequality holds due to the Cauchy–Schwarz inequality and the assumed moment conditions. The same approach and inequalities can be used for the other three terms, i.e., we have $\mathbb{E}\left[|U_{03}|^3\right], \mathbb{E}\left[|U_{04}|^3\right], \mathbb{E}\left[|U_{05}|^3\right] < \infty$. Note that U_{03}, U_{04}, and U_{05} involve higher-order polynomials of $Y_i h(W_i)$ and $\mathrm{Var}\,(h(W_i) \mid Z_i)$ than U_{01}, U_{02}, and thus require assuming bounded 12th moments to ensure the boundedness of their third absolute moments, hence the assumptions in Theorem C.1 that $\mathbb{E}\left[Y^{12}\right] < \infty$ and $\mathbb{E}\left[h^{12}(W)\right] < \infty$.

Regarding (ii) and (iii): recalling the definitions of $\tilde{\sigma}^2$ and L in (C.13), (C.14), we have

$$\begin{aligned}
\tilde{\sigma}_0^2\tilde{\sigma}^2 = \tilde{\sigma}_0^2\|L(U_0)\|_2 &= \frac{1}{\mathbb{E}\left[h^2(W)\right]}\mathbb{E}\left[\left(U_{i1} - \frac{\mathbb{E}\left[Yh(W)\right]}{2\mathbb{E}\left[h^2(W)\right]}U_{i2}\right)^2\right] \\
&= \mathbb{E}\left[\left(U_{i1} - \frac{\mathbb{E}\left[Yh(W)\right]}{2}U_{i2}\right)^2\right] \qquad \text{(C.17)} \\
&= \mathbb{E}\left[\left(R_i - \mathbb{E}\left[Yh(W)\right] - \frac{\mathbb{E}\left[Yh(W)\right]}{2}(\mathrm{Var}\,(h(W_i) \mid Z_i) - 1)\right)^2\right] \\
&= \mathbb{E}\left[(A+B)^2\right], \qquad \text{(C.18)}
\end{aligned}$$

where the third equality holds since $\mathbb{E}\left[h^2(W)\right] = 1$ as assumed without loss of generality, the fourth one comes from (C.8), and the last one is by rearranging with A, B defined as:

$$\begin{aligned}
A &:= Y_i h(W_i) - \mathbb{E}\left[Y_i h(W_i) \mid Z_i\right], \qquad \text{(C.19)} \\
B &:= \mathbb{E}\left[Y_i h(W_i) \mid Z_i\right] - \mathbb{E}\left[Yh(W)\right] - \frac{\mathbb{E}\left[Yh(W)\right]}{2}(\mathrm{Var}\,(h(W_i) \mid Z_i) - 1). \qquad \text{(C.20)}
\end{aligned}$$

The above terms A, B have equivalent expressions as the terms A, B defined in the proof of Theorem 2.3 (see (A.22), (A.23)). Note $\mathbb{E}\left[(A+B)^2\right] > 0$, as proved over the course of derivations from (A.20) to the end of the proof of Theorem 2.3. Due to (C.18), we then have $\tilde{\sigma}_0^2\tilde{\sigma}^2$ in this proof is nonzero, thus finish showing (ii).

Now we will verify $\tilde{\sigma} = 1$. According to (C.17), we equivalently write down

$$\begin{aligned}
\tilde{\sigma}_0^2\tilde{\sigma}^2 &= \mathbb{E}\left[\left(U_{i1} - \frac{\mathbb{E}\left[Yh(W)\right]}{2}U_{i2}\right)^2\right] \\
&= \mathbb{E}\left[\left(\left(1, -\frac{\mathbb{E}\left[Yh(W)\right]}{2}\right)(U_{i1}, U_{i2})^\top\right)^2\right] \\
&= \boldsymbol{a}^\top \Sigma_U \boldsymbol{a}, \qquad \text{(C.21)}
\end{aligned}$$

where $\boldsymbol{a}^\top := \left(1, -\frac{\mathbb{E}[Yh(W]}{2}\right)$ and Σ_U is the covariance matrix for the random vector U_i, which can be explicitly written as

$$\Sigma_U = \begin{pmatrix} \mathrm{Var}\,(Yh(W)) & \mathrm{Cov}\,(Yh(W), \mathrm{Var}\,(h(W) \mid Z)) \\ \mathrm{Cov}\,(Yh(W), \mathrm{Var}\,(h(W) \mid Z)) & \mathrm{Var}\,(\mathrm{Var}\,(h(W) \mid Z)) \end{pmatrix}.$$

We immediately have $\tilde{\sigma}_0^2\tilde{\sigma}^2 = \boldsymbol{a}^\top \Sigma_U \boldsymbol{a} = H_2(\mathbf{0}) = \tilde{\sigma}_0^2$ due to the expression of $\tilde{\sigma}_0^2$ in (C.12), (C.21) and $\mathbb{E}\left[h^2(W)\right] = 1$ as assumed; hence, $\tilde{\sigma} = 1$.

Having verified (i), (ii) and (iii), we thus prove the Berry–Esseen-type bound in (C.5), which completes the proof for case (IV). Therefore, the asymptotic coverage validity with a rate of $1/\sqrt{n}$ for the lower confidence bounds produced by Algorithm 1 has been established. □

C.1.2 Theorem C.2

Proof of Theorem C.2. Similarly as in the proofs of Theorem 2.3 and Theorem C.1, we immediately have coverage validity when $\mu(X,Z) \in \mathscr{A}(Z)$. Otherwise, it suffices to show

$$\inf_{K>1} \mathbb{P}\left(\frac{\bar{R}}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n}} \le f(\mu)\right) \ge 1 - \alpha - C/\sqrt{n} \tag{C.22}$$

for some constant C, where the sample mean $(\bar{R}, \bar{V})$ and sample covariance matrix $\hat{\Sigma}$ are defined the same way as in Algorithm 1 except that R_i, V_i are replaced by their Monte Carlo estimators R_i^K, V_i^K as defined below:

$$R_i^K = Y_i\left(\mu(X_i, Z_i) - \frac{1}{K}\sum_{k=1}^K \mu(X_i^{(k)}, Z_i)\right),$$

$$V_i^K = \frac{1}{K-1}\sum_{k=1}^K \left(\mu(X_i^{(k)}, Z_i) - \frac{1}{K}\sum_{k=1}^K \mu(X_i^{(k)}, Z_i)\right)^2,$$

Recall that the proof in Appendix A.1.2 considers 4 cases then deals with them separately. Essentially we can conduct similar analysis, but to avoid lengthy derivations, we focus on Case IV. Note we also make the extra assumption $\mathbb{E}\left[\mathrm{Var}\left(Y(\mu(X,Z) - \mathbb{E}\left[\mu(X,Z) \mid Z\right]) \mid Z\right)\right] > 0$ to simplify the proof.

In the proof of Theorem 2.5, we have the following asymptotic normality result:

$$\sqrt{n}\left(\frac{\frac{1}{n}\sum_{i=1}^n R_i^K}{\sqrt{\frac{1}{n}\sum_{i=1}^n V_i^K}} - f(\mu)\right) \xrightarrow{d} \mathcal{N}\left(0, \tilde{\sigma}_0^2\right).$$

To establish (C.22), we follow the proof strategy of Theorem C.1. Specifically, we apply the Berry–Esseen bound for nonlinear statistics (see Theorem C.6 in Appendix C.1.1).

Again we first introduce some new notations for the following proof: let random vectors $\{U_i\}_{i=1}^n = \{(U_{i1}, U_{i2}, U_{i3}, U_{i4}, U_{i5})\}_{i=1}^n \overset{i.i.d.}{\sim} U_0 = (U_{01}, U_{02}, U_{03}, U_{04}, U_{05})$ to be

$$U_{i1} = R_i^K - \mathbb{E}\left[Yh(W)\right], \quad U_{i2} = V_i^K - \mathbb{E}\left[h^2(W)\right], \tag{C.23}$$

$$U_{i3} = (R_i^K)^2 - \mathbb{E}\left[(R_i^K)^2\right], \quad U_{i4} = (V_i^K)^2 - \mathbb{E}\left[(V_i^K)^2\right], \quad U_{i5} = R_i^K V_i^K - \mathbb{E}\left[R_i^K V_i^K\right].$$

Note by the construction of the null samples, $X_i^{(k)}$ satisfy the two properties in (A.45) and (A.46) and we have (A.47), (A.48) hold. Recall (A.44) in the proof of Theorem 2.5 states $\mathbb{E}\left[R_i^K\right] = \mathbb{E}\left[Yh(W)\right], \mathbb{E}\left[V_i^K\right] = \mathbb{E}\left[h^2(W)\right]$, hence $\mathbb{E}\left[U_{i1}\right] = \mathbb{E}\left[U_{i2}\right] = 0$. Straightforwardly, $\mathbb{E}\left[U_{i3}\right] = \mathbb{E}\left[U_{i4}\right] = \mathbb{E}\left[U_{i5}\right] = 0$. Thus we have $\mathbb{E}\left[U_0\right] = \mathbf{0}$. Now we denote $\bar{U} = (\bar{U}_1, \bar{U}_2, \bar{U}_3, \bar{U}_4, \bar{U}_5) = \frac{1}{n}\sum_{i=1}^n U_i$ and rewrite the following expression,

$$\frac{1}{s}\left(\frac{\frac{1}{n}\sum_{i=1}^n R_i^K}{\sqrt{\frac{1}{n}\sum_{i=1}^n V_i^K}} - f(\mu)\right) := H(\bar{U}) := \frac{H_1(\bar{U}_1, \bar{U}_2)}{\sqrt{H_2(\bar{U})}},$$

where s is similarly defined as in Algorithm 1 except that R_i, V_i are replaced by R_i^K, V_i^K. Here $H(x) = H(x_1, x_2, x_3, x_4, x_5) : \mathbb{R}^5 \to \mathbb{R}$ is the same as in the proof of Theorem C.1. Therefore the smoothness condition, i.e., Condition (C.5), holds by the same argument as in Appendix C.1.1. The continuous linear functional L is also defined the same way. To apply Theorem C.6, it remains to verify the following moment bound conditions on U_0 and $L(U_0)$,

$$0 < \tilde{\sigma} := \|L(U_0)\|_2 < \infty,$$
$$\nu_2 := \|U_0\|_2, \ \nu_3 := \|U_0\|_3 < \infty,$$
$$\varsigma_3 := \frac{\|L(U_0)\|_3}{\tilde{\sigma}} < \infty.$$

Note that $\nu_3^3 = \|U_0\|_3^3 = \mathbb{E}\left[|U_{01}|^3\right] + \mathbb{E}\left[|U_{02}|^3\right] + \mathbb{E}\left[|U_{03}|^3\right] + \mathbb{E}\left[|U_{04}|^3\right] + \mathbb{E}\left[|U_{05}|^3\right]$ and we can bound $(\varsigma_3\tilde{\sigma})^3$ similarly as in the proof of Theorem C.1:

$$\begin{aligned}(\varsigma_3\tilde{\sigma})^3 &= \mathbb{E}\left[|L(U_0)|^3\right] = \mathbb{E}\left[\left|\frac{1}{\sqrt{\mathbb{E}\left[h^2(W)\right]}}U_{01} - \frac{\mathbb{E}\left[Yh(W)\right]}{2(\sqrt{\mathbb{E}\left[h^2(W)\right]})^3}U_{02}\right|^3\right] \\ &\leq 2^{3-1}\left(A\frac{1}{(\sqrt{\mathbb{E}\left[h^2(W)\right]})^3}\mathbb{E}\left[|U_{01}|^3\right] + \frac{(\mathbb{E}\left[Yh(W)\right])^3}{8(\sqrt{\mathbb{E}\left[h^2(W)\right]})^9}\mathbb{E}\left[|U_{02}|^3\right]\right), \end{aligned} \tag{C.24}$$

Due to the fact that the finiteness of higher moments implies that of lower moments and (C.24), we only need to show

(i) $\mathbb{E}\left[|U_{01}|^3\right], \mathbb{E}\left[|U_{02}|^3\right], \mathbb{E}\left[|U_{03}|^3\right], \mathbb{E}\left[|U_{04}|^3\right], \mathbb{E}\left[|U_{05}|^3\right] < \infty$

(ii) $\tilde{\sigma}_0^2 = H_2(\mathbf{0}) > 0$

(iii) $\tilde{\sigma}^2 = \|L(U_0)\|_2 > 0$

under the stated moment conditions. For (iii), we have $\tilde{\sigma}^2 = 1$, due to the derivations in the proof of Theorem C.1. Hence we will focus on the first two conditions in the following. Appendix C.1.1 verifies (i) and (ii) for any given $K > 1$. In this proof, we will actually show

$$\sup_{K>1}\mathbb{E}\left[|U_{0j}|^3\right] < \infty, \quad \forall j \in [5], \quad \inf_{K>1}\tilde{\sigma}_0^2 > 0.$$

Note the definitions of $U_0 = (U_{01}, U_{02}, U_{03}, U_{04}, U_{05})$ and $\tilde{\sigma}_0^2$ depend on K. To simplify notations, we do not make this dependence explicit. By the definitions in (C.23), we bound U_{01}, U_{02} as below:

$$\begin{aligned}\mathbb{E}\left[|U_{01}|^3\right] = \mathbb{E}\left[|U_{i1}|^3\right] &= \mathbb{E}\left[|R_i^K - \mathbb{E}\left[Yh(W)\right]|^3\right] \\ &\leq 2^{3-1}\left(\mathbb{E}\left[|R_i^K|^3\right] + (\mathbb{E}\left[Yh(W)\right])^3\right), \\ \mathbb{E}\left[|U_{02}|^3\right] = \mathbb{E}\left[|U_{i2}|^3\right] &= \mathbb{E}\left[|V_i^K - \mathbb{E}\left[h^2(W)\right]|^3\right] \\ &\leq 2^{3-1}\left(\mathbb{E}\left[|V_i^K|^3\right] + (\mathbb{E}\left[h^2(W)\right])^3\right),\end{aligned}$$

where the inequalities hold due to the C_r inequality. Recalling in the proof of Theorem 2.5, we show $\mathbb{E}\left[|R_i^K|^2\right] < \infty, \mathbb{E}\left[|V_i^K|^2\right] < \infty$ under the condition $\mathbb{E}\left[Y^4\right], \mathbb{E}\left[h^4(W)\right] < \infty$ over the course of derivations from (A.50) to the end of that proof. The derivations are mainly based on the C_r inequality and the Bahr–Esseen inequality in Dharmadhikari et al. (1969). Using the same bounding strategy, we can show $\mathbb{E}\left[|R_i^K|^3\right], \mathbb{E}\left[|V_i^K|^3\right] < \infty$ when assuming $\mathbb{E}\left[Y^6\right], \mathbb{E}\left[h^6(W)\right] < \infty$. Hence we obtain $\sup_{K>1}\mathbb{E}\left[|U_{01}|^3\right], \sup_{K>1}\mathbb{E}\left[|U_{02}|^3\right] < \infty$ under the above moment conditions. And nearly identical derivations as in bounding $\mathbb{E}\left[|U_{01}|^3\right]$ and $\mathbb{E}\left[|U_{02}|^3\right]$ suffice to show $\sup_{K>1}\mathbb{E}\left[|U_{03}|^3\right]$, $\sup_{K>1}\mathbb{E}\left[|U_{04}|^3\right]$, $\sup_{K>1}\mathbb{E}\left[|U_{05}|^3\right] < \infty$ under the stronger moment boundedness conditions $\mathbb{E}\left[Y^{12}\right] < \infty, \mathbb{E}\left[h^{12}(W)\right] < \infty$ stated in Theorem 2.5.

Regarding (ii), we notice that

$$\tilde{\sigma}_0^2 \geq \mathbb{E}\left[(A+B)^2\right] \geq \mathbb{E}\left[\mathrm{Var}\left(Yh(W)\,|\,Z\right)\right], \tag{C.25}$$

where the first inequality holds due to (A.62), A, B are defined as (A.22) and (A.23) in the proof of Theorem 2.3, and the second inequality holds by (A.24). The above lower bound for $\tilde{\sigma}_0^2$ does not depend on K and implies the positiveness of $\inf_{K>1}\tilde{\sigma}_0$ under the assumed condition $\mathbb{E}\left[\mathrm{Var}\left(Yh(W)\,|\,Z\right)\right] = \mathbb{E}\left[\mathrm{Var}\left(Y(\mu(X,Z) - \mathbb{E}\left[\mu(X,Z)\,|\,Z\right])\,|\,Z\right)\right] > 0$.

Therefore, we obtain the Berry–Esseen bound for nonlinear statistics by applying Theorem C.6. Finally we conclude the asymptotic coverage with a rate of $n^{-1/2}$, i.e.,

$$\inf_{K>1}\mathbb{P}\left(L_{n,K}^{\alpha}(\mu) \leq \mathcal{I}\right) \geq 1 - \alpha - Cn^{-1/2},$$

where the constant C only depends on the moments of Y and $h(X,Z) = \mu(X,Z) - \mathbb{E}\left[\mu(X,Z)\,|\,Z\right]$. □

D Applicability of the Model-X assumption

Model-X floodgate assumes knowing the distribution of $P_{X|Z}$. This may not always hold in practice, but in some important instances, $P_{X|Z}$ may be (A) known due to experimental randomization, (B) well-modeled a priori due to domain expertise, or (C) accurately estimated from a large unlabeled data set. For example, (A) holds in the high-dimensional experiments of conjoint analysis (Luce and Tukey, 1964; Hainmueller and Hopkins, 2014), (B) holds in the study of the microbiome where accurate covariate simulators exist (Ren et al., 2016), and a combination of (B) and (C) hold in genomics, where the model-X framework has been repeatedly and successfully applied for controlled variable selection (Sesia et al., 2019; Katsevich and Sabatti, 2019; Sesia et al., 2020b; Bates et al., 2020; Sesia et al., 2020a).

We also quantify the robustness of our inferences to this assumption in Appendix E and show it can be relaxed to parametric models (Section 3.2), and indeed model-X approaches have shown promising empirical performance in a number of applications in which it is unclear whether any of (A), (B), or (C) hold, such as bacterial classification from spectroscopic data (Chia et al., 2020) and single cell regulatory screening (Katsevich and Roeder, 2020).

E Robustness

To explain how the floodgate idea is not tied to the model-X assumption, a double-robustness type result (Lemma 2.3) is presented in Remark 2.3.1. It involves an approximated floodgate functional (2.8) and says that the inferential statements are valid as long as either of the models of $X \mid Z$ or $Y \mid Z$ is correctly specified. For ease of exposition, Algorithm 1 and Theorem 2.3 focus on a particular floodgate procedure which requires knowing $P_{X|Z}$. However, it is still of interest to study the robustness of floodgate (in Algorithm 1) to misspecification of $P_{X|Z}$. Specifically, we consider the case when the true distribution $P_{X|Z}$ used in floodgate is replaced by an approximation $Q_{X|Z}$.

Notationally, let $Q = P_{Y|X,Z} \times Q_{X|Z} \times P_Z$ (we need not consider misspecification in the distributions of Z or $Y \mid X, Z$ since these are not inputs to floodgate), and let f^Q be an analogue of f with certain expectations replaced by expectations over Q (we will denote such expectations by $\mathbb{E}_Q[\cdot]$); see Equation (E.5) for a formal definition. It is not hard to see that floodgate with input $Q_{X|Z}$ produces an asymptotically-valid LCB for $f^Q(\mu)$, from which we immediately draw the following conclusions.

First, if μ does not actually depend on X, i.e., $\mathrm{Var}_Q(\mu(X,Z) \mid Z) \overset{a.s.}{=} 0$, then $f^Q(\mu) = 0$ regardless of Q and floodgate is trivially asymptotically-valid. Second, when μ does depend on X, floodgate's inference will still be approximately valid as long as $f^Q(\mu) - f(\mu) \approx 0$, and this difference can be bounded by, for instance, the χ^2 divergence between $P_{X|Z}$ and $Q_{X|Z}$. The third, and perhaps most interesting, conclusion is that the gap between $\mathcal{I}$ and $f(\mu)$ grants floodgate an *extra* layer of robustness as long as $\mathcal{I} - f(\mu)$ is large compared to $f^Q(\mu) - f(\mu)$. Thus even if $Q_{X|Z}$ is a bad approximation of $P_{X|Z}$, floodgate's inference may be saved if $f(\mu)$ is an *even worse* approximation of $\mathcal{I}$, and this latter approximation is related to that of μ for $\mu^\star$. To make this last relation precise, we quantify μ's approximation of $\mu^\star$ by focusing on a particular representative of S_μ: for any $\mu : \mathbb{R}^p \to \mathbb{R}$,

$$\bar{\mu}(x,z) = \sqrt{\frac{\mathbb{E}\left[\mathrm{Var}(\mu^\star(X,Z) \mid Z)\right]}{\mathbb{E}\left[\mathrm{Var}(\mu(X,Z) \mid Z)\right]}} \Big(\mu(x,z) - \mathbb{E}\left[\mu(X,Z) \mid Z = z\right]\Big) + \mathbb{E}\left[\mu^\star(X,Z) \mid Z = z\right], \tag{E.1}$$

where $0/0 = 0$. We can think of $\bar{\mu}$ as a generally accurate representative from S_μ, in that it takes μ and corrects its conditional mean and expected conditional variance to match $\mu^\star$. Note that $\bar{\mu} = \mu^\star$ whenever $\mu^\star \in S_\mu$, which includes anytime $\mathcal{I} = 0$. Since the LCB from floodgate with input $Q_{X|Z}$ is asymptotically-valid for $f^Q(\mu)$ under certain moment conditions and the proof can be done similarly as Theorem 2.3, we will focus on quantifying the difference between $f^Q(\mu)$ and $\mathcal{I}$ in the following robustness result.

Theorem E.1 (Floodgate robustness). *For data $\{(Y_i, X_i, Z_i)\}_{i=1}^n$ i.i.d. draws from P satisfying $\mathbb{E}[Y^4] < \infty$, a sequence of working regression functions $\mu_n : \mathbb{R}^p \to \mathbb{R}$ such that for some C and all n either*

$\mathrm{Var}_{Q^{(n)}}(\mu_n(X,Z)\,|\,Z) \overset{a.s.}{=} 0$ *or* $\frac{\max\left\{\mathbb{E}\left[\mu_n^4(X,Z)\right], \mathbb{E}_{Q^{(n)}}\left[\mu_n^4(X,Z)\right]\right\}}{\mathbb{E}\left[\mathrm{Var}_{Q^{(n)}}(\mu_n(X,Z)\,|\,Z)\right]^2} \le C$*, and a sequence of conditional distributions* $Q^{(n)}_{X|Z}$*, the difference between* $f^{Q^{(n)}}(\mu)$ *and* $\mathcal{I}$ *can be controlled as*

$$\Delta_n = f^{Q^{(n)}}(\mu_n) - \mathcal{I} \le c_1 \sqrt{\mathbb{E}\left[\chi^2\left(P_{X|Z} \,\|\, Q^{(n)}_{X|Z}\right)\right]} - c_2\,\mathbb{E}\left[(\bar{\mu}_n(X,Z) - \mu^\star(X,Z))^2\right] \tag{E.2}$$

for some positive c_1 *and* c_2 *that depend on* P*, where* $\chi^2(\cdot\,\|\,\cdot)$ *denotes the* χ^2 *divergence.*

The proof of Theorem E.1 can be found in Appendix E.1. Equation (E.2) formalizes that larger MSE of $\bar{\mu}_n$ actually *improves* robustness, although we remind the reader once again that when $\mathcal{I} = 0$, the MSE of $\bar{\mu}_n$ is always zero by construction in Equation (E.1). Given the $n^{-1/2}$-rate half-width lower-bound for floodgate, a sufficient condition for asymptotically-exact coverage is

$$\sqrt{\mathbb{E}\left[\chi^2\left(P_{X|Z} \,\|\, Q^{(n)}_{X|Z}\right)\right]} = o\left(n^{-1/2} + \mathbb{E}\left[(\bar{\mu}_n(X,Z) - \mu^\star(X,Z))^2\right]\right). \tag{E.3}$$

When $Q^{(n)}_{X|Z}$ is a standard parametric estimator based on N_n independent samples, the left-hand side has a $O(N_n^{-1/2})$ rate. Thus if $N_n \gg \min\{n, \mathbb{E}\left[(\bar{\mu}_n(X,Z) - \mu^\star(X,Z))^2\right]^{-2}\}$, then floodgate's coverage will be asymptotically-exact. For certain parametric models for $X \mid Z$, Section 3.2 shows how to modify floodgate to attain asymptotically-exact inference without the need for estimation at all.

Theorem E.1 treats the sequence $Q^{(n)}_{X|Z}$ as fixed, which of course means $Q^{(n)}_{X|Z}$ can be estimated from any data that is independent of the data floodgate is applied to. This means the same data can be used to estimate μ_n and $Q^{(n)}_{X|Z}$. For $Q^{(n)}_{X|Z}$ however, this strict separation may not be necessary in practice, and in our simulations we found floodgate to be quite robust to estimating $Q^{(n)}_{X|Z}$ on samples that included those used as input to floodgate; see Section 4.5.

Another layer of robustness beyond that addressed in this section can be injected by replacing $P_{X|Z}$ in floodgate with $P_{X|Z,T}$ for some random variable T. For instance, floodgate's model-X assumption can be formally relaxed to only needing to know a fixed-dimensional model for $P_{X|Z}$ by conditioning on T that is a sufficient statistic for that model; see Section 3.2 for details. More generally, conditioning on T that is a function of $\{(X,Z)\}_{i=1}^n$ may induce some degree of robustness, as conditioning on the order statistics of the X_i can in conditional independence testing (Berrett et al., 2020).

E.1 Proofs in Appendix E

In the case where the conditional distribution of X given Z is specified as $Q_{X|Z}$ (in the following, we often denote the true conditional distribution by $P := P_{X|Z}$ and the specified conditional distribution by $Q := Q_{X|Z}$ without causing confusion), the floodgate functional with input $Q_{X|Z}$ is denoted by $f^Q(\mu)$. Note that $f(\mu)$ can be rewritten with explicit subscripts as below (here we use the equivalent expression of $f(\mu)$ in (A.7) and expand $h(W)$).

$$f(\mu) = \frac{\mathbb{E}_P\left[Y\left(\mu(X,Z) - \mathbb{E}_P\left[\mu(X,Z)\,|\,Z\right]\right)\right]}{\sqrt{\mathbb{E}_{P_Z}\left[\mathrm{Var}_P\left(\mu(X,Z)\,|\,Z\right)\right]}} \tag{E.4}$$

Therefore, $f^Q(\mu)$ admits the following expression:

$$f^Q(\mu) := \frac{\mathbb{E}_P\left[Y\left(\mu(X,Z) - \mathbb{E}_Q\left[\mu(X,Z)\,|\,Z\right]\right)\right]}{\sqrt{\mathbb{E}_{P_Z}\left[\mathrm{Var}_Q\left(\mu(X,Z)\,|\,Z\right)\right]}}. \tag{E.5}$$

Denote $\omega(x,z) := \frac{dP_{X|Z}(x|z)}{dQ_{X|Z}(x|z)}$. Note that $\omega(x,z)$ is the ratio of conditional densities if we are in the continuous case; $\omega(x,z)$ is the ratio of conditional probability mass function in discrete case. Then we can quantify the difference between $f(\mu)$ and $f^Q(\mu)$ as in Lemma E.2.

Lemma E.2. *Assuming $\mathbb{E}\left[Y^4\right] < \infty$, consider two joint distributions P, Q over (X, Z), defined as $P(x, z) = P_{X|Z}(x|z)P_Z(z), Q(x, z) = Q_{X|Z}(x|z)P_Z(z)$. If we denote $\mathcal{U}$ to be the class of functions $\mu : \mathbb{R}^p \to \mathbb{R}$ satisfying one of the following conditions:*

- $\mu(X, Z) \in \mathscr{A}(Z)$;
- $\max\{\mathbb{E}_P\left[\mu^4(X, Z)\right], \mathbb{E}_Q\left[\mu^4(X, Z)\right]\}/(\mathbb{E}_{P_Z}\left[\mathrm{Var}_Q\left(\mu(X, Z) \mid Z\right)\right])^2 \le c_0$.

for some constants c_0, then we have the following bounds

$$\Delta(P, Q) := \sup_{\mu \in \mathcal{U}} |\theta^Q(\mu) - f(\mu)| \le C\sqrt{\mathbb{E}_{P_Z}\left[\chi^2\left(P_{X|Z} \| Q_{X|Z}\right)\right]} \tag{E.6}$$

for some constant C only depending on $\mathbb{E}\left[Y^4\right]$ and c_0, where the χ^2 divergence between two distributions P, Q on the probability space Ω is defined as $\chi^2\left(P \| Q\right) := \int_\Omega (\frac{dP}{dQ} - 1)^2 dQ$.

When the $X \mid Z$ model is misspecified, the inferential validity will not hold in general, without adjustment on the lower confidence bound. Lemma E.2 gives a quantitative characterization about how much we need to adjust.

Proof of Lemma E.2. When the support of Q does not contain the support of P, the χ^2 divergence between P and Q is infinite, which immediately proves (E.6). From now, we work with the case where the support of Q contains the support of P. When $\mu(X, Z) \in \mathscr{A}(Z)$, $f(\mu) = f^Q(\mu) = 0$, thus the statement holds. Now we deal with the nontrivial case where $\mathbb{E}_{P_Z}\left[\mathrm{Var}_Q\left(\mu(X, Z) \mid Z\right)\right] > 0$. Without loss of generality, we assume $\mathbb{E}_{P_Z}\left[\mathrm{Var}_Q\left(\mu(X, Z) \mid Z\right)\right] = 1$ for the following proof (since floodgate is invariate to positive scaling of μ). Then the stated moment conditions on μ imply

$$\mathbb{E}_P\left[\mu^4(X, Z)\right], \mathbb{E}_Q\left[\mu^4(X, Z)\right] \le c_0. \tag{E.7}$$

First we simplify $f(\mu)$ and $f^Q(\mu)$ into

$$\begin{aligned} f(\mu) &= \frac{\mathbb{E}_P\left[\mu^\star(X, Z)\left(\mu(X, Z) - \mathbb{E}_{P_{X|Z}}\left[\mu(X, Z) \mid Z\right]\right)\right]}{\sqrt{\mathbb{E}_{P_Z}\left[\mathrm{Var}_{P_{X|Z}}\left(\mu(X, Z) \mid Z\right)\right]}} = \frac{\mathbb{E}_P\left[\mu^\star(W)\left(\mu(W) - \mathbb{E}_P\left[\mu(W) \mid Z\right]\right)\right]}{\sqrt{\mathbb{E}_{P_Z}\left[\mathrm{Var}_P\left(\mu(W) \mid Z\right)\right]}} \\ f^Q(\mu) &= \frac{\mathbb{E}_P\left[\mu^\star(X, Z)\left(\mu(X, Z) - \mathbb{E}_{Q_{X|Z}}\left[\mu(X, Z) \mid Z\right]\right)\right]}{\sqrt{\mathbb{E}_{P_Z}\left[\mathrm{Var}_{Q_{X|Z}}\left(\mu(X, Z) \mid Z\right)\right]}} = \frac{\mathbb{E}_P\left[\mu^\star(W)\left(\mu(W) - \mathbb{E}_Q\left[\mu(W) \mid Z\right]\right)\right]}{\sqrt{\mathbb{E}_{P_Z}\left[\mathrm{Var}_Q\left(\mu(W) \mid Z\right)\right]}} \end{aligned}$$

due to (A.4), where we denote $W = (X, Z)$ (thus $w = (x, z)$). Noticing the following facts

$$\left|\frac{a}{\sqrt{b}} - \frac{c}{\sqrt{d}}\right| = \left|\frac{a\sqrt{d} - c\sqrt{b}}{\sqrt{bd}}\right| \le \frac{a}{\sqrt{bd}}\left|\sqrt{b} - \sqrt{d}\right| + \frac{1}{\sqrt{d}}|a - c| \le \frac{a}{\sqrt{b}} \cdot \frac{1}{d}|b - d| + \frac{1}{\sqrt{d}}|a - c|,$$

we let a, c to be the numerators of $f(\mu)$ and $f^Q(\mu)$ respectively and $\sqrt{b}, \sqrt{d}$ to be their denominators. Before dealing with $|b - d|$ and $|c - d|$, we have the following bounds on the terms $a/\sqrt{b}$ and $1/d$.

$$a/\sqrt{b} = f(\mu) \le \mathcal{I} \le (\mathbb{E}_P\left[Y^4\right])^{1/4}, \quad 1/d = 1/\mathbb{E}_{P_Z}\left[\mathrm{Var}_Q\left(\mu(X, Z) \mid Z\right)\right] = 1, \tag{E.8}$$

where the first equality is due to Lemma 2.2 and the second one is by applying Jensen's inequality $(\mathbb{E}_{P_Z}\left[\mathrm{Var}_P\left(\mathbb{E}\left[Y \mid X, Z\right] \mid Z\right)\right] \le \mathbb{E}_{P_Z}\left[\mathbb{E}_P\left[(\mathbb{E}\left[Y \mid X, Z\right])^2 \mid Z\right]\right] \le \mathbb{E}\left[Y^2\right] \le \sqrt{\mathbb{E}\left[Y^4\right]})$. The equality holds by assumption. Now it suffices to consider bounding $|b-d|$ and $|c-d|$ in terms of the expected χ^2 divergence between $P_{X|Z}$ and $Q_{X|Z}$. We have the following equations for $|a-c|$:

$$\begin{aligned}
|a-c| &= |\mathbb{E}_P\left[\mu^\star(W)\left(\mu(W)-\mathbb{E}_P\left[\mu(W)\,|\,Z\right]\right)\right]-\mathbb{E}_P\left[\mu^\star(W)\left(\mu(W)-\mathbb{E}_Q\left[\mu(W)\,|\,Z\right]\right)\right]| \\
&= |\mathbb{E}_P\left[\mu^\star(W)\left(\mathbb{E}_P\left[\mu(W)\,|\,Z\right]-\mathbb{E}_Q\left[\mu(W)\,|\,Z\right]\right)\right]| \\
&= |\mathbb{E}_{P_Z}\left[\mathbb{E}_P\left[\mu^\star(W)\,|\,Z\right]\left(\mathbb{E}_P\left[\mu(W)\,|\,Z\right]-\mathbb{E}_Q\left[\mu(W)\,|\,Z\right]\right)\right]|\,. \qquad \text{(E.9)}
\end{aligned}$$

Now we rewrite $|\mathbb{E}_P\left[\mu(W)\,|\,Z\right]-\mathbb{E}_Q\left[\mu(W)\,|\,Z\right]|$ in the form of integral then bound it as

$$\begin{aligned}
|\mathbb{E}_P\left[\mu(W)\,|\,Z\right]-\mathbb{E}_Q\left[\mu(W)\,|\,Z\right]| &= \left|\int \mu(x,Z)(1-\omega(x,Z))dQ_{X|Z}(x\mid Z)\right| \\
&\leq \sqrt{\mathbb{E}_{Q_{X|Z}}\left[\mu^2(X,Z)\,|\,Z\right]}\sqrt{\int(1-w(x,Z))^2dQ_{X|Z}(x\mid Z)} \\
&= \sqrt{\mathbb{E}_{Q_{X|Z}}\left[\mu^2(W)\,|\,Z\right]}\sqrt{\chi^2\left(P_{X|Z}\|Q_{X|Z}\right)}, \qquad \text{(E.10)}
\end{aligned}$$

where $\omega(x,Z)=\frac{dP_{X|Z}(x|Z)}{dQ_{X|Z}(x|Z)}$ and the above inequality is from the Cauchy–Schwarz inequality. Hence we can plug (E.10) into (E.9) and further bound $|a-c|$ by

$$\begin{aligned}
|a-c| \leq\ & \mathbb{E}_{P_Z}\left[\mathbb{E}_{P_{X|Z}}\left[\mu^\star(W)\,|\,Z\right]\sqrt{\mathbb{E}_{Q_{X|Z}}\left[\mu^2(W)\,|\,Z\right]}\sqrt{\chi^2\left(P_{X|Z}\|Q_{X|Z}\right)}\right] \\
\leq\ & \sqrt{\mathbb{E}_{P_Z}\left[(\mathbb{E}_{P_{X|Z}}\left[\mu^\star(W)\,|\,Z\right])^2\mathbb{E}_{Q_{X|Z}}\left[\mu^2(W)\,|\,Z\right]\right]}\cdot\sqrt{\mathbb{E}_{P_Z}\left[\chi^2\left(P_{X|Z}\|Q_{X|Z}\right)\right]}. \qquad \text{(E.11)}
\end{aligned}$$

For the first part of the product in (E.11), we can apply the Cauchy–Schwarz inequality and Jensen's inequality and bound it by $(\mathbb{E}_P\left[(\mu^\star)^4(W)\right]\mathbb{E}_Q\left[\mu^4(W)\right])^{1/4}$, which is upper bounded by some constant under the stated condition $\mathbb{E}\left[Y^4\right]<\infty$ and $\mathbb{E}_Q\left[\mu^4(X,Z)\right]\leq c_0$ (from (E.7)). Regarding $|b-d|$, we have

$$\begin{aligned}
|b-d| &= |\mathbb{E}_{P_Z}\left[\mathrm{Var}_P\left(\mu(W)\,|\,Z\right)\right]-\mathbb{E}_{P_Z}\left[\mathrm{Var}_Q\left(\mu(X,Z)\,|\,Z\right)\right]| \\
&\leq \left|\mathbb{E}_{P_Z}\left[(\mathbb{E}_P\left[\mu(W)\,|\,Z\right])^2-(\mathbb{E}_Q\left[\mu(W)\,|\,Z\right])^2\right]\right| \\
&\quad +\left|\mathbb{E}_{P_Z}\left[\mathbb{E}_P\left[\mu^2(W)\,|\,Z\right]-\mathbb{E}_Q\left[\mu^2(W)\,|\,Z\right]\right]\right|. \qquad \text{(E.12)}
\end{aligned}$$

Similarly as (E.10), we obtain

$$\left|\mathbb{E}_P\left[\mu^2(W)\,|\,Z\right]-\mathbb{E}_Q\left[\mu^2(W)\,|\,Z\right]\right|\leq\sqrt{\mathbb{E}_{Q_{X|Z}}\left[\mu^4(W)\,|\,Z\right]}\sqrt{\chi^2\left(P_{X|Z}\|Q_{X|Z}\right)}.$$

Then under the moment bounds $\mathbb{E}_Q\left[\mu^4(X,Z)\right]\leq c_0$ in (E.7), we show the second term in (E.12) is upper bounded by $\sqrt{c_0\mathbb{E}_{P_Z}\left[\chi^2\left(P_{X|Z}\|Q_{X|Z}\right)\right]}$. Regarding the first term in (E.12), we can write

$$(\mathbb{E}_P\left[\mu(W)\,|\,Z\right])^2-(\mathbb{E}_Q\left[\mu(W)\,|\,Z\right])^2=\left(\mathbb{E}_P\left[\mu(W)\,|\,Z\right]-\mathbb{E}_Q\left[\mu(W)\,|\,Z\right]\right)\left(\mathbb{E}_P\left[\mu(W)\,|\,Z\right]+\mathbb{E}_Q\left[\mu(W)\,|\,Z\right]\right)$$

then apply similar strategies in deriving (E.9) and (E.11) to control the above term under $C\sqrt{\mathbb{E}_{P_Z}\left[\chi^2\left(P_{X|Z}\|Q_{X|Z}\right)\right]}$ for some constant C. And this will make use of the moment bound conditions $\mathbb{E}_P\left[\mu^4(X,Z)\right]$,$\mathbb{E}_Q\left[\mu^4(X,Z)\right]\leq c_0$ in (E.7). Finally we establish the bound in (E.6). □

Proof of Theorem E.1. First notice that Δ_n can be decomposed into two parts:

$$\Delta_n=f^{Q^{(n)}}(\mu_n)-\mathcal{I}=(f^{Q^{(n)}}(\mu_n)-f(\mu_n))-(\mathcal{I}-f(\mu_n)). \qquad \text{(E.13)}$$

In the following, we will deal with $f^{Q^{(n)}}(\mu_n) - f(\mu_n)$ and $\mathcal{I} - f(\mu_n)$ separately. Applying Lemma E.2 to P, $Q^{(n)}$ and μ_n under the stated conditions gives

$$(f^{Q^{(n)}}(\mu_n) - f(\mu_n)) \leq c_1 \sqrt{\mathbb{E}\left[\chi^2\left(P_{X|Z} \,\|\, Q^{(n)}_{X|Z}\right)\right]} \tag{E.14}$$

for some constant c_1 only depending on $\mathbb{E}\left[Y^4\right]$ and c_0. Regarding the term $\mathcal{I} - f(\mu_n)$, we recall the derivations in the proof of Theorem 2.6, specifically (A.69) and (A.70), then obtain

$$\mathcal{I} - f(\mu_n) \geq \frac{\mathbb{E}\left[(\bar{h}_n(W) - h^\star(W))^2\right]}{2\mathcal{I}} = \frac{\mathbb{E}\left[(\bar{\mu}_n(W) - \mu^\star(W))^2\right]}{2\mathcal{I}}, \tag{E.15}$$

where the equality holds by the definition of $h^\star$, $\bar{\mu}_n$ and $\bar{h}_n$. Combining (E.13), (E.14) and (E.15) yields (E.2). □

F Details of extending the mMSE gap

F.1 Taking the supremum over transformations

Drawing inspiration from the maximum correlation coefficient (Hirschfeld, 1935), taking the supremum of the mMSE gap over transformations of Y leads to other desirable properties. For a set $\mathcal{G}$ of functions g mapping Y to its sample space, let $\mathcal{I}_{\mathcal{G}} = \sup_{g\in\mathcal{G}} \mathcal{I}_{\text{sf}}(g(Y))$, where $\mathcal{I}_{\text{sf}}(g(Y))$ denotes the scale-free version of the mMSE gap when Y is replaced by $g(Y)$. Then for any fixed function $g \in \mathcal{G}$, floodgate's LCB for $\mathcal{I}_{\text{sf}}(g(Y))$ is also an asymptotically valid LCB for $\mathcal{I}_{\mathcal{G}}$. And like μ, g can be chosen based on an independent split of the data to make the gap between $\mathcal{I}_{\text{sf}}(g(Y))$ and $\mathcal{I}_{\mathcal{G}}$ as small as possible. If $\mathcal{G}$ forms a group, then it is immediate that $\mathcal{I}_{\mathcal{G}}$ takes the same value when $g(Y)$ is used as the response, for any $g \in \mathcal{G}$, i.e., $\mathcal{I}_{\mathcal{G}}$ is invariant to any transformation $g \in \mathcal{G}$ of Y. For instance, we might choose $\mathcal{G}$ to be the group of all strictly monotone functions, or of all bijections. Regardless of whether $\mathcal{G}$ is a group or not, if it is large enough that it contains all bounded continuous functions then, by the Portmanteau Theorem, $\mathcal{I}_{\mathcal{G}}$ will be zero if *and only if* $Y \perp\!\!\!\perp X \mid Z$. That is, for sufficiently large $\mathcal{G}$, $\mathcal{I}_{\mathcal{G}}$ satisfies the key property of the MOVI in Azadkia and Chatterjee (2019) and floodgate provides asymptotically valid inference for it. A natural choice[1] of $\mathcal{G}$ satisfying such property is $\{\mathbb{1}_{\{y\leq t\}} : t \in \mathbb{R}\}$ as

$$\mathcal{I}_{\mathcal{G}} = \sup_{t\in\mathbb{R}} \frac{\mathbb{E}\left[\text{Var}\left(\mathbb{E}\left[\mathbb{1}_{\{Y\leq t\}} \,|\, X, Z\right] \,|\, Z\right)\right]}{\text{Var}\left(\mathbb{1}_{\{Y\leq t\}}\right)}.$$

The above quantity is related to the measure of conditional dependence in Azadkia and Chatterjee (2019) as both involve $\mathbb{E}\left[\text{Var}\left(\mathbb{E}\left[\mathbb{1}_{\{Y\leq t\}} \,|\, X, Z\right] \,|\, Z\right)\right]$.

F.2 Extending via the RKHS framework

A reviewer pointed out a very interesting work (Huang et al., 2020) which came out after our arXiv preprint. To handle X, Y, Z from general topological spaces, Huang et al. (2020) proposes the kernel partial correlation coefficient (KPC) to measure conditional dependence and provides consistent estimation methods. Huang et al. (2020) mentioned the numerator of KPC with a linear kernel equals to the mMSE gap considered in our paper. In this section, we discuss how to extend the floodgate inferential approach via reproducing kernel Hilbert spaces (RKHS) to apply to the KPC. For ease of exposition, we focus on the numerator of KPC and call it the average kernel maximum mean discrepancy (AKMMD). Note that the AKMMD with a characteristic kernel will be zero if *and only if* $Y \perp\!\!\!\perp X \mid Z$.

[1] We are grateful to an anonymous reviewer for suggesting this choice.

Recall the equivalent expression of the mMSE gap in (2.4)

$$\mathcal{I}^2 = \mathbb{E}\left[(\mathbb{E}[Y \mid X, Z] - \mathbb{E}[Y \mid Z])^2\right],$$

where $\mathbb{E}[Y \mid X, Z]$ can be viewed as the kernel embedding of $P_{Y|X,Z}$ under a special linear kernel. Then $\mathcal{I}^2$ essentially quantifies the distance between $P_{Y|X,Z}$ and $P_{Y|Z}$ via the maximum mean discrepancy (MMD). To extend this idea using a general kernel, we introduce some new notations and preliminary concepts about RKHS. Suppose (Y, X, Z) take values in some topological space $\mathcal{Y} \times \mathcal{X} \times \mathcal{Z}$ and let P be the joint distribution over (Y, X, Z). The marginal distribution of Y is denoted by P_Y. Sometimes this subscript is dropped when doing so does not cause confusion. We use the bold $\boldsymbol{\mu}$ notation for kernel mean embeddings, which should be differentiated from the working regression function in the main text. Denote by $\mathcal{H}_{\mathcal{Y}}$ an RKHS with kernel $\mathcal{K}(\cdot, \cdot)$ on the space $\mathcal{Y}$, where $\mathcal{K} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a symmetric and positive semidefinite function such that $\mathcal{K}(\cdot, y)$ is measurable function on $\mathcal{Y}, \forall\ y \in \mathcal{Y}$. The inner product and norm on the RKHS $\mathcal{H}_{\mathcal{Y}}$ are denoted by $\langle \cdot, \cdot \rangle_{\mathcal{H}_{\mathcal{Y}}}$ and $|| \cdot ||_{\mathcal{H}_{\mathcal{Y}}}$, with the subscripts often dropped for simplicity. The kernel reproducing property implies that $h(y) = \langle \mathcal{K}(\cdot, y), h \rangle_{\mathcal{H}_{\mathcal{Y}}}$. First we introduce the definitions of the kernel mean embedding and the MMD (Deb et al., 2020; Huang et al., 2020).

Definition F.1 (Kernel mean embedding)**.** *Suppose* $Y \sim P_Y$ *and* $\mathbb{E}_P\left[\sqrt{\mathcal{K}(Y, Y)}\right] < \infty$. *There exists (Deb et al., 2020; Huang et al., 2020) a unique* $\boldsymbol{\mu}_P \in \mathcal{H}_{\mathcal{Y}}$ *satisfying*

$$\langle \boldsymbol{\mu}_P, h \rangle_{\mathcal{H}_{\mathcal{Y}}} = \mathbb{E}_P[h(Y)], \quad \textit{for all } h \in \mathcal{H}_{\mathcal{Y}},$$

which is called the kernel mean embedding of P_Y *into* $\mathcal{H}_{\mathcal{Y}}$.

Definition F.2 (Maximum mean discrepancy)**.** *We measure the distance between two distributions* P_1, P_2 *via the MMD (with respect to the kernel* $\mathcal{K}(\cdot, \cdot)$*), defined as*

$$\mathrm{MMD}(P_1, P_2) := ||\boldsymbol{\mu}_{P_1} - \boldsymbol{\mu}_{P_2}||_{\mathcal{H}_{\mathcal{Y}}}.$$

It also has the following equivalent representation (Deb et al., 2020; Huang et al., 2020):

$$\mathrm{MMD}^2(P_1, P_2) := \mathbb{E}\left[\mathcal{K}(U, U')\right] + \mathbb{E}\left[\mathcal{K}(V, V')\right] - 2\mathbb{E}\left[\mathcal{K}(U, V)\right],$$

where $U, U' \overset{\text{i.i.d.}}{\sim} P_1$, $V, V' \overset{\text{i.i.d.}}{\sim} P_2$ *and* $U \perp\!\!\!\perp V$.

Now we are ready to define the AKMMD.

Definition F.3 (average kernel maximum mean discrepancy)**.** *The* average kernel maximum mean discrepancy *for variable* X *is defined as*

$$\mathcal{I}_{\mathcal{K}}^2 = \mathbb{E}\left[\mathrm{MMD}^2(P_{Y|X,Z}, P_{Y|Z})\right] \tag{F.1}$$

whenever all the above expectations exist.

We also present its alternative expression in terms of the kernel:

$$\mathcal{I}_{\mathcal{K}}^2 = \mathbb{E}\left[\mathcal{K}(Y_2, \tilde{Y}_2)\right] - \mathbb{E}\left[\mathcal{K}(Y_1, \tilde{Y}_1)\right] = \mathbb{E}\left[\mathbb{E}\left[\mathcal{K}(Y_2, \tilde{Y}_2) \mid X, Z\right]\right] - \mathbb{E}\left[\mathbb{E}\left[\mathcal{K}(Y_1, \tilde{Y}_1) \mid Z\right]\right],$$

where $Y_1, \tilde{Y}_1, Y_2, \tilde{Y}_2$ are defined as below

$$\begin{aligned} Y_1 \mid X \sim P_{Y|Z}, \quad \tilde{Y}_1 \mid X \sim P_{Y|Z}, \quad \text{and } Y_1 \perp\!\!\!\perp \tilde{Y}_1 \mid X, \\ (X, Z) \sim P_{X,Z}, \quad Y_2 \mid X, Z \sim P_{Y|X,Z}, \quad \tilde{Y}_2 \mid X, Z \sim P_{Y|X,Z}, \quad \text{and } Y_2 \perp\!\!\!\perp \tilde{Y}_2 \mid X, Z. \end{aligned}$$

The floodgate functional constitutes a deterministic lower bound for the mMSE gap for any working regression function μ. As we are now dealing with mean embeddings with a general kernel, we will replace the role of μ with $Q_{Y|X,Z}$, an estimate of the full conditional distribution of $Y \mid X, Z$ (as opposed to just its conditional mean). Let $Q = Q_{Y|X,Z} \times P_{X,Z}$ and the associated conditional distribution of Y given Z by $Q_{Y|Z}$. For notational simplicity, $Q_{Y|X,Z}$ and $Q_{Y|Z}$ are both sometimes abbreviated simply as Q. Given any non-random conditional distribution $Q_{Y|X,Z}$, we consider the kernel floodgate functional

$$f_{\mathcal{K}}(Q) := \frac{\mathbb{E}\left[\mathcal{K}(Y, Y_2^Q)\right] - \mathbb{E}\left[\mathcal{K}(Y, Y_1^Q)\right]}{\sqrt{\mathbb{E}\left[\mathcal{K}(Y_2^Q, \tilde{Y}_2^Q)\right] - \mathbb{E}\left[\mathcal{K}(Y_2^Q, Y_1^Q)\right]}}, \tag{F.2}$$

where the involved random variables are defined through

$$\begin{aligned}
(X, Z) \sim P_{X,Z}, \quad Y \mid X, Z \sim P_{Y|X,Z}, \quad Y \mid Z \sim P_{Y|Z} \\
Y_2^Q, \tilde{Y}_2^Q \mid X, Z \overset{i.i.d.}{\sim} Q_{Y|X,Z}, \quad Y \perp\!\!\!\perp (Y_2^Q, \tilde{Y}_2^Q) \mid X, Z, \\
Y_1^Q \mid Z \overset{i.i.d.}{\sim} Q_{Y|Z}, \quad Y_1^Q \perp\!\!\!\perp (X, Y, Y_2^Q, \tilde{Y}_2^Q) \mid Z.
\end{aligned} \tag{F.3}$$

Lemma F.4 shows $f_{\mathcal{K}}$ tightly satisfies the lower-bounding property, as f does in Lemma 2.2. The proof can be found in Appendix F.3.

Lemma F.4. *For any Q such that $f_{\mathcal{K}}(Q)$ exists, we have $f_{\mathcal{K}}(Q) \leq \mathcal{I}_{\mathcal{K}}$, with equality when $Q = P_{Y|X,Z}$.*

Therefore, we can provide an LCB for $\mathcal{I}_{\mathcal{K}}$ via a LCB for $f_{\mathcal{K}}(Q)$ with some choice of Q. Since the definition of $f_{\mathcal{K}}(Q)$ involves null Y samples such as $Y_2^Q, \tilde{Y}_2^Q, Y_1^Q$, we will follow (F.3) to generate null samples of Y then construct i.i.d. unbiased estimates of the numerator and the denominator of $f_{\mathcal{K}}(Q)$ respectively. Based on the CLT and the delta method, we can derive asymptotically valid LCBs for $f_{\mathcal{K}}(Q)$. This idea is spelled out in Algorithm 2.

Algorithm 2 Kernel floodgate

Input: Data $\{(Y_i, W_i)\}_{i=1}^n$, a chosen kernel $\mathcal{K}(\cdot,\cdot)$, a estimated conditional distribution of $P_{Y|X,Z}$, denoted by $Q_{Y|X,Z}$, resampling number M, $P_{X|Z}$, number of null replicates K, and a confidence level $\alpha \in (0,1)$.

1: For each $i \in [n]$, draw $\{Y_{2,i}^{(m)}\}_{m=1}^M$ from $Q_{Y|X,Z}$ given (X_i, Z_i); given Z_i, draw i.i.d. null samples $\{\tilde{X}_i^{(k)}\}_{k=1}^K$ from $P_{X|Z}$, then draw $\{Y_{1,i}^{(k,m)}\}_{m=1}^M$ from $Q_{Y|X,Z}$ given $(X_i, \tilde{Z}_i^{(k)})$ for each $k \in [K]$. Denote $Y_{2,i}^{(m)} = Y_{1,i}^{(0,m)}$ for each $m \in [M]$.

2: Compute

$$\begin{aligned}
R_i &= \frac{1}{M}\sum_{m=1}^M \mathcal{K}\left(Y_i, Y_{2,i}^{(m)}\right) - \frac{1}{KM}\sum_{k=1}^K\sum_{m=1}^M \mathcal{K}\left(Y_i, Y_{1,i}^{(k,m)}\right) \\
V_i &= \frac{2}{(K+1)M(M-1)}\sum_{k=0}^K \sum_{1\leq m_1<m_2\leq M} \mathcal{K}\left(Y_{1,i}^{(k,m_1)}, Y_{1,i}^{(k,m_2)}\right) \\
&\quad - \frac{2}{K(K+1)M^2}\sum_{m_1,m_2=1}^M \sum_{0\leq k_1<k_2\leq K} \mathcal{K}\left(Y_{1,i}^{(k_1,m_1)}, Y_{1,i}^{(k_2,m_2)}\right)
\end{aligned}$$

for each $i \in [n]$, and their sample mean $(\bar{R}, \bar{V})$ and sample covariance matrix $\hat{\Sigma}$, and compute $s^2 = \frac{1}{\bar{V}}\left[\left(\frac{\bar{R}}{2\bar{V}}\right)^2 \hat{\Sigma}_{22} + \hat{\Sigma}_{11} - \frac{\bar{R}}{\bar{V}}\hat{\Sigma}_{12}\right]$.

Output: Lower confidence bound $L_n^\alpha(\mu) = \max\left\{\frac{\bar{R}}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n}}, 0\right\}$, with the convention that $0/0 = 0$.

F.3 Proofs in Appendix F.2

Proof of Lemma F.4. Recall the form of the kernel floodgate functional in (F.2)

$$f_{\mathcal{K}}(Q) = \frac{\mathbb{E}\left[\mathcal{K}(Y, Y_2^Q)\right] - \mathbb{E}\left[\mathcal{K}(Y, Y_1^Q)\right]}{\sqrt{\mathbb{E}\left[\mathcal{K}(Y_2^Q, \tilde{Y}_2^Q)\right] - \mathbb{E}\left[\mathcal{K}(Y_2^Q, Y_1^Q)\right]}} := \frac{\mathrm{II}_1}{\sqrt{\mathrm{II}_2}},$$

where $X, Z, Y, Y_2^Q, \tilde{Y}_2^Q, Y_1^Q$ are defined as

$$(X, Z) \sim P_{X,Z}, \quad Y \mid X, Z \sim P_{Y|X,Z}, \quad Y \mid Z \sim P_{Y|Z} \tag{F.4}$$

$$Y_2^Q, \tilde{Y}_2^Q \mid X, Z \overset{i.i.d.}{\sim} Q_{Y|X,Z}, \quad Y \perp\!\!\!\perp (Y_2^Q, \tilde{Y}_2^Q) \mid X, Z, \tag{F.5}$$

$$Y_1^Q \mid Z \overset{i.i.d.}{\sim} Q_{Y|Z}, \quad Y_1^Q \perp\!\!\!\perp (X, Y, Y_2^Q, \tilde{Y}_2^Q) \mid Z. \tag{F.6}$$

Denote the true conditional distributions $P_{Y|X,Z}, P_{Y|Z}$ by F, G respectively, the estimated conditional distributions $Q_{Y|X,Z}, Q_{Y|Z}$ by $F_\mathfrak{q}, G_\mathfrak{q}$ respectively, and the kernel mean embeddings of those conditional distributions by $\boldsymbol{\mu}_F, \boldsymbol{\mu}_G, \boldsymbol{\mu}_{F_\mathfrak{q}}, \boldsymbol{\mu}_{G_\mathfrak{q}}$. First notice

$$\langle \boldsymbol{\mu}_F, \boldsymbol{\mu}_{F_\mathfrak{q}} \rangle_{\mathcal{H}_\mathcal{Y}} = \left\langle \mathbb{E}\left[\mathcal{K}(\cdot, Y) \mid X, Z\right], \mathbb{E}\left[\mathcal{K}(\cdot, Y_2^Q) \mid X, Z\right] \right\rangle_{\mathcal{H}_\mathcal{Y}} = \mathbb{E}\left[\mathcal{K}(Y, Y_2^Q) \mid X, Z\right]$$

by (F.4), (F.5) and the definition of the kernel embedding. Similarly, we have the following equalities,

$$\mathbb{E}\left[\mathcal{K}(Y, Y_2^Q)\right] = \mathbb{E}\left[\mathbb{E}\left[\mathcal{K}(Y, Y_2^Q) \mid X, Z\right]\right] = \mathbb{E}\left[\langle \boldsymbol{\mu}_F, \boldsymbol{\mu}_{F_\mathfrak{q}} \rangle_{\mathcal{H}_\mathcal{Y}}\right], \tag{F.7}$$

$$\mathbb{E}\left[\mathcal{K}(Y, Y_1^Q)\right] = \mathbb{E}\left[\mathbb{E}\left[\mathcal{K}(Y, Y_1^Q) \mid Z\right]\right] = \mathbb{E}\left[\langle \boldsymbol{\mu}_G, \boldsymbol{\mu}_{G_\mathfrak{q}} \rangle_{\mathcal{H}_\mathcal{Y}}\right], \tag{F.8}$$

$$\mathbb{E}\left[\mathcal{K}(Y_2^Q, \tilde{Y}_2^Q)\right] = \mathbb{E}\left[\mathbb{E}\left[\mathcal{K}(Y_2^Q, \tilde{Y}_2^Q) \mid X, Z\right]\right] = \mathbb{E}\left[\langle \boldsymbol{\mu}_{F_\mathfrak{q}}, \boldsymbol{\mu}_{F_\mathfrak{q}} \rangle_{\mathcal{H}_\mathcal{Y}}\right], \tag{F.9}$$

$$\mathbb{E}\left[\mathcal{K}(Y_2^Q, Y_1^Q)\right] = \mathbb{E}\left[\mathbb{E}\left[\mathcal{K}(Y_2^Q, Y_1^Q) \mid X, Z\right]\right] = \mathbb{E}\left[\langle \boldsymbol{\mu}_{F_\mathfrak{q}}, \boldsymbol{\mu}_{G_\mathfrak{q}} \rangle_{\mathcal{H}_\mathcal{Y}}\right], \tag{F.10}$$

where we also apply the law of total expectation. Note that the subscripts for the expectation in the above equations are abbreviated. In addition to the these equalities, our derivation also relies on a key result $\mathbb{E}\left[\langle \boldsymbol{\mu}_G, \boldsymbol{\mu}_{F_\mathfrak{q}} \rangle\right] = \mathbb{E}\left[\langle \boldsymbol{\mu}_G, \boldsymbol{\mu}_{G_\mathfrak{q}} \rangle\right]$. Consider $\tilde{Y}$ satisfying $\tilde{Y} \mid X, Z \sim P_{Y|Z}$, $\tilde{Y} \perp\!\!\!\perp Y_2^Q \mid X, Z$, $\tilde{Y} \perp\!\!\!\perp Y_1^Q \mid Z$, then we prove the key result as below,

$$\begin{aligned} \mathbb{E}\left[\langle \boldsymbol{\mu}_G, \boldsymbol{\mu}_{F_\mathfrak{q}} \rangle_{\mathcal{H}_\mathcal{Y}}\right] &= \mathbb{E}_{X,Z}\left[\mathbb{E}\left[\mathcal{K}(\tilde{Y}, Y_2^Q) \mid X, Z\right]\right] \\ &= \mathbb{E}\left[\mathcal{K}(\tilde{Y}, Y_2^Q)\right] \\ &= \mathbb{E}_Z\left[\mathbb{E}\left[\mathcal{K}(\tilde{Y}, Y_2^Q) \mid Z\right]\right] \\ &= \mathbb{E}_Z\left[\mathbb{E}\left[\mathcal{K}(\tilde{Y}, Y_1^Q) \mid Z\right]\right] = \mathbb{E}\left[\langle \boldsymbol{\mu}_G, \boldsymbol{\mu}_{G_\mathfrak{q}} \rangle_{\mathcal{H}_\mathcal{Y}}\right], \end{aligned} \tag{F.11}$$

where the first and the last equalities hold by the definition of the kernel mean embedding, the second and the third equalities hold by the law of total expectation, and the fourth equality holds by the definitions of $Y_1^Q, Y_2^Q, \tilde{Y}$.

Therefore we can rewrite the numerator of $f_{\mathcal{K}}(Q)$ as

$$
\begin{aligned}
\mathrm{II}_1 &= \mathbb{E}\left[\mathcal{K}(Y, Y_2^Q)\right] - \mathbb{E}\left[\mathcal{K}(Y, Y_1^Q)\right] \\
&= \mathbb{E}\left[\langle \boldsymbol{\mu}_F, \boldsymbol{\mu}_{F_\mathrm{q}} \rangle_{\mathcal{H}_\mathcal{Y}}\right] - \mathbb{E}\left[\langle \boldsymbol{\mu}_F, \boldsymbol{\mu}_{G_\mathrm{q}} \rangle_{\mathcal{H}_\mathcal{Y}}\right] \\
&= \mathbb{E}\left[\langle \boldsymbol{\mu}_F, \boldsymbol{\mu}_{F_\mathrm{q}} - \boldsymbol{\mu}_{G_\mathrm{q}} \rangle_{\mathcal{H}_\mathcal{Y}}\right] \\
&= \mathbb{E}\left[\langle \boldsymbol{\mu}_F - \boldsymbol{\mu}_G, \boldsymbol{\mu}_{F_\mathrm{q}} - \boldsymbol{\mu}_{G_\mathrm{q}} \rangle_{\mathcal{H}_\mathcal{Y}}\right] + \mathbb{E}\left[\langle \boldsymbol{\mu}_G, \boldsymbol{\mu}_{F_\mathrm{q}} - \boldsymbol{\mu}_{G_\mathrm{q}} \rangle_{\mathcal{H}_\mathcal{Y}}\right] \\
&= \mathbb{E}\left[\langle \boldsymbol{\mu}_F - \boldsymbol{\mu}_G, \boldsymbol{\mu}_{F_\mathrm{q}} - \boldsymbol{\mu}_{G_\mathrm{q}} \rangle_{\mathcal{H}_\mathcal{Y}}\right] + \mathbb{E}\left[\langle \boldsymbol{\mu}_G, \boldsymbol{\mu}_{F_\mathrm{q}} \rangle_{\mathcal{H}_\mathcal{Y}}\right] - \mathbb{E}\left[\langle \boldsymbol{\mu}_G, \boldsymbol{\mu}_{G_\mathrm{q}} \rangle_{\mathcal{H}_\mathcal{Y}}\right] \\
&= \mathbb{E}\left[\langle \boldsymbol{\mu}_F - \boldsymbol{\mu}_G, \boldsymbol{\mu}_{F_\mathrm{q}} - \boldsymbol{\mu}_{G_\mathrm{q}} \rangle_{\mathcal{H}_\mathcal{Y}}\right] \\
&\le \mathbb{E}\left[||\boldsymbol{\mu}_F - \boldsymbol{\mu}_G||_{\mathcal{H}_\mathcal{Y}} ||\boldsymbol{\mu}_{F_\mathrm{q}} - \boldsymbol{\mu}_{G_\mathrm{q}}||_{\mathcal{H}_\mathcal{Y}}\right] \\
&\le \sqrt{\mathbb{E}\left[||\boldsymbol{\mu}_F - \boldsymbol{\mu}_G||^2_{\mathcal{H}_\mathcal{Y}}\right]} \sqrt{\mathbb{E}\left[||\boldsymbol{\mu}_{F_\mathrm{q}} - \boldsymbol{\mu}_{G_\mathrm{q}}||^2_{\mathcal{H}_\mathcal{Y}}\right]},
\end{aligned} \tag{F.12}
$$

where the first line holds due to (F.7) and (F.8), the second to the fourth equalities hold by rearranging, the fifth equality holds due to (F.11), the last two inequalities hold by the Cauchy–Schwarz inequality. Regarding the denominator of $f_{\mathcal{K}}(Q)$, we rewrite II_2 in terms of the kernel embedding

$$
\begin{aligned}
\mathrm{II}_2 &= \mathbb{E}\left[\mathcal{K}(Y_2^Q, \tilde{Y}_2^Q)\right] - \mathbb{E}\left[\mathcal{K}(Y_2^Q, Y_1^Q)\right] \\
&= \mathbb{E}\left[\langle \boldsymbol{\mu}_{F_\mathrm{q}}, \boldsymbol{\mu}_{F_\mathrm{q}} \rangle_{\mathcal{H}_\mathcal{Y}}\right] - \mathbb{E}\left[\langle \boldsymbol{\mu}_{G_\mathrm{q}}, \boldsymbol{\mu}_{F_\mathrm{q}} \rangle_{\mathcal{H}_\mathcal{Y}}\right] \\
&= \mathbb{E}\left[\langle \boldsymbol{\mu}_{F_\mathrm{q}}, \boldsymbol{\mu}_{F_\mathrm{q}} \rangle_{\mathcal{H}_\mathcal{Y}}\right] + \mathbb{E}\left[\langle \boldsymbol{\mu}_{G_\mathrm{q}}, \boldsymbol{\mu}_{G_\mathrm{q}} \rangle_{\mathcal{H}_\mathcal{Y}}\right] - 2\mathbb{E}\left[\langle \boldsymbol{\mu}_{G_\mathrm{q}}, \boldsymbol{\mu}_{F_\mathrm{q}} \rangle_{\mathcal{H}_\mathcal{Y}}\right] \\
&= \mathbb{E}\left[||\boldsymbol{\mu}_{F_\mathrm{q}} - \boldsymbol{\mu}_{G_\mathrm{q}}||^2_{\mathcal{H}_\mathcal{Y}}\right],
\end{aligned} \tag{F.13}
$$

where the second equality holds due to (F.9) and (F.10) and the third equality holds since $\mathbb{E}\left[\langle \boldsymbol{\mu}_{G_\mathrm{q}}, \boldsymbol{\mu}_{F_\mathrm{q}} \rangle_{\mathcal{H}_\mathcal{Y}}\right] = \mathbb{E}\left[\langle \boldsymbol{\mu}_{G_\mathrm{q}}, \boldsymbol{\mu}_{G_\mathrm{q}} \rangle_{\mathcal{H}_\mathcal{Y}}\right]$ can be similarly derived as (F.11). As $\mathcal{I}_{\mathcal{K}}^2 = \mathbb{E}\left[\mathrm{MMD}^2(P_{Y|X,Z}, P_{Y|Z})\right] = \mathbb{E}\left[||\boldsymbol{\mu}_F - \boldsymbol{\mu}_G||^2_{\mathcal{H}_\mathcal{Y}}\right]$, we have $f_{\mathcal{K}}(Q) \le \mathcal{I}_{\mathcal{K}}$ by combining (F.2), (F.12), and (F.13). □

G Transporting inference to other covariate distributions

To present how to perform inference on a target population whose covariate distribution differs from the distribution the study samples are drawn from, let Q denote the target distribution for all the random variables (Y, X, Z), but assume that $Q_{Y|X,Z} = P_{Y|X,Z}$ and that $Q_{X|Z}$ and the likelihood ratio Q_Z/P_Z are known (note this last requirement is trivially satisfied if only $X \mid Z$ changes between the study and target distributions, i.e., we know $Q_Z = P_Z$). Overloading notation slightly, let Q and P also denote the real-valued densities of random variables under their respective distributions (so, e.g., $P(Y = y \mid Z = z)$ denotes the density of $Y \mid Z = z$ under P evaluated at the value y), which we assume to exist. We can now define a weighted analogue of the floodgate functional (2.6):

$$
f^w(\mu) = \frac{\mathbb{E}_P[(Y - \mu(\tilde{X}, Z))^2 w(X, Z) w_1(\tilde{X}, Z) - (Y - \mu(X, Z))^2 w(X, Z)]}{\sqrt{2\mathbb{E}_P[(\mu(X, Z) - \mu(\tilde{X}, Z))^2 w(X, Z) w_1(\tilde{X}, Z)]}}, \tag{G.1}
$$

where $w(x, z) = w_0(z) w_1(x, z)$, $w_0(z) = \frac{Q(Z=z)}{P(Z=z)}$, $w_1(x, z) = \frac{Q(X=x \mid Z=z)}{P(X=x \mid Z=z)}$, and $\tilde{X} \sim P_{X|Z}$ conditionally independently of Y and X. The following Lemma certifies that f^w satisfies property (a) of a floodgate functional for $\mathcal{I}_Q^2 = \mathbb{E}_Q\left[\mathrm{Var}_Q\left(\mathbb{E}_Q[Y \mid X, Z] \mid Z\right)\right]$, the mMSE gap with respect to Q.

Lemma G.1. *If $Q_{Y|X,Z} = P_{Y|X,Z}$, then for any μ such that $f^w(\mu)$ exists, $f^w(\mu) \le \mathcal{I}_Q$, with equality when $\mu = \mu^\star$.*

The proof is immediate from Lemma 2.2 if we notice that the ratio of the joint distribution of $(Y, X, \tilde{X}, Z)$ under the two populations equals

$$\frac{Q(Y,X,Z)Q(\tilde{X} \mid Z)}{P(Y,X,Z)P(\tilde{X} \mid Z)} = \frac{Q(Y \mid X,Z)}{P(Y \mid X,Z)} \frac{Q(X,Z)}{P(X,Z)} \frac{Q(\tilde{X} \mid Z)}{P(\tilde{X} \mid Z)} = w_1(\tilde{X}, Z)w(X,Z), \tag{G.2}$$

where the last equality follows from $P_{Y|X,Z} = Q_{Y|X,Z}$. Floodgate property (b) of f^w can be established in the same way as for f by computing weighted versions of R_i and V_i from Algorithm 1 according to the weights in Equation (G.1), applying the central limit theorem, and combining them with the delta method.

H Algorithm details for inference on the MACM gap

Recall the construction of the floodgate functional ((3.2) in Section 3.1):

$$f_{\ell_1}(\mu) = 2\mathbb{P}\big(Y(\mu(\tilde{X}, Z) - \mathbb{E}[\mu(X,Z) \mid Z]) < 0\big) - 2\mathbb{P}\big(Y(\mu(X,Z) - \mathbb{E}[\mu(X,Z) \mid Z]) < 0\big).$$

We can define random variables which are i.i.d. and unbiased for $f_{\ell_1}(\mu)$ then construct CLT-based confidence bounds, as formalized in Algorithm 3. Algorithm 3 involves computing the terms $\mathbb{E}[\mu(X_i, Z_i) \mid Z_i]$

Algorithm 3 Floodgate for the MACM gap

Input: Data $\{(Y_i, X_i, Z_i)\}_{i=1}^n$, $P_{X|Z}$, a working regression function $\mu : \mathbb{R}^p \to \mathbb{R}$, and a confidence level $\alpha \in (0,1)$.

Let $U_i = \mu(X_i, Z_i) - \mathbb{E}[\mu(X_i, Z_i) \mid Z_i]$ and compute

$$R_i = \begin{cases} \mathbb{P}\left(U_i < 0 \mid Z_i\right) - \mathbb{1}_{\{U_i < 0\}} & \text{if } Y_i = 1 \\ \mathbb{P}\left(U_i > 0 \mid Z_i\right) - \mathbb{1}_{\{U_i > 0\}} & \text{if } Y_i = -1 \end{cases}$$

for $i \in [n]$, and compute its sample mean $\bar{R}$ and sample variance s^2.

return Lower confidence bound $L_n^\alpha(\mu) = 2\max\left\{\bar{R} - \frac{z_\alpha s}{\sqrt{n}}, 0\right\}$.

---

and evaluating the CDF of the conditional distribution $\mu(X,Z) \mid Z = z$ at the value $\mathbb{E}[\mu(X_i, Z_i) \mid Z_i]$, which is not analytically possible in general. Unlike in Section 2.4, where users can replace $\mathbb{E}\left[\mu(X,Z) \mid Z\right]$ and $\mathrm{Var}\left(\mu(X,Z) \mid Z\right)$ by their Monte Carlo estimators without it impacting asymptotic normality, we need slightly more assumptions when inferring the MACM gap due to the discontinuous indicator functions in the definition of $f_{\ell_1}(\mu)$. Before stating the required assumptions, we introduce some notation, all of which is specific to a given working regression function μ.

$$\begin{aligned}
&U := \mu(X,Z), \quad g(z) := \mathbb{E}[\mu(X,Z) \mid Z = z], \\
&G_z(u) := \mathbb{P}\left(U < u \mid Z = z\right), \quad F_z(u) := \mathbb{P}\left(U \le u \mid Z = z\right). \\
&\varsigma(z) := \sqrt{\mathrm{Var}\left(\mu(X,Z) \mid Z = z\right)}, \\
&C_{u,z,y} := \frac{\max\{\left|G_{z,y}(u) - G_{z,y}(g(z))\right|, \left|F_{z,y}(u) - F_{z,y}(g(z))\right|\}}{|u - g(z)|}
\end{aligned} \tag{H.1}$$

where $F_{z,y}(u)$ is the CDF of $\mu(X,Z) \mid Z = z, Y = y$ evaluated at u, $G_{z,y}(u)$ is the limit from the left of the same CDF at u, and with the convention for $C_{u,z,y}$ that $0/0 = 0$ (so it is well-defined when $u = g(z)$). Now we are ready to state Assumption H.1.

Assumption H.1. *Assume the joint distribution over (Y, X, Z) and the nonrandom function $\mu : \mathbb{R}^p \to \mathbb{R}$ satisfy the following on a set of values of $Y = y, Z = z$ of probability 1:*

(a) *There exists a* $\delta_{z,y} > 0$ *and finite* $C_{z,y}$ *such that*

$$C_{u,z,y} \leq C_{z,y} \quad \text{when } |u - g(z)| \leq \varsigma(z)\delta_{z,y}.$$

(b) *The above* $C_{z,y}$ *and* $\delta_{z,y}$ *satisfy*

$$\mathbb{E}\left[C_{Z,Y}^2\right] < \infty, \ \mathbb{E}\left[\frac{1}{\delta_{Z,Y}}\right] < \infty.$$

(c) $\mathbb{E}\left[\varsigma^2(Z)\right] < \infty, \ \mathbb{E}\left[\frac{\mathbb{E}\left[|\mu(X,Z)-\mathbb{E}[\mu(X,Z) \mid Z]|^3 \mid Z\right]}{\varsigma^3(Z)}\right] < \infty.$

These assumptions are placed because we have to construct the Monte Carlo estimator of $\mathbb{E}[\mu(X,Z) \mid Z]$ then plug it into the discontinuous indicator functions in $f_{\ell_1}(\mu)$. Assumptions H.1(a) and H.1(b) are smoothness requirements on the the CDF of $\mu(X,Z) \mid Z, Y$ around $\mathbb{E}[\mu(X,Z) \mid Z]$. Assumption H.1($c$) specifies mild moment bound conditions on $\mu(X,Z)$. To see that they are actually sensible, we consider the example of logistic regression and walk through those assumptions in Appendix H.1.

Assume that we can sample $(M+K)$ copies of X_i from $P_{X_i|Z_i}$ conditionally independently of X_i and Y_i, which are denoted by $\{\tilde{X}_i^{(m)}\}_{m=1}^M$, $\{\tilde{X}_i^{(k)}\}_{k=1}^K$, and thus replace $g(Z_i)$ (i.e. $\mathbb{E}[\mu(X_i,Z_i) \mid Z_i]$) and R_i, respectively, by the sample estimators

$$g^M(Z_i) = \frac{1}{M}\sum_{m=1}^M \mu(\tilde{X}_i^{(m)}, Z_i), \ R_i^{M,K} = \frac{1}{K}\sum_{k=1}^K \left(\mathbb{1}_{\left\{Y_i(\mu(\tilde{X}_i^{(k)},Z_i)-g^M(Z_i))<0\right\}}\right) - \mathbb{1}_{\{Y_i(\mu(X_i,Z_i))-g^M(Z_i))<0\}}$$

Theorem H.2. *Under the same setting as in Theorem 3.3, if either (i)* $\mathbb{E}[\mathrm{Var}(\mu(X,Z) \mid Z)] = 0$ *or (ii)* $\mathbb{E}\left[\mathrm{Var}\left(\mathbb{1}_{\{Y\cdot[\mu(X,Z)-\mathbb{E}[\mu(X,Z) \mid Z]]<0\}} \mid Z, Y\right)\right] > 0$ *holds together with Assumption H.1 and* $n/M = o(1)$*, then* $L_{n,M,K}^{\alpha}(\mu)$ *computed by replacing* $g(Z_i)$ *and* R_i *with* $g^M(Z_i)$ *and* $R_i^{M,K}$*, respectively, in Algorithm 3 satisfies*

$$\mathbb{P}\left(L_{n,M,K}^{\alpha}(\mu) \leq \mathcal{I}_{\ell_1}\right) \geq 1 - \alpha + o(1).$$

The proof can be found in Appendix H.2. Intuitively when we construct a lot more null samples to estimate the term $g(Z_i)$, our inferential validity improves. Formally, when $n^2/M = O(1)$, we can improve the asymptotic miscoverage to $O(n^{-1/2})$. Note that we only place a rate assumption on M (but put no requirement on K).

H.1 Illustration of assumption H.1

We consider the joint distribution over W to be p-dimensional multivariate Gaussian with $X = W_j, Z = W_{-j}$ for some $1 \leq j \leq p$, and Y follows a generalized linear model with logistic link. That is,

$$W \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \quad \mu^\star(W) = 2\mathbb{P}(Y=1 \mid W) - 1, \quad \text{where } \mathbb{P}(Y=1 \mid W) = \frac{\exp(W\beta^\star)}{1+\exp(W\beta^\star)}, \ \beta^\star \in \mathbb{R}^p.$$

Choosing logistic regression as the fitting algorithm, we have $U := \mu(X,Z)$ takes the following form

$$U := \mu(W) = \frac{2\exp(W\beta)}{1+\exp(W\beta)} - 1$$

where $\beta \in \mathbb{R}^p$ is the fitted regression coefficient vector and $\beta_j \neq 0$ whenever $\mathbb{E}[\mathrm{Var}(\mu(X,Z) \mid Z)] > 0$. Conditional on Z, U follows a logit-normal distribution (defined as the logistic function transformation of normal random variable) up to constant shift and scaling. Note that the probability density function (PDF) of logit-normal distribution with parameters a, σ is

$$h_{\text{logit}}(u) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(\text{logit}(u)-a)^2}{2\sigma^2}\right)\frac{1}{u(1-u)}, \quad u \in (0,1) \tag{H.2}$$

where $\text{logit}(u) = \log(u/(1-u))$ is the logit function. Note $h_{\text{logit}}(u)$ is bounded over its support. Regarding the PDF of $U \mid Z = z, Y = 1$, which is denoted as $h_{z,1}(u)$, we first notice the following expression

$$h(x \mid Z = z, Y = 1) = \frac{h(x \mid Z = z)\mathbb{P}\left(Y = 1 \,|\, W = w\right)}{\int h(x \mid Z = z)\mathbb{P}\left(Y = 1 \,|\, W = w\right) dx} \tag{H.3}$$

where $w_j = x, w_{\text{-}j} = z$, $h(x \mid Z = z, Y = 1)$ and $h(x \mid Z = z, Y = 1)$ denote the density functions of $X \mid Z = z, Y = 1$ and $X \mid Z = z$. Since $\text{logit}(z)$ is one-to-one mapping, we have $f_{z,1}(z)$ (up to constant shift and scaling) takes the form similar to (H.3)

$$h_{z,1}(u) = \frac{h_{\text{logit}}(u)\mathbb{P}\left(Y = 1 \,|\, W = w\right)}{\int h_{\text{logit}}(u)\mathbb{P}\left(Y = 1 \,|\, W = w\right) dx} \tag{H.4}$$

where $w = (x, z) = \mu^{-1}(u)$, and we denote the PDF of $U \mid Z = z$ as $h_{\text{logit}}(u)$ without causing confusion (the parameters of $h_{\text{logit}}(u)$ depend on z, β). Therefore we can show $h_{z,1}(z)$ is bounded (similarly for $h_{z,-1}(z)$).

The boundedness of $h_{z,y}(u)$ implies that the corresponding CDF $F_{z,y}$ ($F_{z,y} = G_{z,y}$ in this case) satisfies a Lipschitz condition over its support. Hence $\delta_{z,y}$ can be chosen to be greater than some positive constant uniformly, so that $\mathbb{E}\left[\frac{1}{\delta_{Z,Y}}\right] < \infty$ holds. Though the Lipschitz constant does depend on z, β, it is easy to verify $\mathbb{E}\left[C_{Z,Y}^2\right] < \infty$, thus assumption (b) holds. And assumption (c) is just a regular moment condition.

H.2 Proofs in Appendix H

Proof of Theorem H.2. Similar to the proof of Theorem 3.3, it suffices to deal with the case where $\mu(X, Z) \notin \mathscr{A}(Z)$ and prove

$$\mathbb{P}\left(L_{n,M,K}^{\alpha}(\mu) \leq f_{\ell_1}(\mu)\right) \geq 1 - \alpha + o(1). \tag{H.5}$$

Note that in Algorithm 3, $\mathbb{E}\left[R_i\right] = f_{\ell_1}(\mu)/2$. But when $g(Z_i)$ (i.e., $\mathbb{E}[\mu(X_i, Z_i) \,|\, Z_i]$) and R_i are replaced by $g^M(Z_i)$ and $R_i^{M,K}$, respectively, in Algorithm 3, we do not have $\mathbb{E}\left[R_i^{M,K}\right]$ equal to $f_{\ell_1}(\mu)/2$ anymore. Note that $f_{\ell_1}(\mu)/2$ equals the following

$$f_{\ell_1}(\mu)/2 = \mathbb{E}\left[\mathbb{1}_{\{Y \cdot [\mu(\tilde{X}, Z) - \mathbb{E}[\mu(X,Z) \,|\, Z]] < 0\}}\right] - \mathbb{E}\left[\mathbb{1}_{\{Y \cdot [\mu(X,Z) - \mathbb{E}[\mu(X,Z) \,|\, Z]] < 0\}}\right], \tag{H.6}$$

and $R_i^{M,K}$ is defined as

$$R_i^{M,K} = \frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{1}_{\left\{Y_i(\mu(\tilde{X}_i^{(k)}, Z_i) - g^M(Z_i)) < 0\right\}}\right) - \mathbb{1}_{\{Y_i(\mu(X_i, Z_i)) - g^M(Z_i)) < 0\}} \tag{H.7}$$

Remark the value of $\mathbb{E}\left[R_i^{M,K}\right]$ does not depend on K, hence we simplify the notation into R_i^M without causing confusion. Actually we can show as $M \to \infty$, $\mathbb{E}\left[R_i^M\right] \to f_{\ell_1}(\mu)/2$. Indeed, we need to show $\sqrt{n}|\mathbb{E}\left[R_i^M\right] - f_{\ell_1}(\mu)/2| = o(1)$ in order to prove (H.5). Also remark that in Section 3.1, it is mentioned that under a stronger condition $n^2/M = O(1)$ (which will imply $\sqrt{n}|\mathbb{E}\left[R_i^M\right] - f_{\ell_1}(\mu)/2| = O(1/\sqrt{n})$), we can additionally establish a rate for $n^{-1/2}$ for the asymptotic coverage validity in Theorem H.2. In either cases, it is reduced to prove

$$\left|\mathbb{E}\left[R_i^M\right] - \frac{f_{\ell_1}(\mu)}{2}\right| = O\left(\frac{1}{\sqrt{M}}\right) \tag{H.8}$$

First we ignore the i subscripts and get rid of the average over K null samples in the definition of $R_i^{M,K}$, then $\mathbb{E}\left[R_i^M\right]$ can be simplified into

$$\mathbb{E}\left[\mathbb{1}_{\left\{Y(\mu(\tilde{X}, Z) - g^M(Z)) < 0\right\}} - \mathbb{1}_{\{Y(\mu(X,Z) - g^M(Z)) < 0\}}\right] \tag{H.9}$$

where $g^M(Z) = \frac{1}{M}\sum_{m=1}^M \mu(\tilde{X}^{(m)}, Z)$. To bound $\left|\mathbb{E}\left[R_i^M\right] - f_{\ell_1}(\mu)/2\right|$, we consider the two terms in (H.6) and separately bound

$$\begin{aligned}
\mathrm{II}_1 &:= \left|\mathbb{E}\left[\mathbb{1}_{\left\{Y(\mu(\tilde{X},Z)-g^M(Z))<0\right\}} - \mathbb{1}_{\{Y\cdot[\mu(\tilde{X},Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]]<0\}}\right]\right|, \\
\mathrm{II}_2 &:= \left|\mathbb{E}\left[\mathbb{1}_{\{Y(\mu(X,Z)-g^M(Z)<0\}} - \mathbb{1}_{\{Y\cdot[\mu(X,Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]]<0\}}\right]\right|.
\end{aligned}$$

Starting from the second term above, we rewrite it as

$$\begin{aligned}
\mathrm{II}_2 = & \left|\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{Y(\mu(X,Z)-g^M(Z))<0\}} - \mathbb{1}_{\{Y\cdot[\mu(X,Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]]<0\}}\,|\,Z, Y, \{\tilde{X}^{(m)}\}_{m=1}^M\right]\right]\right| \\
\leq & \left|\mathbb{E}\left[\mathbb{1}_{\{Y=1\}}\mathbb{E}\left[\mathbb{1}_{\{\mu(X,Z)<g^M(Z)\}} - \mathbb{1}_{\{\mu(X,Z)<\mathbb{E}[\mu(X,Z)\,|\,Z]\}}\,|\,Z, Y, \{\tilde{X}^{(m)}\}_{m=1}^M\right]\right]\right| \\
& + \left|\mathbb{E}\left[\mathbb{1}_{\{Y=-1\}}\mathbb{E}\left[\mathbb{1}_{\{\mu(X,Z)>g^M(Z)\}} - \mathbb{1}_{\{\mu(X,Z)>\mathbb{E}[\mu(X,Z)\,|\,Z]\}}\,|\,Z, Y, \{\tilde{X}^{(m)}\}_{m=1}^M\right]\right]\right| \\
\leq & \ \mathbb{E}\left[\max\{\left|G_{Z,Y}(g^M(Z)) - G_{Z,Y}(g(Z))\right|, \left|F_{Z,Y}(g^M(Z)) - F_{Z,Y}(g(Z))\right|\}\right] \\
:= & \ \mathbb{E}\left[A\right] \qquad \text{(H.10)}
\end{aligned}$$

where the first equality is by the law of total expectation, the first and the second inequality are simply expanding and rearranging. By construction, $\mu(\tilde{X}^{(m)}, Z), m \in [M]$ are i.i.d. random variables conditioning on Z, Y, then by central limit theorem we have

$$\frac{\sqrt{M}(g^M(Z) - g(Z))}{\varsigma(Z)} \xrightarrow{d} \mathcal{N}(0, 1)$$

conditioning on Z, Y. Further we obtain the following from the Berry–Esseen bound i.e. Lemma C.3:

$$\left|\mathbb{P}\left(\left|\frac{\sqrt{M}|g^M(Z) - g(Z)|}{\varsigma(Z)}\right| > \sqrt{M}\delta_{Z,Y} \,\middle|\, Z, Y\right) - \bar{\Phi}(\left|\sqrt{M}\delta_{Z,Y}\right|)\right| \leq \frac{C}{\sqrt{M}} \cdot \frac{\mathbb{E}\left[|\mu^3(X,Z)|\,|\,Z\right]}{\varsigma^3(Z)} \qquad \text{(H.11)}$$

for any $\delta_{Z,Y}$ when conditioning on Z, Y, where $\bar{\Phi}(x) = 1 - \Phi(x)$ and C is some constant which does not depend on the distribution of (Y, X, Z). Regarding (H.10), by considering the event $B := \{|g^M(Z) - g(Z)|/\varsigma(Z) \leq \delta_{Z,Y}\}$, we can decompose (H.10) into

$$\mathbb{E}\left[A\right] = \mathbb{E}\left[A\mathbb{1}_{\{B\}}\right] + \mathbb{E}\left[A\mathbb{1}_{\{B^c\}}\right] \qquad \text{(H.12)}$$

For the first term, we have

$$\begin{aligned}
\mathbb{E}\left[A\mathbb{1}_{\{B\}}\right] &\leq \mathbb{E}\left[C_{g^M(Z),Z,Y}\left|g^M(Z) - g(Z)\right|\mathbb{1}_{\{B\}}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[C_{g^M(Z),Z,Y}\left|g^M(Z) - g(Z)\right|\mathbb{1}_{\{B\}} \,\middle|\, Z, Y\right]\right] \\
&\leq \mathbb{E}\left[C_{Z,Y}\mathbb{E}\left[\left|g^M(Z) - g(Z)\right|\,|\,Z, Y\right]\right] \\
&\leq \mathbb{E}\left[C_{Z,Y}\sqrt{\mathbb{E}\left[\left|g^M(Z) - g(Z)\right|^2 \,\middle|\, Z, Y\right]}\right] \qquad \text{(H.13)}
\end{aligned}$$

where the first inequality is by the definition of $C_{u,z,y}$, the first equality is from the law of total expectation, the second inequality holds by (a) in Assumption H.1 and the last inequality holds due to the Cauchy–Schwarz inequality. Remember we have $g^M(Z) = \frac{1}{M}\sum_{m=1}^M \mu(\tilde{X}^{(m)}, Z)$ where $\mu(\tilde{X}^{(m)}, Z), m \in [M]$ are i.i.d. random variables with mean $g(Z)$ when conditioning on Z, Y, hence (H.13) equals

$$\mathbb{E}\left[C_{Z,Y}\sqrt{\frac{\varsigma^2(Z)}{M}}\right] \leq \frac{1}{\sqrt{M}}\sqrt{\mathbb{E}\left[C_{Z,Y}^2\right]}\sqrt{\mathbb{E}\left[\varsigma^2(Z)\right]} = O\left(\frac{1}{\sqrt{M}}\right)$$

where the first inequality is from the Cauchy–Schwarz inequality and the second one holds by (b) and (c) in Assumption H.1. Now we have showed

$$\mathbb{E}\left[A\mathbb{1}_{\{B\}}\right] = O\left(\frac{1}{\sqrt{M}}\right), \tag{H.14}$$

it suffices to prove the same rate for $\mathbb{E}\left[A\mathbb{1}_{\{B^c\}}\right]$:

$$\begin{aligned}
\mathbb{E}\left[A\mathbb{1}_{\{B^c\}}\right] &\leq 2\,\mathbb{P}\left(B^c\right) \\
&= 2\,\mathbb{E}\left[\mathbb{P}\left(B^c \,|\, Z\right)\right] \\
&= 2\,\mathbb{E}\left[\mathbb{P}\left(\sqrt{M}|g^M(Z) - g(Z)|/\varsigma(Z) > \sqrt{M}\delta_{Z,Y} \,|\, Z\right)\right] \\
&\leq 2\mathbb{E}\left[\bar{\Phi}(\left|\sqrt{M}\delta_{Z,Y}\right|) + \frac{C}{\sqrt{M}} \cdot \frac{\mathbb{E}\left[|\mu^3(X,Z)|\,|\,Z\right]}{\varsigma^3(Z)}\right] \\
&\leq 2\mathbb{E}\left[\frac{2}{\sqrt{2\pi}}\frac{\exp\{-M\delta_{Z,Y}^2\}}{\sqrt{M}\delta_{Z,Y}} + \frac{C}{\sqrt{M}} \cdot \frac{\mathbb{E}\left[|\mu^3(X,Z)|\,|\,Z\right]}{\varsigma^3(Z)}\right]
\end{aligned}$$

where the first inequality holds since $F_{z,y}(u), G_{z,y}(u)$ are bounded between 0 and 1, the first equality is due to the law of total expectation, the second equality is from the definition of the event B, the second inequality holds due to (H.11) and the last inequality is a result of Mill's Ratio, see Proposition 2.1.2 in Vershynin (2018). Under (b) and (c) in Assumption H.1, the following holds

$$\mathbb{E}\left[A\mathbb{1}_{\{B^c\}}\right] = O\left(\frac{1}{\sqrt{M}}\right). \tag{H.15}$$

Finally we prove

$$\left|\mathbb{E}\left[\mathbb{1}_{\{Y(\mu(X,Z)-g^M(Z))<0\}} - \mathbb{1}_{\{Y\cdot[\mu(X,Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]]<0\}}\right]\right| = O\left(\frac{1}{\sqrt{M}}\right).$$

Regarding the term

$$\mathrm{II}_1 = \left|\mathbb{E}\left[\mathbb{1}_{\{Y(\mu(\tilde{X},Z)-g^M(Z))<0\}} - \mathbb{1}_{\{Y\cdot[\mu(\tilde{X},Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]]<0\}}\right]\right|$$

All of the steps are the same except that the CDF (and its limit) of the conditional distribution $X \mid Z, Y$ are replaced by those of $X \mid Z$, i.e. $F_z(u)$ and $G_z(u)$ as defined in (H.1). Hence it suffices to notice the following derivations for $F_z(u)$:

$$\begin{aligned}
F_z(u) = \mathbb{P}\left(U \leq u \,|\, Z = z\right) &= \mathbb{E}_{Y|Z=z}\left[\mathbb{P}\left(U \leq u \,|\, Z = z, Y\right) | \, Z = z\right] \\
&= \mathbb{E}_{Y|Z=z}\left[F_{z,Y}(u) \,|\, Z = z\right],
\end{aligned}$$

and similarly for $G_z(u)$. Together with the definition of $C_{u,z,y}$ and (a) in Assumption H.1, the above equations yield

$$\max\{|F_z(u) - F_z(g(z))|\,, |G_z(u) - G_z(g(z))|\} \leq C_{z,y}|u - g(z)|$$

over the region $|u - g(z)| \leq \varsigma(z)\delta_{z,y}$. Then the other steps follow as those of proving the term II_2. Finally, we obtain a rate of $O\left(\frac{1}{\sqrt{M}}\right)$ for $\left|\mathbb{E}\left[R_i^M\right] - f_{\ell_1}(\mu)/2.\right|$.

In the following, we prove the stronger version of (H.5), i.e.,

$$\mathbb{P}\left(L^{\alpha}_{n,M,K}(\mu) \leq f_{\ell_1}(\mu)\right) \geq 1 - \alpha - O\left(\frac{1}{\sqrt{n}}\right), \tag{H.16}$$

when assuming $n^2/M = O(1)$. For this it suffices to establish the following Berry–Esseen bound:

$$\Delta := \sup_{t\in\mathbb{R}} \left| \mathbb{P}\left(\sqrt{n}\left(\frac{\bar{R} - f_{\ell_1}(\mu)/2}{s} \right) \le t \right) - \Phi(t) \right| = O\left(\frac{1}{\sqrt{n}} \right),$$

where $\bar{R}$ and s are defined similarly as in Algorithm 3 except that $g(Z_i)$ and R_i are replaced with $g^M(Z_i)$ and $R_i^{M,K}$, respectively. Notice that

$$\begin{aligned} \Delta &= \sup_{t\in\mathbb{R}} \left| \mathbb{P}\left(\sqrt{n}\left(\frac{\bar{R} - \mathbb{E}\left[R_i^M\right]}{s} \right) \le t + \sqrt{n}\frac{(\mathbb{E}\left[R_i^M\right] - f_{\ell_1}(\mu)/2)}{s} \right) - \Phi(t) \right| \\ &\le \sup_{t\in\mathbb{R}} \left| \mathbb{P}\left(\sqrt{n}\left(\frac{\bar{R} - \mathbb{E}\left[R_i^M\right]}{s} \right) \le t \right) - \Phi(t) \right| + \sup_{t\in\mathbb{R}} \left| \Phi\left(t + \sqrt{n}\frac{(\mathbb{E}\left[R_i^M\right] - f_{\ell_1}(\mu)/2)}{s} \right) - \Phi(t) \right| \\ &:= \Delta_1 + \Delta_2 \end{aligned}$$

Since the first derivative of $\Phi(t)$ is bounded by $1/\sqrt{2\pi}$ over $\mathbb{R}$, we have

$$\Delta_2 \le \frac{\sqrt{n}}{\sqrt{2\pi}} \frac{|f_{\ell_1}(\mu)/2 - \mathbb{E}\left[R_i^M\right]|}{\sqrt{\mathrm{Var}\left(R_i^M\right)}} \cdot (\sqrt{\mathrm{Var}\left(R_i^M\right)}/s)$$

by Taylor expansion. Note that as a result of (H.8), we have

$$\sqrt{n}|\mathbb{E}\left[R_i^M\right] - f_{\ell_1}(\mu)/2| = O(1/\sqrt{n}). \tag{H.17}$$

Then it suffices to prove $\Delta_1 = O(1/\sqrt{n})$ and $\mathrm{Var}\left(R_i^M\right) > 0$ (since s is simply the sample mean estimator of $\mathrm{Var}\left(R_i^M\right)$ thus consistent). $\Delta_1 = O(1/\sqrt{n})$ holds when applying the triangular array version of the Berry–Esseen bound in Lemma C.4 (note that the result is stated in a way such that the bound clearly applies to the triangular array with i.i.d. rows $\{R_i^{M,K}\}_{i=1}^n$ for each M). The only thing we need to deal with is to verify the following uniform moment conditions:

(i) $\sup_{M,K} \mathbb{E}\left[\left| R_i^{M,K} - \mathbb{E}\left[R_i^{M,K}\right] \right|^3 \right] < \infty$,

(ii) $\inf_{M,K} \mathrm{Var}\left(R_i^{M,K}\right) > 0$.

where we go back to the original notation $R_i^{M,K}$ from the simplified one R_i^M since the above moments do depend on both M and K. Since $R_i^{M,K}$ is always bounded, (i) holds. Regarding (ii), notice that we have the following

$$\begin{aligned} &\mathrm{Var}\left(R_i^{M,K}\right) \\ &= \mathbb{E}\left[\mathrm{Var}\left(R_i^{M,K} \mid Z_i, Y_i, \{\tilde{X}_i^{(m)}\}_{m=1}^M \right) \right] + \mathrm{Var}\left(\mathbb{E}\left[R_i^{M,K} \mid Z_i, Y_i, \{\tilde{X}_i^{(m)}\}_{m=1}^M \right] \right) \\ &\ge \mathbb{E}\left[\mathrm{Var}\left(R_i^{M,K} \mid Z_i, Y_i, \{\tilde{X}_i^{(m)}\}_{m=1}^M \right) \right] \\ &= \mathbb{E}\left[\mathrm{Var}\left(\frac{1}{K}\sum_{k=1}^K \left(\mathbb{1}_{\left\{Y_i(\mu(\tilde{X}_i^{(k)}, Z_i) - g^M(Z_i)) < 0\right\}} \right) - \mathbb{1}_{\{Y_i(\mu(X_i, Z_i)) - g^M(Z_i)) < 0\}} \,\middle|\, Z_i, Y_i, \{\tilde{X}_i^{(m)}\}_{m=1}^M \right) \right] \\ &\ge \mathbb{E}\left[\mathrm{Var}\left(\mathbb{1}_{\{Y_i(\mu(X_i, Z_i)) - g^M(Z_i)) < 0\}} \,\middle|\, Z_i, Y_i, \{\tilde{X}_i^{(m)}\}_{m=1}^M \right) \right] := \sigma_M^2 \end{aligned} \tag{H.18}$$

where the first equality is due to the law of total expectation, the second equality is by the definition of $R_i^{M,K}$, the second inequality holds since $\{\tilde{X}_i^{(k)}\}_{k=1}^K \perp\!\!\!\perp X_i \mid Z_i, Y_i, \{\tilde{X}_i^{(m)}\}_{m=1}^M$ due to the construction of

$\{\tilde{X}_i^{(k)}\}_{k=1}^K$ and the variance of first term is non-negative. Before dealing with (H.18), notice the stated condition

$$\sigma_0^2 := \mathbb{E}\left[\mathrm{Var}\left(\mathbb{1}_{\{Y_i(\mu(X_i,Z_i))-g(Z_i))<0\}} \,\middle|\, Z_i, Y_i\right)\right] > 0$$

Thus to establish (ii), it suffices to show $\sigma_M^2 \to \sigma_0^2$ as $M \to \infty$. Recall the derivations in (H.10) for bounding the term II_2, we can similarly bound $|\sigma_M^2 - \sigma_0^2|$ by the following quantity:

$$\begin{aligned} |\sigma_M^2 - \sigma_0^2| &\le \mathbb{E}\left[3\max\{\left|G_{Z,Y}(g^M(Z)) - G_{Z,Y}(g(Z))\right|, \left|F_{Z,Y}(g^M(Z)) - F_{Z,Y}(g(Z))\right|\}\right] \\ &= 3\mathbb{E}\left[A\right] = 3(\mathbb{E}\left[A\mathbb{1}_{\{B\}}\right] + \mathbb{E}\left[A\mathbb{1}_{\{B^c\}}\right]) = O\left(\frac{1}{\sqrt{M}}\right). \end{aligned}$$

where the last equality holds due to the results (H.14) and (H.15) from previous derivations for the term II_2. Finally we conclude (H.16), which immediately implies a weaker version of the result, i.e.the statement of Theorem H.2. □

I Co-sufficient floodgate details

The strategy described in Section 3.2 is formalized in Algorithm 4 (under the simplifying assumption that the number of batches, n_2, evenly divides the sample size n).

Algorithm 4 Co-sufficient floodgate

Input: The inputs of Algorithm 1, a sufficient statistic functional $\mathcal{T}$, and a batch size n_2.

1: Let $n_1 = n/n_2$ and for $m \in [n_1]$, denote $(\boldsymbol{X}_m, \boldsymbol{Z}_m) = \{X_i, Z_i\}_{i=(m-1)n_2+1}^{mn_2}$, and let $\boldsymbol{T}_m = \mathcal{T}(\boldsymbol{X}_m, \boldsymbol{Z}_m)$.

2: For $m \in [n_1]$, compute

$$R_m = \frac{1}{n_2}\sum_{i=(m-1)n_2+1}^{mn_2} Y_i\left(\mu(X_i,Z_i) - \mathbb{E}\left[\mu(X_i,Z_i)\,|\,\boldsymbol{Z}_m, \boldsymbol{T}_m\right]\right),$$

$$V_m = \frac{1}{n_2}\sum_{i=(m-1)n_2+1}^{mn_2} \mathrm{Var}\left(\mu(X_i,Z_i)\,|\,\boldsymbol{Z}_m, \boldsymbol{T}_m\right),$$

their sample mean $(\bar{R}, \bar{V})$, their sample covariance matrix $\hat{\Sigma}$, and $s^2 = \frac{1}{\bar{V}}\left[\left(\frac{\bar{R}}{2\bar{V}}\right)^2\hat{\Sigma}_{22} + \hat{\Sigma}_{11} - \frac{\bar{R}}{\bar{V}}\hat{\Sigma}_{12}\right]$.

3: **return** Lower confidence bound $L_n^{\alpha,\mathcal{T}}(\mu) = \max\left\{\frac{\bar{R}}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n_1}}, 0\right\}$, with the convention that $0/0 = 0$.

I.1 Monte Carlo analogue of co-sufficient floodgate

Similarly as in Section 2, when the conditional expectations in Algorithm 4 do not have closed-form expressions, Monte Carlo provides a general approach: within each batch, we can sample K copies $\tilde{\boldsymbol{X}}_m^{(k)}$ of $\boldsymbol{X}_m$ from the conditional distribution $\boldsymbol{X}_m\,|\,\boldsymbol{Z}_m, \boldsymbol{T}_m$, conditionally independently of $\boldsymbol{X}_m$ and $\boldsymbol{y}$ and thus replace R_m and V_m, respectively, by the sample estimators

$$\begin{aligned} (R_m^K, V_m^K) = \frac{1}{n_2}\Bigg(&\sum_{i=(m-1)n_2+1}^{mn_2} Y_i\left(\mu(X_i,Z_i) - \frac{1}{K}\sum_{k=1}^K \mu(\tilde{X}_i^{(k)}, Z_i)\right), \\ &\sum_{i=(m-1)n_2+1}^{mn_2} \frac{1}{K-1}\sum_{k=1}^K\left(\mu(X_i^{(k)}, Z_i) - \frac{1}{K}\sum_{k=1}^K \mu(\tilde{X}_i^{(k)}, Z_i)\right)^2\Bigg) \end{aligned}$$

We defer to future work a proof of validity of the Monte Carlo analogue of co-sufficient floodgate following similar techniques as Theorem 2.5.

I.2 Proofs in Appendix I

Lemma I.1. *Under the moment conditions* $\mathbb{E}\left[\mu^2(X,Z)\right], \mathbb{E}\left[(\mu^\star)^2(X,Z)\right] < \infty$, *we can quantify the gap between* $f(\mu)$ *and* $f_n^{\mathcal{T}}(\mu)$ *as below.*

$$f(\mu) - f_n^{\mathcal{T}}(\mu) = O\left(\max\{\mathrm{II}(\mu), \mathrm{II}(\mu^\star)\}\right) \tag{I.1}$$

where $\mathrm{II}(\mu) = \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_{\boldsymbol{T}|\boldsymbol{Z}}\left(\mathbb{E}\left[\mu(X_i,Z_i) \mid \boldsymbol{Z},\boldsymbol{T}\right]\right)\right]$.

When this lemma is used in the proof of Proposition 3.5 and 3.6, the natural sufficient statistic and $f_n^{\mathcal{T}}(\mu)$ are actually defined based on the batch $\mathcal{B}_m$ whose sample size is n_2. We do not carry these in the above notation, but use generic $(\boldsymbol{X},\boldsymbol{Z})$ instead, where $(\boldsymbol{X},\boldsymbol{Z}) = \{(X_i,Z_i)\}_{i=1}^n$.

Proof of Lemma I.1. Recall the definition of $f(\mu)$ and $f_n^{\mathcal{T}}(\mu)$,

$$\begin{aligned} f(\mu) &= \frac{\mathbb{E}\left[\mathrm{Cov}(\mu^\star(X,Z), \mu(X,Z) \mid Z)\right]}{\sqrt{\mathbb{E}\left[\mathrm{Var}(\mu(X,Z) \mid Z)\right]}}, &\text{(I.2)}\\ f_n^{\mathcal{T}}(\mu) &= \frac{\mathbb{E}\left[\mathrm{Cov}(\mu^\star(X_i,Z_i), \mu(X_i,Z_i) \mid \boldsymbol{Z},\boldsymbol{T})\right]}{\sqrt{\mathbb{E}\left[\mathrm{Var}(\mu(X_i,Z_i) \mid \boldsymbol{Z},\boldsymbol{T})\right]}}, &\text{(I.3)} \end{aligned}$$

then denote $W_i = (X_i,Z_i)$, $h(W_i) := \mu(W_i) - \mathbb{E}\left[\mu(W_i) \mid Z_i\right]$, $h^{\mathcal{T}}(W_i) := \mu^\star(W_i) - \mathbb{E}\left[\mu^\star(W_i) \mid \boldsymbol{Z},\boldsymbol{T}\right]$ and assume $\mathbb{E}\left[h^2(W_i)\right] = 1$ without loss of generality. First notice a simple fact $|\frac{a}{b} - \frac{c}{d}| = \frac{|ad-bc|}{bd} = \frac{|ad-cd+cd-bc|}{bd} \le \frac{|a-c|}{b} + \frac{c|b-d|}{bd}$ for $a,b,c,d > 0$, then let the numerator and denominator of $f(\mu)$ in (I.2) to be a,b respectively (similarly denote c,d for $f_n^{\mathcal{T}}(\mu)$ in (I.3)). And we have

$$\max\{\frac{1}{b}, \frac{c}{bd}\} \le 1 + f_n^{\mathcal{T}}(\mu) \le 1 + f_n^{\mathcal{T}}(\mu^\star) \le 1 + f(\mu^\star) \le 1 + \mathbb{E}\left[(\mu^\star)^2(X,Z)\right] < \infty,$$

hence it suffices to bound $|a-c|$ and $|b-d|$. First we have the following

$$\begin{aligned} a - c &= \mathbb{E}\left[\mathrm{Cov}\left(\mu^\star(W_i), \mu(W_i) \mid \boldsymbol{Z}\right)\right] - \mathbb{E}\left[\mathrm{Cov}\left(\mu^\star(W_i), \mu(W_i) \mid \boldsymbol{Z},\boldsymbol{T}\right)\right] &\text{(I.4)}\\ &= \mathbb{E}\left[\mathrm{Cov}\left(\mathbb{E}\left[\mu^\star(W_i) \mid \boldsymbol{Z},\boldsymbol{T}\right], \mathbb{E}\left[\mu(W_i) \mid \boldsymbol{Z},\boldsymbol{T}\right] \mid \boldsymbol{Z}\right)\right]\\ &= \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Cov}_{\boldsymbol{T}|\boldsymbol{Z}}\left(\mathbb{E}\left[\mu^\star(W_i) \mid \boldsymbol{Z},\boldsymbol{T}\right], \mathbb{E}\left[\mu(W_i) \mid \boldsymbol{Z},\boldsymbol{T}\right]\right)\right]. \end{aligned}$$

where the first equality holds due to the independence among *i.i.d.* samples $(\boldsymbol{X},\boldsymbol{Z}) = \{(X_i,Z_i)\}_{i=1}^n$. For the second equality, we apply the law of total covariance to the covariance term $\mathrm{Cov}\left(\mu^\star(W_i), \mu(W_i) \mid \boldsymbol{Z}\right)$ then cancel out the second term of the first line, leading to the term in the second line. Finally we spell out the randomness of the expectation and covariance through explicit subscripts in the last inequality. They by applying Cauchy–Schwarz inequality, we obtain

$$|a-c| \le \sqrt{\mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_{\boldsymbol{T}|\boldsymbol{Z}}\left(\mathbb{E}\left[\mu^\star(W_i) \mid \boldsymbol{Z},\boldsymbol{T}\right]\right)\right]}\sqrt{\mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_{\boldsymbol{T}|\boldsymbol{Z}}\left(\mathbb{E}\left[\mu(W_i) \mid \boldsymbol{Z},\boldsymbol{T}\right]\right)\right]} \tag{I.5}$$

Regarding the term $|b-d|$, we have

$$\begin{aligned} |b-d| &= \left|\sqrt{\mathbb{E}\left[h^2(W_i)\right]} - \sqrt{\mathbb{E}\left[(h^{\mathcal{T}})^2(W_i)\right]}\right|\\ &= \frac{\left|\mathbb{E}\left[h^2(W_i)\right] - \mathbb{E}\left[(h^{\mathcal{T}})^2(W_i)\right]\right|}{\sqrt{\mathbb{E}\left[h^2(W_i)\right]} + \sqrt{\mathbb{E}\left[(h^{\mathcal{T}})^2(W_i)\right]}}\\ &\le \frac{\left|\mathbb{E}\left[h^2(W_i)\right] - \mathbb{E}\left[(h^{\mathcal{T}})^2(W_i)\right]\right|}{\sqrt{\mathbb{E}\left[h^2(W_i)\right]}}\\ &\le \mathbb{E}\left[\mathrm{Var}\left(\mu(W_i) \mid \boldsymbol{Z}\right)\right] - \mathbb{E}\left[\mathrm{Var}\left(\mu(W_i) \mid \boldsymbol{Z},\boldsymbol{T}\right)\right]\\ &= \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_{\boldsymbol{T}|\boldsymbol{Z}}\left(\mathbb{E}\left[\mu(W_i) \mid \boldsymbol{Z},\boldsymbol{T}\right]\right)\right] &\text{(I.6)} \end{aligned}$$

where we use the assumption $\mathbb{E}\left[h^2(W_i)\right]=1$ and the definition of $h, h^{\mathcal{T}}$ in the second inequality. The last equality holds as a result of applying the law of total variance to the variance term $\operatorname{Var}\left(\mu(W_i)\,|\,\boldsymbol{Z}\right)$ then getting the second term of line 4 cancelled out. Finally, combining (I.5) and (I.6) establishes the bound in (I.1). □

I.2.1 Proposition 3.5

Proof of Proposition 3.5. Throughout the proof, the natural sufficient statistic and $f_n^{\mathcal{T}}(\mu)$ are defined based on the batch $\mathcal{B}_m$ whose sample size is n_2. But we will abbreviate the notation dependence on it for simplicity and use a generic n instead of n_2 to avoid carrying too many subscripts, without causing any confusion. Now we present a roadmap of this proof.

(i) due to Lemma I.1, it suffices to bound the term $\mathrm{II}(\mu)$, $\mathrm{II}(\mu^\star)$ in (I.1).

(ii) we bound $\mathrm{II}(\mu)$, $\mathrm{II}(\mu^\star)$ with the same strategy. Specifically, we will show

$$\mathrm{II}(\mu)=O\left(\mathbb{E}_{Z_i}\left[\mathbb{E}_F\left[\mu^2(W_i)\right]\mathbb{E}\left[h_{ii}\,|\,Z_i\right]\right]\right)$$

and similarly for $\mathrm{II}(\mu^\star)$ under the stated model, where F denotes the conditional distribution of $X_i|\boldsymbol{Z}$, and h_{ii} is the ith diagonal term of the hat matrix $\boldsymbol{H}$, which is defined later. This terminology comes from the fact that we can treat X_j as response variable, $(1,Z)$ as predictors, the natural sufficient statistic for this low dimensional multivariate Gaussian distribution is equivalent to the OLS estimator.

(iii) Regarding the term $\mathbb{E}\left[h_{ii}\,|\,Z_i\right]$ above, we can carefully bound it by $1/(n-1)+\mathbb{E}\left[\boldsymbol{\Xi}\,|\,Z_i\right]$, where $\boldsymbol{\Xi}$ is defined in (I.16).

(iv) Simply expanding $\mathbb{E}\left[\boldsymbol{\Xi}\,|\,Z_i\right]$ into three terms: $\mathrm{III}_1, \mathrm{III}_2, \mathrm{III}_3$, which are defined in (I.17), (I.18) and (I.18), we will show $\mathrm{III}_2=0$ and figure out the stochastic representation of $\mathrm{III}_1, \mathrm{III}_3$, which turns out to be related to chi-squared, Wishart and inverse-Wishart random variables.

(v) Cauchy–Schwarz inequalities together with some properties of those random variables (chi-squared, Wishart and inverse-Wishart) and the stated moment conditions finally gives us the result in (3.4).

Having proved Lemma I.1, now we directly start with step (ii). Notice the following

$$\begin{aligned}
\mathrm{II}(\mu) &= \mathbb{E}_{\boldsymbol{Z}}\left[\operatorname{Var}_{\boldsymbol{T}|\boldsymbol{Z}}\left(\mathbb{E}\left[\mu(W_i)\,|\,\boldsymbol{Z},\boldsymbol{T}\right]\right)\right] \\
&= \mathbb{E}_{\boldsymbol{Z}}\left[\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\left(\mathbb{E}_F\left[\mu(W_i)\right]-\mathbb{E}_{F_{\boldsymbol{T}}}\left[\mu(W_i)\right]\right)^2\right]\right] \\
&= \mathbb{E}_{\boldsymbol{Z}}\left[\operatorname{Var}_F\left(\mu(W_i)\right)\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\frac{\left(\mathbb{E}_F\left[\mu(W_i)\right]-\mathbb{E}_{F_{\boldsymbol{T}}}\left[\mu(W_i)\right]\right)^2}{\operatorname{Var}_F\left(\mu(W_i)\right)}\right]\right] \\
&\leq \mathbb{E}_{\boldsymbol{Z}}\left[\operatorname{Var}_F\left(\mu(W_i)\right)\min\left\{\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\chi^2(F_{\boldsymbol{T}}\|F)\right],2\right\}\right] \qquad \text{(I.7)}
\end{aligned}$$

where the second equality is just rewriting the conditional variance, with F denoting the conditional distribution $X_i|\boldsymbol{Z}$ and $F_{\boldsymbol{T}}$ denoting the conditional distribution $X_i|\boldsymbol{Z},\boldsymbol{T}$. Here we abbreviate the subscript dependence on i for notation simplicity. The third equality holds since $\operatorname{Var}_F\left(\mu(W_i)\right)\in\mathscr{A}(\boldsymbol{Z})$. Regarding the last inequality, we make use of the variational representation of χ^2-divergence:

$$\chi^2(P\|Q)=\sup_{\mu}\frac{\left(\mathbb{E}_P(\mu)-\mathbb{E}_Q(\mu)\right)^2}{\operatorname{Var}_Q(\mu)}$$

and the fact that

$$\begin{aligned}
& \mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\frac{(\mathbb{E}_F\left[\mu(W_i)\right]-\mathbb{E}_{F_{\boldsymbol{T}}}\left[\mu(W_i)\right])^2}{\mathrm{Var}_F\left(\mu(W_i)\right)}\right] \\
\leq\ & \frac{\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\mathbb{E}_F\left[\mu^2(W_i)\right]\right]+\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\mathbb{E}_{F_{\boldsymbol{T}}}\left[\mu^2(W_i)\right]\right]-2\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\mathbb{E}_{F_{\boldsymbol{T}}}\left[\mu(W_i)\right]\mathbb{E}_F\left[\mu(W_i)\right]\right]}{\mathrm{Var}_F\left(\mu(W_i)\right)} \\
=\ & \frac{\mathbb{E}_F\left[\mu^2(W_i)\right]+\mathbb{E}_F\left[\mu^2(W_i)\right]-2(\mathbb{E}_F\left[\mu(W_i)\right])^2}{\mathrm{Var}_F\left(\mu(W_i)\right)} \\
=\ & \frac{2\mathrm{Var}_F\left(\mu(W_i)\right)}{\mathrm{Var}_F\left(\mu(W_i)\right)}=2
\end{aligned}$$

where the first inequality is from expanding the quadratic term and the fact $(\mathbb{E}_F\left[\mu(W_i)\right])^2 \leq \mathbb{E}_F\left[\mu^2(W_i)\right]$, $(\mathbb{E}_{F_{\boldsymbol{T}}}\left[\mu(W_i)\right])^2 \leq \mathbb{E}_{F_{\boldsymbol{T}}}\left[\mu^2(W_i)\right]$, the first equality holds as a result of the tower property of conditional expectation and $\mathbb{E}_F\left[\mu(W_i)\right] \in \mathscr{A}(\boldsymbol{Z})$. Denote $u_i = (1, Z_i)^\top$ and the following n by p matrix by $\boldsymbol{U}$:

$$\boldsymbol{U} = \begin{pmatrix} u_1^\top \\ \vdots \\ u_n^\top \end{pmatrix} = (\mathbf{1}, \boldsymbol{Z}) \tag{I.8}$$

Recall that the sufficient statistic (here we ignore the batching index)

$$\boldsymbol{T} = (\sum_{i\in[n]} X_i, \sum_{i\in[n]} X_i Z_i) = \boldsymbol{U}^\top \boldsymbol{X},$$

under the stated multivariate Gaussian model, we know $\boldsymbol{X} \mid \boldsymbol{Z} \sim \mathcal{N}\left(\boldsymbol{U}\gamma, \sigma^2 \boldsymbol{I}_n\right)$, then the conditional distribution of $(X_i, \boldsymbol{T}) \mid \boldsymbol{Z}$ can be specified as below

$$\begin{pmatrix} X_i \\ \boldsymbol{T} \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} (1, Z_i)\gamma \\ \boldsymbol{U}^\top \boldsymbol{U}\gamma \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & e_i^\top \boldsymbol{U} \\ \boldsymbol{U}^\top e_i^\top & \boldsymbol{U}^\top \boldsymbol{U} \end{bmatrix}\right) \tag{I.9}$$

where $e_i \in \mathbb{R}^n$, $(e_1, \cdots, e_n)$ forms the standard orthogonal basis. Noticing the above joint distribution is multivariate Gaussian, we can immediately derive the conditional distribution as below,

$$X_i \mid \boldsymbol{Z}, \boldsymbol{T} \sim \mathcal{N}\left(e_i^\top \boldsymbol{U}(\boldsymbol{U}^\top \boldsymbol{U})^{-1}\boldsymbol{U}^\top \boldsymbol{X}, \sigma^2(1 - e_i^\top \boldsymbol{U}(\boldsymbol{U}^\top \boldsymbol{U})^{-1}\boldsymbol{U}^\top e_i)\right).$$

Denote $\boldsymbol{H} = \boldsymbol{U}(\boldsymbol{U}^\top \boldsymbol{U})^{-1}\boldsymbol{U}^\top$, which is the “hat” matrix. Now we compactly write down the following two conditional distributions:

$$\begin{aligned}
F_{\boldsymbol{T}} &: X_i \mid \boldsymbol{Z}, \boldsymbol{T} \sim \mathcal{N}\left(e_i^\top \boldsymbol{H}\boldsymbol{X}, \sigma^2(1 - h_{ii})\right) \\
F &: X_i \mid \boldsymbol{Z} \sim \mathcal{N}\left((1, Z_i)\gamma, \sigma^2\right)
\end{aligned}$$

Note the sufficient statistic $\boldsymbol{T}$ is equivalent to

$$\hat{\gamma}^{OLS} = (\boldsymbol{U}^\top \boldsymbol{U})^{-1}\boldsymbol{U}^\top \boldsymbol{X}$$

whenever $\boldsymbol{U}^\top \boldsymbol{U}$ is nonsingular. Here $\hat{\gamma}^{OLS}$ is the OLS estimator for γ (when treating X as response variable, $(1, Z)$ as predictors). Simply, we have

$$\hat{\gamma}^{OLS} \sim \mathcal{N}\left(\gamma, \sigma^2(\boldsymbol{U}^\top \boldsymbol{U})^{-1}\right)$$

Now we are ready to calculate $\chi^2(F_{\boldsymbol{T}} \| F)$. First,

$$\begin{aligned}
e_i^\top \boldsymbol{H}\boldsymbol{X} - (1, Z_i)\gamma &= e_i^\top \boldsymbol{U}\hat{\gamma}^{OLS} - (1, Z_i)\gamma \\
&= e_i^\top \boldsymbol{U}(\hat{\gamma}^{OLS} - \gamma) \sim \mathcal{N}(0, \sigma^2 h_{ii})
\end{aligned} \tag{I.10}$$

Since $2\sigma^2 > \sigma^2(1-h_{ii})$, applying Lemma I.2 yields the following

$$\begin{aligned}
\chi^2(F_{\boldsymbol{T}} \| F) &= \frac{1}{2}\left[\frac{1}{\sqrt{1-h_{ii}^2}} \exp\left\{\frac{(e_i^\top \boldsymbol{H}\boldsymbol{X} - (1, Z_i)\gamma)^2}{\sigma^2(1+h_{ii})}\right\} - 1\right] \\
&\leq \frac{1}{\sqrt{1-h_{ii}}} \exp\left\{\frac{(e_i^\top \boldsymbol{H}\boldsymbol{X} - (1, Z_i)\gamma)^2}{\sigma^2(1+h_{ii})}\right\} - 1 \\
&= \frac{1}{\sqrt{1-h_{ii}}} \exp\left\{\frac{h_{ii}G^2}{1+h_{ii}}\right\} - 1 \qquad \text{(I.11)}
\end{aligned}$$

where $G \sim \mathcal{N}(0,1)$ is independent from $\boldsymbol{X}$ and the last equality holds due to (I.10). Plugin (I.11) back to (I.7), we have

$$\begin{aligned}
\mathrm{II}(\mu) &\leq \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_F(\mu(W_i)) \min\left\{\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\chi^2(F_{\boldsymbol{T}} \| F)\right], 2\right\}\right] \\
&\leq \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_F(\mu(W_i)) \min\left\{\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\frac{1}{\sqrt{1-h_{ii}}} \exp\left\{\frac{h_{ii}G^2}{1+h_{ii}}\right\} - 1\right], 2\right\}\right]
\end{aligned}$$

Note the moment generating function for χ_1^2 random variable is $\frac{1}{\sqrt{1-2t}}$ when $t < 1/2$. Since the expectation of $\exp\left\{\frac{h_{ii}G^2}{1+h_{ii}}\right\}$ does not always exist, we consider two events E and E^c such that conditional on the event E, the expectation exists and the probability of event E^c is small. More specifically, define the event $E = \{h_{ii} < \frac{1}{2}\}$, which implies

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\frac{1}{\sqrt{1-h_{ii}}} \exp\left\{\frac{h_{ii}G^2}{1+h_{ii}}\right\}\right] - 1 &= \frac{1}{\sqrt{1-h_{ii}}\sqrt{1-2h_{ii}/(1+h_{ii})}} - 1 \\
&= \frac{\sqrt{1+h_{ii}}}{1-h_{ii}} - 1 \\
&\leq \frac{1+h_{ii}}{1-h_{ii}} - 1 \\
&\leq 4h_{ii}
\end{aligned}$$

hence we can bound $\mathrm{II}(\mu)$ by the summation of the following two terms:

$$\mathrm{II}_1 := \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_F(\mu(W_i))\, \mathbb{1}_{\{E\}} \cdot 4h_{ii}\right], \quad \mathrm{II}_2 := \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_F(\mu(W_i))\, \mathbb{1}_{\{E^c\}} \cdot 2\right]$$

Regarding II_1, the following holds:

$$\mathrm{II}_1 \leq 4\, \mathbb{E}_{Z_i}\left[\mathbb{E}_F\left[\mu^2(W_i)\right] \mathbb{E}\left[h_{ii} \,|\, Z_i\right]\right],$$

where we apply the tower property of conditional expectation and $\mathrm{Var}_F(\mu(W_i)) \leq \mathbb{E}_F\left[\mu^2(W_i)\right] \in \mathscr{A}(Z_i)$ Regarding II_2, we have

$$\begin{aligned}
\mathrm{II}_2 &= 2\, \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_F(\mu(W_i))\, \mathbb{1}_{\{E^c\}}\right] \\
&= 2\, \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_F(\mu(W_i))\, \mathbb{E}\left[\mathbb{1}_{\{E^c\}} \,|\, Z_i\right]\right] \\
&\leq 2\, \mathbb{E}_{Z_i}\left[\mathbb{E}_F\left[\mu^2(W_i)\right] \mathbb{P}\left(h_{ii} \geq \frac{1}{2} \,|\, Z_i\right)\right] \\
&\leq 4\, \mathbb{E}_{Z_i}\left[\mathbb{E}_F\left[\mu^2(W_i)\right] \mathbb{E}\left[h_{ii} \,|\, Z_i\right]\right]
\end{aligned}$$

where the second equality comes from the tower property of conditional expectation and $\mathrm{Var}_F(\mu(W_i)) \in \mathscr{A}(Z_i)$ and the last inequality holds due to Markov's inequality. Now we can compactly write down the following bound for $\mathrm{II}(\mu)$,

$$\mathrm{II}(\mu) \leq \mathrm{II}_1 + \mathrm{II}_2 \leq 8\, \mathbb{E}_{Z_i}\left[\mathbb{E}_F\left[\mu^2(W_i)\right] \mathbb{E}\left[h_{ii} \,|\, Z_i\right]\right], \qquad \text{(I.12)}$$

Similarly we obtain $\mathrm{II}(\mu^\star) = O\left(\mathbb{E}_{Z_i}\left[\mathbb{E}_F\left[(\mu^\star)^2(W_i)\right]\mathbb{E}\left[h_{ii}\,|\,Z_i\right]\right]\right)$. Now we proceed step (iii), i.e. calculating $\mathbb{E}\left[h_{ii}\,|\,Z_i\right]$. Notice h_{ii} is the ith diagonal term of the "hat" matrix, which involves $\{w_i\}_{i=1}^n$. In order to bound the conditional expectation of h_{ii} given Z_i in a sharp way, we carefully expand h_{ii} and try to get w_i separated from $\{w_m\}_{m\neq i}$. Recall the definition of $\boldsymbol{U} = (\mathbf{1}, \boldsymbol{Z})$ in (I.8), we can rewrite

$$\boldsymbol{U}^\top\boldsymbol{U} = \sum_{m\neq i} u_m u_m^\top + u_i u_i^\top, \quad \boldsymbol{A} := \sum_{m\neq i} u_m u_m^\top$$

Note that $h_{ii} = u_i^\top(\boldsymbol{U}^\top\boldsymbol{U})^{-1}u_i$ since $\boldsymbol{H} = \boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{U}^\top$, hence we have

$$h_{ii} = u_i^\top(\boldsymbol{A} + u_i u_i^\top)^{-1}u_i$$

As $n > p$, $\boldsymbol{A}$ is almost surely positive definite thus invertible, then applying Sherman–Morrison formula to $\boldsymbol{A}$ and $u_i u_i^\top$ yields the following

$$h_{ii} = u_i^\top\boldsymbol{A}^{-1}u_i - \frac{(u_i^\top\boldsymbol{A}^{-1}u_i)^2}{1 + u_i^\top\boldsymbol{A}^{-1}u_i} \leq u_i^\top\boldsymbol{A}^{-1}u_i. \tag{I.13}$$

Since $\boldsymbol{A}$ also involves the unit vector $\mathbf{1}_{n-1}$, it is easier when we first project $\boldsymbol{Z}_{\text{-}i}$ on $\mathbf{1}_{n-1}$ then work with the orthogonal complement. Bearing this idea in mind, we denote $\boldsymbol{\Omega} = (\mathbf{1}_{n-1}, \boldsymbol{Z}_{\text{-}i})$ which is a $n-1$ by p matrix, then rewrite $\boldsymbol{A}$ as

$$\boldsymbol{A} = \boldsymbol{\Omega}^\top\boldsymbol{\Omega} = \begin{pmatrix} \mathbf{1}_{n-1}^\top\mathbf{1}_{n-1} & \mathbf{1}_{n-1}^\top\boldsymbol{Z}_{\text{-}i} \\ \boldsymbol{Z}_{\text{-}i}^\top\mathbf{1}_{n-1} & \boldsymbol{Z}_{\text{-}i}^\top\boldsymbol{Z}_{\text{-}i} \end{pmatrix}$$

where $\mathbf{I}_{n-1}$ is the $(n-1)$ dimensional identity matrix. Denote

$$\overline{\boldsymbol{Z}_{\text{-}i}} := \frac{1}{n-1}\sum_{m\neq i} Z_m = \frac{1}{n-1}\mathbf{1}_{n-1}^\top\boldsymbol{Z}_{\text{-}i} \quad \boldsymbol{\Gamma} := \begin{pmatrix} 1 & -\overline{\boldsymbol{Z}_{\text{-}i}} \\ \mathbf{0} & \mathbf{I}_{n-1} \end{pmatrix}, \tag{I.14}$$

we have

$$\begin{aligned} \boldsymbol{\Omega}\boldsymbol{\Gamma} = (\mathbf{1}_{n-1}, \boldsymbol{Z}_{\text{-}i})\boldsymbol{\Gamma} &= (\mathbf{1}_{n-1}, \boldsymbol{Z}_{\text{-}i} - \mathbf{1}_{n-1}\overline{\boldsymbol{Z}_{\text{-}i}}) \\ &= (\mathbf{1}_{n-1}, (\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}_{\text{-}i}). \end{aligned}$$

where $\boldsymbol{P}_{n-1} = \mathbf{1}_{n-1}\mathbf{1}_{n-1}^\top/(n-1)$ is the projection matrix onto $\mathbf{1}_{n-1}$. Then we immediately have

$$(\boldsymbol{\Omega}\boldsymbol{\Gamma})^\top\boldsymbol{\Omega}\boldsymbol{\Gamma} = \begin{pmatrix} n-1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z}_{\text{-}i}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}_{\text{-}i} \end{pmatrix}$$

since $\boldsymbol{P}_{n-1}\mathbf{1}_{n-1} = \mathbf{1}_{n-1}, (\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\mathbf{1}_{n-1} = \mathbf{0}$ and

$$u_i^\top\boldsymbol{\Gamma} = (1, Z_i)\boldsymbol{\Gamma} = (1, Z_i - \overline{\boldsymbol{Z}_{\text{-}i}}). \tag{I.15}$$

Combining (I.14) with (I.15) yields the following

$$\begin{aligned} u_i^\top\boldsymbol{A}^{-1}u_i &= u_i^\top(\boldsymbol{\Omega}^\top\boldsymbol{\Omega})^{-1}u_i \\ &= u_i^\top\boldsymbol{\Gamma}((\boldsymbol{\Omega}\boldsymbol{\Gamma})^\top\boldsymbol{\Omega}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^\top w_i \\ &= \frac{1}{n-1} + (Z_i - \overline{\boldsymbol{Z}_{\text{-}i}})(\boldsymbol{Z}_{\text{-}i}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}_{\text{-}i})^{-1}(Z_i - \overline{\boldsymbol{Z}_{\text{-}i}})^\top, \end{aligned}$$

which together with (I.13) implies $\mathbb{E}\left[h_{ii}\,|\,Z_i\right] \leq \mathbb{E}\left[u_i^\top\boldsymbol{A}^{-1}u_i\,|\,Z_i\right] = 1/(n-1) + \mathbb{E}\left[\boldsymbol{\Xi}\,|\,Z_i\right]$, where

$$\boldsymbol{\Xi} = (Z_i - \overline{\boldsymbol{Z}_{\text{-}i}})(\boldsymbol{Z}_{\text{-}i}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}_{\text{-}i})^{-1}(Z_i - \overline{\boldsymbol{Z}_{\text{-}i}})^\top. \tag{I.16}$$

As the problem has been reduced to calculating $\mathbb{E}[\boldsymbol{\Xi} \mid Z_i]$, we arrive at the step (iv) now. Write $(\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}_{\text{-}i}) = (\boldsymbol{Z}_i - \boldsymbol{v}_0) - (\overline{\boldsymbol{Z}}_{\text{-}i} - \boldsymbol{v}_0)$, where $\boldsymbol{v}_0$ is the mean of Gaussian random variable Z, we can expand $\mathbb{E}[\boldsymbol{\Xi} \mid Z_i] = \text{III}_1 + \text{III}_2 + \text{III}_3$, where

$$\text{III}_1 = (\boldsymbol{Z}_i - \boldsymbol{v}_0)\mathbb{E}\left[(\boldsymbol{Z}_{\text{-}i}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}_{\text{-}i})^{-1} \mid \boldsymbol{Z}_i\right](\boldsymbol{Z}_i - \boldsymbol{v}_0)^\top \tag{I.17}$$

$$\text{III}_2 = -2(\boldsymbol{Z}_i - \boldsymbol{v}_0)\mathbb{E}\left[(\boldsymbol{Z}_{\text{-}i}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}_{\text{-}i})^{-1}(\overline{\boldsymbol{Z}}_{\text{-}i} - \boldsymbol{v}_0)^\top \mid \boldsymbol{Z}_i\right] \tag{I.18}$$

$$\text{III}_3 = \mathbb{E}\left[(\overline{\boldsymbol{Z}}_{\text{-}i} - \boldsymbol{v}_0)(\boldsymbol{Z}_{\text{-}i}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}_{\text{-}i})^{-1}(\overline{\boldsymbol{Z}}_{\text{-}i} - \boldsymbol{v}_0)^\top \mid \boldsymbol{Z}_i\right] \tag{I.19}$$

Below we are going to show $\text{III}_2 = 0$ and derive $\text{III}_1, \text{III}_3$ carefully. Regarding the term III_1, we exactly write down its stochastic representation. Under the state Gaussian model, we have $\boldsymbol{Z}_{\text{-}i}^\top \sim \mathcal{N}\left(\boldsymbol{v}_0\mathbf{1}_{n-1}^\top, \mathbf{I}_{n-1} \otimes \boldsymbol{\Sigma}_0\right)$, then $(\boldsymbol{Z}_{\text{-}i}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}_{\text{-}i})^{-1}$ follows an inverse Wishart distribution i.e.

$$(\boldsymbol{Z}_{\text{-}i}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}_{\text{-}i})^{-1} \sim \mathcal{W}_{p-1}^{-1}(\boldsymbol{\Sigma}_0^{-1}, n-2)$$

and $\boldsymbol{Z}_{\text{-}i} \perp\!\!\!\perp \boldsymbol{Z}_i$, hence we can calculate

$$\mathbb{E}\left[(\boldsymbol{Z}_{\text{-}i}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}_{\text{-}i})^{-1} \mid \boldsymbol{Z}_i\right] = \frac{\boldsymbol{\Sigma}_0^{-1}}{n-p-2}.$$

Plug in the above equation into (I.17), we have

$$\text{III}_1 = (\boldsymbol{Z}_i - \boldsymbol{v}_0)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{Z}_i - \boldsymbol{v}_0)^\top = \frac{\boldsymbol{\Phi}}{n-p-2}, \quad \text{where } \boldsymbol{\Phi} \sim \chi^2_{p-1},\ \boldsymbol{\Phi} \perp\!\!\!\perp \boldsymbol{Z}_{\text{-}i}. \tag{I.20}$$

Regarding the term III_2 in (I.18), we first denote $\boldsymbol{Z} = \boldsymbol{Z}_{\text{-}i} - \mathbf{1}_{n-1}\boldsymbol{v}_0$ and notice

$$\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n-1} \otimes \boldsymbol{\Sigma}_0), \quad \mathbf{1}_{n-1}^\top \boldsymbol{Z} = (n-1)(\overline{\boldsymbol{Z}}_{\text{-}i} - \boldsymbol{v}_0), \tag{I.21}$$

then rewrite III_2 as below

$$\text{III}_2 = -2(\boldsymbol{Z}_i - \boldsymbol{v}_0)\mathbb{E}\left[((\boldsymbol{Z} + \mathbf{1}_{n-1}\boldsymbol{v}_0)^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})(\boldsymbol{Z} + \mathbf{1}_{n-1}\boldsymbol{v}_0))^{-1}\frac{(\mathbf{1}_{n-1}^\top \boldsymbol{Z})^\top}{n-1}\right]$$

where we also makes use of the fact that

$$(\boldsymbol{Z}_{\text{-}i}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}_{\text{-}i})^{-1}(\overline{\boldsymbol{Z}}_{\text{-}i} - \boldsymbol{v}_0)^\top \perp\!\!\!\perp \boldsymbol{Z}_i$$

Noticing that $(\mathbf{1}_{n-1}\boldsymbol{v}_0)^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1}) = \mathbf{0}$, we can simplify further

$$\text{III}_2 = -\frac{2}{n-1}(\boldsymbol{Z}_i - \boldsymbol{v}_0)\mathbb{E}\left[(\boldsymbol{Z}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z})^{-1}(\mathbf{1}_{n-1}^\top \boldsymbol{Z})^\top\right] \tag{I.22}$$

Notice in the above equation, $\boldsymbol{Z}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})$ is the orthogonal complement of $\boldsymbol{Z}^\top \mathbf{1}_{n-1}$, which implies independence under the Gaussian distribution assumption, which we will now use to prove the expectation in (I.22) equals zero. Formally, we first have $(\boldsymbol{Z}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1}), \boldsymbol{Z}^\top \mathbf{1}_{n-1})$ are multivariate Gaussian. Introducing the vectorization of matrix and the Kronecker product, we can express in the following way:

$$\text{vec}(\boldsymbol{Z}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})) = (\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1}) \otimes \mathbf{I}_{p-1}\text{vec}(\boldsymbol{Z}^\top), \quad \text{vec}(\boldsymbol{Z}^\top) = \mathbf{1}_{n-1} \otimes \mathbf{I}_{p-1}\text{vec}(\boldsymbol{Z}^\top).$$

Now we are ready to calculate the covariance

$$\begin{aligned} & \text{Cov}\left(\text{vec}(\boldsymbol{Z}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})), \text{vec}(\boldsymbol{Z}^\top \mathbf{1}_{n-1})\right) \\ = & ((\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1}) \otimes \mathbf{I}_{p-1})(\mathbf{I}_{n-1} \otimes \boldsymbol{\Sigma}_0)(\mathbf{1}_{n-1} \otimes \mathbf{I}_{p-1})^\top \\ = & ((\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\mathbf{I}_{n-1}\mathbf{1}_{n-1}) \otimes (\mathbf{I}_{p-1}\boldsymbol{\Sigma}_0\mathbf{I}_{p-1}) = \mathbf{0} \end{aligned}$$

where in above equalities we use the fact $\mathrm{Var}\left(\mathrm{vec}(\boldsymbol{Z}^\top)\right) = \mathbf{I}_{n-1} \otimes \boldsymbol{\Sigma}_0$ in (I.21) and the mixed-product property of the Kronecker product. Therefore

$$\boldsymbol{Z}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1}) \perp\!\!\!\perp \boldsymbol{Z}^\top \mathbf{1}_{n-1} \quad \Longrightarrow \quad \mathrm{III}_2 = 0 \tag{I.23}$$

Regarding the term III_3, first denote $\boldsymbol{\Psi}_1 = \boldsymbol{Z}^\top \boldsymbol{P}_{n-1}\boldsymbol{Z}$ and $\boldsymbol{\Psi}_2 = \boldsymbol{Z}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}$, we obtain two independent Wishart random variables i.e.

$$\boldsymbol{\Psi}_1 \sim \mathcal{W}_{p-1}(\boldsymbol{\Sigma}_0, 1), \quad \boldsymbol{\Psi}_2 \sim \mathcal{W}_{p-1}(\boldsymbol{\Sigma}_0, n-2), \quad \boldsymbol{\Psi}_1 \perp\!\!\!\perp \boldsymbol{\Psi}_2.$$

Then III_3 can be calculated as below

$$\begin{aligned}
\mathrm{III}_3 &= \mathbb{E}\left[(\overline{\boldsymbol{Z}}_{\text{-}i} - \boldsymbol{v}_0)(\boldsymbol{Z}_{\text{-}i}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}_{\text{-}i})^{-1}(\overline{\boldsymbol{Z}}_{\text{-}i} - \boldsymbol{v}_0)^\top \mid \boldsymbol{Z}_i\right] \\
&= \mathbb{E}\left[\mathbf{1}_{n-1}^\top \boldsymbol{Z}(\boldsymbol{Z}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \mathbf{1}_{n-1}\right]/(n-1)^2 \\
&= \mathbb{E}\left[\mathrm{Tr}\left(\mathbf{1}_{n-1}^\top \boldsymbol{Z}(\boldsymbol{Z}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \mathbf{1}_{n-1}\right)\right]/(n-1)^2 \\
&= \mathbb{E}\left[\mathrm{Tr}(\boldsymbol{\Psi}_1\boldsymbol{\Psi}_2^{-1})\right]/(n-1) \\
&= \mathrm{Tr}\,\mathbb{E}\left[\boldsymbol{\Psi}_1\boldsymbol{\Psi}_2^{-1}\right]/(n-1) \\
&= \mathrm{Tr}(\mathbb{E}\left[\boldsymbol{\Psi}_1\right]\mathbb{E}\left[\boldsymbol{\Psi}_2^{-1}\right])/(n-1) \\
&= \mathrm{Tr}(\boldsymbol{\Sigma}_0 \frac{\boldsymbol{\Sigma}_0^{-1}}{n-p-2})/(n-1) \\
&= \frac{p}{(n-1)(n-p-2)}
\end{aligned} \tag{I.24}$$

where the first equality is from (I.19), the second equality is similarly obtained as (I.22), the fourth equality holds by the fact $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ and the definition of $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$, the sixth equality holds due to $\boldsymbol{\Psi}_1 \perp\!\!\!\perp \boldsymbol{\Psi}_2$. So far we have shown $\mathrm{III}_2 = 0$ and figured out the stochastic representation of $\mathrm{III}_2, \mathrm{III}_3$, which are also further simplified using the properties of Wishart and inverse-Wishart random variables. These bring us to the final stage i.e. step (v). Combining (I.13), (I.20), (I.23) and (I.24), we finally obtain

$$\begin{aligned}
\mathbb{E}\left[h_{ii} \mid Z_i\right] &\leq \mathbb{E}\left[u_i^\top \boldsymbol{A}^{-1} u_i \mid Z_i\right] \\
&\leq \frac{1}{n-1} + \mathbb{E}\left[\boldsymbol{\Xi} \mid Z_i\right] \\
&= \frac{1}{n-1} + \mathrm{III}_1 + \mathrm{III}_2 + \mathrm{III}_3 \\
&\leq \frac{1}{n-1} \cdot \frac{n-2}{n-p-2} + \frac{\boldsymbol{\Phi}}{n-p-2}
\end{aligned} \tag{I.25}$$

Recall the bound for $\mathrm{II}(\mu)$ in (I.12), then we apply the Cauchy–Schwarz inequality to $\mathbb{E}\left[\mu^2(W_i) \mid \boldsymbol{Z}_i\right]$ and $\mathbb{E}\left[h_{ii} \mid \boldsymbol{Z}_i\right]$, which yields

$$\begin{aligned}
\mathrm{II}(\mu) &\leq 8\ \mathbb{E}_{Z_i}\left[\mathbb{E}_F\left[\mu^2(W_i)\right]\mathbb{E}\left[h_{ii} \mid Z_i\right]\right] \\
&\leq \frac{8(n-2)\mathbb{E}\left[\mu^2(W_i)\right]}{(n-1)(n-p-2)} + \frac{8\sqrt{\mathbb{E}\left[\boldsymbol{\Phi}^2\right]}}{n-p-2}\sqrt{\mathbb{E}_{\boldsymbol{Z}_i}\left[\mathbb{E}\left[\mu^4(W_i) \mid \boldsymbol{Z}_i\right]\right]} \\
&\leq \frac{8\sqrt{\mathbb{E}\left[\mu^4(X,Z)\right]}}{n-p-2}\left(1 + \sqrt{\mathbb{E}\left[\boldsymbol{\Phi}^2\right]}\right)
\end{aligned} \tag{I.26}$$

where in the above equality, $\boldsymbol{\Phi} \sim \chi_{p-1}^2$ and is independent from $\boldsymbol{Z}_{\text{-}i}$. Since $\mathbb{E}\left[\boldsymbol{\Phi}^2\right] \leq p^2$, under the assumption $\mathbb{E}\left[\mu^4(X,Z)\right] < \infty$, we obtain the following bound on $\mathrm{II}(\mu)$,

$$\mathrm{II}(\mu) = O\left(\frac{p}{n-p-2}\right). \tag{I.27}$$

Replacing the μ function by $\mu^\star$ and applying the assumption $\mathbb{E}\left[(\mu^\star)^4(X,Z)\right] < \infty$, we can establish the same rate for $\Pi(\mu^\star)$. Shifting back to the n_2 notation, we finally establish (3.4), i.e.

$$f(\mu) - f_n^{\mathcal{T}}(\mu) = O\left(\frac{p}{n_2 - p - 2}\right).$$

□

I.2.2 Proposition 3.6

Proof of Proposition 3.6. From the proposition statement, we know the sufficient statistic $\boldsymbol{T}_m$ and $f_n^{\mathcal{T}}(\mu)$ are defined based on the batch $\mathcal{B}_m$ whose sample size is n_2. Again, we will abbreviate the notation dependence for simplicity, i.e. use a generic n instead of n_2, use $\boldsymbol{T}$ and $\boldsymbol{Z}$ instead of $\boldsymbol{T}_m$ and $\boldsymbol{Z}_m$, as we did in the proof of Proposition 3.5. Following the derivations up to (I.7) in the proof of Proposition 3.5, it suffices to deal with the following term:

$$\Pi(\mu) := \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_F\left(\mu(W_i)\right)\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\chi^2(F^{\boldsymbol{T}}\|F)\right]\right].$$

where F denotes the conditional distribution $X_i|\boldsymbol{Z}$ and $F_{\boldsymbol{T}}$ denotes the conditional distribution $X_i|\boldsymbol{Z},\boldsymbol{T}$. Below we will consider quantifying the χ^2 divergence between $F_{\boldsymbol{T}}$ and F, Let k_1, k_2 be $W_{i,j-1}, W_{i,j+1}$ respectively, we can write down the probability mass function of $F_{\boldsymbol{T}}$ and F:

$$F : \mathbb{P}\left(X_i \,|\, \boldsymbol{Z}\right) = \prod_{k=1}^{K}\left(q(k,k_1,k_2)\right)^{\mathbb{1}\left\{X_i=k, W_{i,j-1}=k_1, W_{i,j+1}=k_1\right\}} \tag{I.28}$$

$$F_{\boldsymbol{T}} : \mathbb{P}\left(X_i \,|\, \boldsymbol{Z}, \boldsymbol{T}\right) = \prod_{k=1}^{K}\left(\hat{q}(k,k_1,k_2)\right)^{\mathbb{1}\left\{X_i=k, W_{i,j-1}=k_1, W_{i,j+1}=k_1\right\}} \tag{I.29}$$

where $\hat{q}(k,k_1,k_2) = N(k,k_1,k_2)/N(:,k_1,k_2)$ and $N(:,k_1,k_2) = \sum_{i=1}^{n}\mathbb{1}_{\{W_{i,j-1}=k_1, W_{i,j+1}=k_2\}}$. Recall the definition of χ^2 divergence between two discrete distributions, we have

$$\chi^2(F_{\boldsymbol{T}}\|F) = \sum_{k=1}^{K}\frac{\left(\hat{q}(k,k_1,k_2) - q(k,k_1,k_2)\right)^2}{q(k,k_1,k_2)}$$

Notice that

$$\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\hat{q}(k,k_1,k_2)\right] = q(k,k_1,k_2), \quad \mathrm{Var}_{\boldsymbol{T}|\boldsymbol{Z}}\left(\hat{q}(k,k_1,k_2)\right) = \frac{q(k,k_1,k_2)(1-q(k,k_1,k_2))}{N(:,k_1,k_2)}$$

hence we can calculate the following conditional expectation,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\chi^2(F_{\boldsymbol{T}}\|F)\right] &= \sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\frac{\left(\hat{q}(k,k_1,k_2) - q(k,k_1,k_2)\right)^2}{q(k,k_1,k_2)}\right] \\
&= \sum_{k=1}^{K}\frac{q(k,k_1,k_2)(1-q(k,k_1,k_2))}{N(:,k_1,k_2)q(k,k_1,k_2)} \\
&= \sum_{k=1}^{K}\frac{K-1}{N(:,k_1,k_2)}
\end{aligned} \tag{I.30}$$

where we use the fact $\sum_{k=1}^{K} q(k,k_1,k_2) = 1$ in the last equality. Now $\Pi(\mu)$ can be calculated as below.

$$\begin{aligned}
\Pi(\mu) &= \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_F\left(\mu(W_i)\right)\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\chi^2(F_{\boldsymbol{T}}\|F)\right]\right] \\
&= \mathbb{E}_{Z_i}\left[\mathrm{Var}_F\left(\mu(W_i)\right)\mathbb{E}\left[\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\chi^2(F_{\boldsymbol{T}}\|F)\right] \,|\, Z_i\right]\right] \\
&= \mathbb{E}_{Z_i}\left[\mathrm{Var}_F\left(\mu(W_i)\right)\mathbb{E}\left[\frac{K-1}{N(:,W_{i,j-1},W_{i,j+1})} \,|\, Z_i\right]\right] \\
&= \mathbb{E}_{Z_i}\left[\mathrm{Var}_F\left(\mu(W_i)\right)\mathbb{E}\left[\frac{K-1}{1+N_{n-1}(W_{i,j-1},W_{i,j+1})} \,|\, Z_i\right]\right]
\end{aligned} \tag{I.31}$$

where the second equality comes from the tower property of conditional expectation, the third equality holds due to (I.30) and $k_1 = W_{i,j-1}, k_2 = W_{i,j+1}$. In term of the fourth equality, we simply use the new notation that $N_{n-1}(W_{i,j-1}, W_{i,j+1}) = \sum_{m\neq i}^{n} \mathbb{1}_{\{W_{m,j-1}=W_{i,j-1}, W_{m,j+1}=W_{i,j+1}\}}$. Due to the independence among *i.i.d.* samples $\{W_i\}_{i=1}^n$, we have, when conditioning on $Z_i = W_{i,\text{-}j}$

$$\mathbb{1}_{\{W_{m,j-1}=W_{i,j-1},\ W_{m,j+1}=W_{i,j+1}\}} \overset{i.i.d.}{\sim} \mathrm{Bern}(q(W_{i,j-1}, W_{i,j+1})), \quad m \in [n],\ m \neq i.$$

where $q(W_{i,j-1}, W_{i,j+1}) = \mathbb{P}\left(W_{j-1} = W_{i,j-1}, W_{j+1} = W_{i,j+1} \,|\, Z_i\right)$. Given a binomial random variable $B \sim \mathrm{Bin}(n, q)$, we have the following fact by elementary calculus,

$$\mathbb{E}\left[\frac{1}{1+B}\right] = \frac{1}{(n+1)q} \cdot (1 - (1-q)^{n+1}). \tag{I.32}$$

hence we can bound the term $\Pi(\mu)$ as below

$$\begin{aligned}
\Pi(\mu) &= \frac{K-1}{n}\mathbb{E}_{Z_i}\left[\mathrm{Var}_F\left(\mu(W_i)\right)\frac{1-(1-q(W_{i,j-1}, W_{i,j+1}))^n}{q(W_{i,j-1}, W_{i,j+1})}\right] && \text{(I.33)}\\
&\leq \frac{K-1}{n}\mathbb{E}_{Z_i}\left[\mathrm{Var}_F\left(\mu(W_i)\right)\right]\frac{K^2}{K^2\min\{q(k_1,k_2)\}} && \text{(I.34)}\\
&\leq \frac{K^3}{n}\frac{\mathbb{E}\left[\mu^2(X,Z)\right]}{q_0} && \text{(I.35)}
\end{aligned}$$

where the equality holds as a result of (I.31) and (I.32). And in the second line, we lower bound $q(W_{i,j-1}, W_{i,j+1})$ by $\min\{q(k_1,k_2)\}$. Assuming $K^2 \min\{\mathbb{P}\left(W_{j-1} = k_1, W_{j+1} = k_2\right)\}_{k_1,k_2\in[K]}\} \geq q_0 > 0$ gives us the third line. Then we can establish $\Pi(\mu) = O\left(\frac{K^3}{n}\right)$ (and similarly for $\Pi(\mu^\star)$) under the stated moment condition $\mathbb{E}\left[(\mu)^2(X,Z)\right], \mathbb{E}\left[(\mu^\star)^2(X,Z)\right] < \infty$. Finally, making use of the rate result about $\Pi(\mu), \Pi(\mu^\star)$ and following the same derivation as in Proposition 3.5, we have $f(\mu) - f_n^{\mathcal{T}}(\mu) = O\left(\frac{K^3}{n_2}\right)$, where we shift back to the n_2 notation. □

I.2.3 Ancillary lemmas

Lemma I.2 can be similarly derived as the expression for the Rényi divergence between two multivariate Gaussian distributions in Section 2.2.4 of Gil (2011). For completeness, we still present our proof below.

Lemma I.2. *The χ^2-divergence between $P : \mathcal{N}(\boldsymbol{a}_1, \Sigma_1)$ and $Q : \mathcal{N}(\boldsymbol{a}_2, \Sigma_2)$ equals the following whenever $2\Sigma_2 - \Sigma_1 \succ 0$:*

$$\frac{|\Sigma_2|}{|\Sigma_1|^{\frac{1}{2}}|2\Sigma_2 - \Sigma_1|^{\frac{1}{2}}}\exp\left\{(\boldsymbol{a}_1 - \boldsymbol{a}_2)^\top(2\Sigma_2 - \Sigma_1)^{-1}(\boldsymbol{a}_1 - \boldsymbol{a}_2)\right\} - 1.$$

where $\boldsymbol{a}_1, \boldsymbol{a}_2 \in \mathbb{R}^d$, $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d\times d}$, $\Sigma \succ 0$ means a matrix Σ is positive definite and $|\Sigma|$ denotes its determinant.

Proof of Lemma I.2. According to the definition of the χ^2-divergence, we have

$$\chi^2(P\|Q) := \int \left(\frac{dP}{dQ}\right)^2 dQ - 1 = \int \frac{p^2(x)}{q(x)}dx - 1, \tag{I.36}$$

where $p(x), q(x)$ are the Gaussian density functions. For multivariate Gaussian random variable with mean $\boldsymbol{a} \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$, the density function equals the following

$$f(x) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(x-\boldsymbol{a})^\top\Sigma^{-1}(x-\boldsymbol{a})\right\}, \quad x \in \mathbb{R}^d. \tag{I.37}$$

Hence we can calculate the χ^2-divergence as below,

$$\begin{aligned}\chi^2(P\|Q) &= \frac{|\Sigma_2|^{\frac{1}{2}}}{|\Sigma_1|}\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{\frac{d}{2}}}\exp\left\{-\frac{1}{2}(x-\boldsymbol{a}_1)^\top(2\Sigma_1^{-1})(x-\boldsymbol{a}_1)+\frac{1}{2}(x-\boldsymbol{a}_2)^\top\Sigma_2^{-1}(x-\boldsymbol{a}_2)\right\}dx-1\\ &:= \frac{|\Sigma_2|^{\frac{1}{2}}}{|\Sigma_1|}\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{\frac{d}{2}}}\exp\{\mathrm{II}_1+\mathrm{II}_2+\mathrm{II}_3\}\,dx-1, \qquad (\mathrm{I.38})\end{aligned}$$

where the first equality holds following the definition in (I.36) and the second equality comes from expanding the term in the exponent and combining, together with the following new notations:

$$\mathrm{II}_1 := -\frac{1}{2}x^\top(2\Sigma_1^{-1}-\Sigma_2^{-1})x \qquad (\mathrm{I.39})$$

$$\mathrm{II}_2 := -\frac{1}{2}\cdot(-2x^\top)(2\Sigma_1^{-1}\boldsymbol{a}_1-\Sigma_2^{-1}\boldsymbol{a}_2) \qquad (\mathrm{I.40})$$

$$\mathrm{II}_3 := -\frac{1}{2}(2\boldsymbol{a}_1^\top\Sigma_1^{-1}\boldsymbol{a}_1-\boldsymbol{a}_2\Sigma_2^{-1}\boldsymbol{a}_2) \qquad (\mathrm{I.41})$$

Let $\Sigma_\star^{-1}=2\Sigma_1^{-1}-\Sigma_2^{-1}$, $\Sigma_\star^{-1}\boldsymbol{a}_\star=2\Sigma_1^{-1}\boldsymbol{a}_1-\Sigma_2^{-1}\boldsymbol{a}_2$ (since we assume the positive definiteness of $2\Sigma_2-\Sigma_1$, which implies $2\Sigma_1^{-1}-\Sigma_2^{-1}\succ 0$, hence $\Sigma_\star$ and $\boldsymbol{a}_\star$ are well-defined), then we have

$$(\Sigma_1^{-1}\Sigma_\star\Sigma_2^{-1})^{-1}=\Sigma_2\Sigma_\star^{-1}\Sigma_1=2\Sigma_2-\Sigma_1 \qquad (\mathrm{I.42})$$

$$2\Sigma_\star\Sigma_1^{-1}-\mathrm{I}_d=\Sigma_\star(2\Sigma_1^{-1}-\Sigma_\star^{-1})=\Sigma_\star\Sigma_2^{-1} \qquad (\mathrm{I.43})$$

$$\begin{aligned}\frac{1}{2}\boldsymbol{a}_\star^\top\Sigma_\star^{-1}\boldsymbol{a}_\star &= \frac{1}{2}(2\Sigma_1^{-1}\boldsymbol{a}_1-\Sigma_2^{-1}\boldsymbol{a}_2)^\top\Sigma_\star(2\Sigma_1^{-1}\boldsymbol{a}_1-\Sigma_2^{-1}\boldsymbol{a}_2)\\ &= 2\boldsymbol{a}_1^\top\Sigma_1^{-1}\Sigma_\star\Sigma_1^{-1}\boldsymbol{a}_1-2\boldsymbol{a}_1^\top\Sigma_1^{-1}\Sigma_\star\Sigma_2^{-1}\boldsymbol{a}_2+\frac{1}{2}\boldsymbol{a}_2^\top\Sigma_2^{-1}\Sigma_\star\Sigma_2^{-1}\boldsymbol{a}_2\\ &= 2\boldsymbol{a}_1^\top\Sigma_1^{-1}\Sigma_\star\Sigma_1^{-1}\boldsymbol{a}_1-2\boldsymbol{a}_1^\top(2\Sigma_2-\Sigma_1)^{-1}\boldsymbol{a}_2+\frac{1}{2}\boldsymbol{a}_2^\top\Sigma_2^{-1}\Sigma_\star\Sigma_2^{-1}\boldsymbol{a}_2 \qquad (\mathrm{I.44})\end{aligned}$$

where the first and the second line hold by the definition of $\Sigma_\star$, the second equality holds since $\Sigma_\star^{-1}=\Sigma_\star^{-1}\Sigma_\star\Sigma_\star^{-1}$, the third line is simply from expanding and the last equality comes from (I.42). The above equations will be used a lot for the incoming derivations. Now the term in the exponent can be written as

$$\begin{aligned}&\mathrm{II}_1+\mathrm{II}_2+\mathrm{II}_3\\ &= -\frac{1}{2}(x^\top\Sigma_\star^{-1}x-2x^\top\Sigma_\star^{-1}\boldsymbol{a}_\star)+\mathrm{II}_3\\ &= -\frac{1}{2}(x-\boldsymbol{a}_\star)^\top\Sigma_\star^{-1}(x-\boldsymbol{a}_\star)+\frac{1}{2}\boldsymbol{a}_\star^\top\Sigma_\star^{-1}\boldsymbol{a}_\star-\frac{1}{2}(2\boldsymbol{a}_1^\top\Sigma_1^{-1}\boldsymbol{a}_1-\boldsymbol{a}_2\Sigma_2^{-1}\boldsymbol{a}_2)\\ &= \lambda(x)+\boldsymbol{a}_1^\top\Sigma_1^{-1}(2\Sigma_\star\Sigma_1^{-1}-\mathrm{I}_d)\boldsymbol{a}_1-2\boldsymbol{a}_1^\top(2\Sigma_2-\Sigma_1)^{-1}\boldsymbol{a}_2+\frac{1}{2}\boldsymbol{a}_2^\top\Sigma_2^{-1}(\Sigma_\star\Sigma_2^{-1}+\mathrm{I}_d)\boldsymbol{a}_2\\ &= \lambda(x)+\boldsymbol{a}_1^\top\Sigma_1^{-1}\Sigma_\star\Sigma_2^{-1}\boldsymbol{a}_1-2\boldsymbol{a}_1^\top(2\Sigma_2-\Sigma_1)^{-1}\boldsymbol{a}_2+\boldsymbol{a}_2^\top\Sigma_2^{-1}\Sigma_\star\Sigma_1^{-1}\boldsymbol{a}_2\\ &= \lambda(x)+\boldsymbol{a}_1^\top(2\Sigma_2-\Sigma_1)^{-1}\boldsymbol{a}_1-2\boldsymbol{a}_1^\top(2\Sigma_2-\Sigma_1)^{-1}\boldsymbol{a}_2+\boldsymbol{a}_2^\top(2\Sigma_2-\Sigma_1)^{-1}\boldsymbol{a}_2\\ &= \lambda(x)+(\boldsymbol{a}_1-\boldsymbol{a}_2)^\top(2\Sigma_2-\Sigma_1)^{-1}(\boldsymbol{a}_1-\boldsymbol{a}_2):=\lambda(x)+Q(\boldsymbol{a}_1,\boldsymbol{a}_2,\Sigma_1,\Sigma_2) \qquad (\mathrm{I.45})\end{aligned}$$

where the first equality holds by the definition of $\Sigma_\star$, $\boldsymbol{a}_\star$ and (I.39), (I.40), and the second equality holds due to (I.41). Regarding the third equality, we denote the term which depends on x by $\lambda(x):=-\frac{1}{2}(x-\boldsymbol{a}_\star)^\top\Sigma_\star^{-1}(x-\boldsymbol{a}_\star)$. As for the other constant terms in the third line, we simply combine (I.44) with the expansion of the term II_3 and rearrange them into three terms: $\boldsymbol{a}_1^\top(\cdot)\boldsymbol{a}_1$, $\boldsymbol{a}_1^\top(\cdot)\boldsymbol{a}_2$ and $\boldsymbol{a}_2^\top(\cdot)\boldsymbol{a}_2$. The fourth equality holds as a result of applying (I.43) twice and the last equality is simply from rearranging.

Since only the term $\lambda(x)$ depends on x, we can simplify the χ^2-divergence into the following

$$\begin{aligned}
\chi^2(P\|Q) &= \frac{|\Sigma_2|^{\frac{1}{2}}}{|\Sigma_1|}\exp\left\{Q(\boldsymbol{a}_1,\boldsymbol{a}_2,\Sigma_1,\Sigma_2)\right\}\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{\frac{d}{2}}}\exp\left\{\lambda(x)\right\}dx - 1\\
&= \frac{|\Sigma_2|^{\frac{1}{2}}}{|\Sigma_1|}\exp\left\{Q(\boldsymbol{a}_1,\boldsymbol{a}_2,\Sigma_1,\Sigma_2)\right\}\int_{\mathbb{R}^d}\frac{|\Sigma_\star|^{\frac{1}{2}}}{(2\pi)^{\frac{d}{2}}|\Sigma_\star|^{\frac{1}{2}}}\exp\left\{\lambda(x)\right\}dx - 1\\
&= \frac{|\Sigma_2|^{\frac{1}{2}}}{|\Sigma_1|}|\Sigma_\star|^{\frac{1}{2}}\exp\left\{Q(\boldsymbol{a}_1,\boldsymbol{a}_2,\Sigma_1,\Sigma_2)\right\} - 1\\
&= \frac{|\Sigma_2|}{|\Sigma_1|^{\frac{1}{2}}}|\Sigma_1^{-1}\Sigma_\star\Sigma_2^{-1}|^{\frac{1}{2}}\exp\left\{Q(\boldsymbol{a}_1,\boldsymbol{a}_2,\Sigma_1,\Sigma_2)\right\} - 1\\
&= \frac{|\Sigma_2|}{|\Sigma_1|^{\frac{1}{2}}|2\Sigma_2-\Sigma_1|^{\frac{1}{2}}}\exp\left\{(\boldsymbol{a}_1-\boldsymbol{a}_2)^\top(2\Sigma_2-\Sigma_1)^{-1}(\boldsymbol{a}_1-\boldsymbol{a}_2)\right\} - 1
\end{aligned}$$

where the first equality comes from (I.38) and (I.45), the third equality holds due to the definition of $\lambda(x)$ and the fact that $\int f(x)dx = 1$, where $f(x)$ is the Gaussian density function with the mean $\boldsymbol{a}_\star$ and covariance matrix $\Sigma_\star$), the fourth equality holds by making use of the properties of determinant and the last equality holds as a result of (I.42). □

J Further simulation details

Source code for conducting floodgate in our simulation studies can be found at `https://github.com/LuZhangH/floodgate`.

J.1 Nonlinear model setup

Consider W which follows a Gaussian copula distribution with $X = W_{j_0}, Z = W_{\text{-}j_0}$ for some j_0 $(1 \le j_0 \le p)$, i.e.,

$$W^{\text{latent}} \sim AR(1),\ W_j = 2\varphi(X_j^{\text{latent}}) - 1,\ \ \forall\ 1 \le j \le p. \tag{J.1}$$

Hence the marginal distribution for W_j is Unif$[-1, 1]$ (in fact, these are the inputs to the fitting methods we use in floodgate, not the AR(1) latent variables W^{latent}). We consider the following conditional model for Y given W, with standard Gaussian noise,

$$\mu^\star(x,z) = \mu^\star(w) := \sum_{j\in S^1} g_j(w_j) + \sum_{(j,l)\in S^2} g_j(w_j)g_l(w_l) + \sum_{(j,l,m)\in S^3} g_j(w_j)g_l(w_l)g_m(w_m) \tag{J.2}$$

where each function $g_j(x)$ is randomly chosen from the following:

$$\sin(\pi x), \cos(\pi x), \sin(\pi x/2), \cos(\pi x)I(x>0), x\sin(\pi x), x, |x|, x^2, x^3, \exp(x)-1. \tag{J.3}$$

S^1 basically contains the main effect terms, while S^2 contain the pairs of variables with first order interactions. Tuples of variables involving second order interaction are denoted by S^3. For a given amplitude, (J.2) is scaled by the amplitude value divided by $\sqrt{n}$.

Now we describe the construction of S^1, S^2, S^3. First we randomly pick 30 variables into $S_\star$ and initialize $S_{\text{wl}} = S_\star$. 15 of them will be randomly assigned into S^1 and removed from S_{wl}. Among these 15 variables in S^1, we further choose 10 variables into 5 pairs randomly, which will be included in S^2. Regarding the other pairs in S^2, each time we randomly pick 2 variables from $S_\star$ with the unscaled weight being $2|S_{\text{wl}}|/|S_\star|$ for variables in S_{wl}, $|S_\star \setminus S_{\text{wl}}|/|S_\star|$ for the others, then add them as a pair into S^2. Once picked, the variables will be removed from S_{wl}. This process iterates until $|S_{\text{wl}}| \le 5$. Regarding the construction of S^3, each time we randomly pick 3 variables from $S_\star$ with the unscaled weight being $1.5|S_{\text{wl}}|/|S_\star|$ for variables in S_{wl}, $|S_\star \setminus S_{\text{wl}}|/|S_\star|$ for the others, then add them as a tuple into S^3. Once picked, the variables will be removed from S_{wl}. This process iterates until $|S_{\text{wl}}| = 0$.

J.2 Implementation details of fitting algorithms

Regarding how to obtain the working regression function, there will be four different fitting algorithms for non-binary responses:

- *LASSO*: We fit a linear model by 10-fold cross-validated LASSO and output a working regression function. The subsequent inference step will be quite fast. First, as implied by Algorithm 1, $L_n^\alpha(\mu)$ will be set to zero for unselected variables, without any computation. Second, as alluded to in Section 2.4, we can analytically compute the conditional quantities in Algorithm 1.
- *Ridge*: We again use 10-fold cross-validation to choose the penalty parameter for Ridge regression. It is also fast to perform floodgate on, due to the second point mentioned above.
- *SAM*: We consider additive modelling, for example the sparse additive models (SAM) proposed in Ravikumar et al. (2009). As suggested by the name, it carries out sparse penalization and our method will assign $L_n^\alpha(\mu) = 0$ to unselected variables, as in *lasso*.
- *Random Forest*: Random forest (Breiman, 2001) is included as a purely nonlinear machine learning algorithm. While random forest do not generally conduct variable selection, we rank variables based on the heuristic importance measure and use the top 50 variables to run Algorithm 1 and set $L_n^\alpha(\mu) = 0$ for the remaining ones. Remark this is only for the concern of speed and does not have any negative impact on the inferential validity.

There are two additional fitting algorithms for binary responses: logistic regression with L1 regularization and L2 regularization, denoted by *Binom_LASSO* and *Binom_Ridge* respectively. Both use 10-fold cross-validation to choose the penalty parameter.

J.3 Implementation details of ordinary least squares

When the conditional model of $Y \mid X, Z$ is linear, i.e., $\mathbb{E}[Y \mid X, Z] = X\beta + Z\theta$ with $(\beta, \theta) \in \mathbb{R}^p$ the coefficients, the mMSE gap for X is closely related to its linear coefficient, formally

$$\mathcal{I} = |\beta| \sqrt{\mathbb{E}[\mathrm{Var}(X \mid Z)]}.$$

When the sample size n is greater than the number of variables p, ordinary least squares (OLS) can provide valid confidence intervals for β. However, there does not seem to exist a non-conservative way to transform the OLS confidence interval for β into a confidence bound for $|\beta|$. So instead, we provide OLS with further oracle information: the sign of β (we only compare half-widths of non-null covariates, and hence never construct OLS LCBs when $\beta = 0$). In particular, if [LCI, UCI] denotes a standard OLS 2-sided, equal-tailed $1 - 2\alpha$ confidence interval for β, then the OLS LCB for $\mathcal{I}$ we use is

$$\mathrm{LCB}_{\mathrm{OLS}} = \begin{cases} \mathrm{LCI}\sqrt{\mathbb{E}[\mathrm{Var}(X \mid Z)]} & \text{if } \beta > 0 \\ -\mathrm{UCI}\sqrt{\mathbb{E}[\mathrm{Var}(X \mid Z)]} & \text{if } \beta < 0 \end{cases} \tag{J.4}$$

which guarantees exact $1 - \alpha$ coverage of $\mathcal{I}$ for any nonzero value of β. We again emphasize that, in order to construct this interval, OLS uses the oracle information of the sign of β (this information is not available to floodgate in our simulations).

J.4 Plots deferred from the main paper

J.4.1 Effect of sample splitting proportion

The corresponding coverage plots of Figure 1 are given in Figure 8. Figures 9 and 10 are additional plots with different simulation parameters specified in the captions. Figures 8 and 10 show that in the simulations in Section 4.2, the coverage of floodgate is consistently at or above the nominal 95% level.

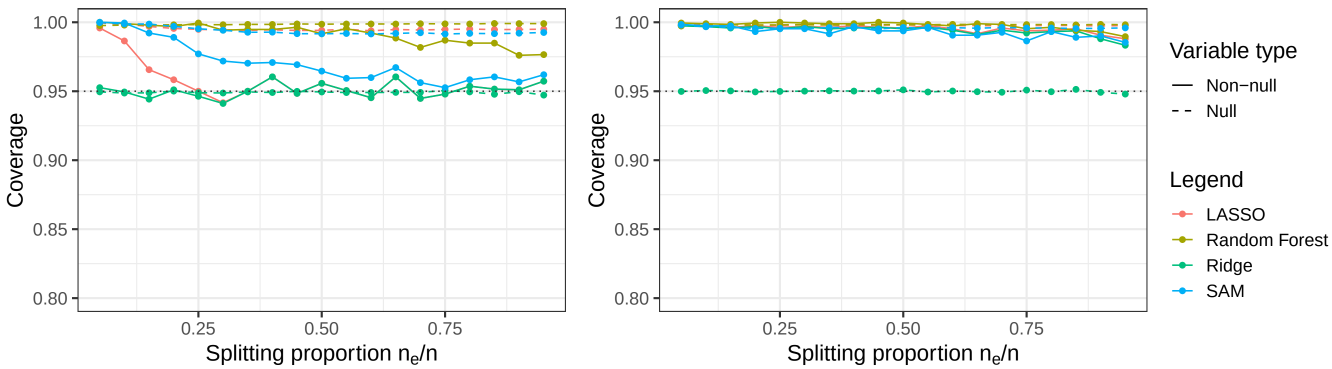

Figure 8: Coverage for the the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 4.2. The coefficient amplitude is chosen to be 10 for the left panel and the sample size n equals 3000 in the right panel; see Section 4.1 for remaining details. Standard errors are below 0.007 (left) and 0.003 (right).

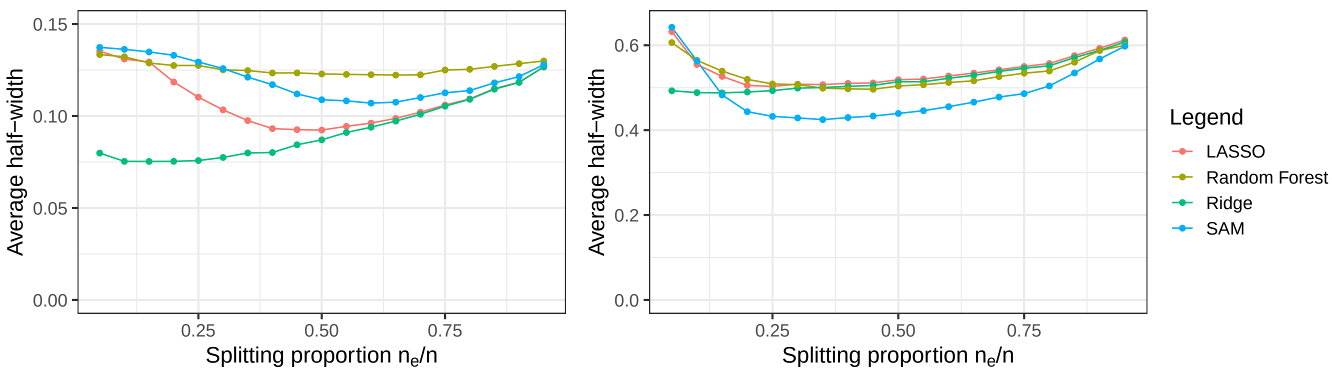

Figure 9: Average half-widths for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 4.2. The coefficient amplitude is chosen to be 5 for the left panel and the sample size n equals 1000 in the right panel; see Section 4.1 for remaining details. Standard errors are below 0.002 (left) and 0.01 (right).

J.4.2 Effect of covariate dimension

The corresponding coverage plots of Figure 2 are given in Figure 11. Figures 12 and 13 are additional plots with different simulation parameters specified in the captions. Figures 11 and 13 show that in these simulations, the coverage of floodgate is consistently at or above the nominal 95% level.

J.4.3 Comparison with Williamson et al. (2020)

The corresponding coverage plot of Figure 3 is given in Figure 14, where we see both methods have coverages above the nominal level. In addition to the example in Section 4.4, we also compare floodgate with W20b in the higher-dimensional setting of the left panel of Figure 2. Due to the computational challenge of running Williamson et al. (2020)'s method, we only consider the two most efficient algorithms (LASSO and Ridge) among the four described in Appendix J.2. Figure 15 shows W20b to have slightly less consistent coverage than floodgate, but also reinforces the general picture from the lower-dimensional simulation in Section 4.4 that W20b's LCBs are quite close to zero compared with floodgate's.

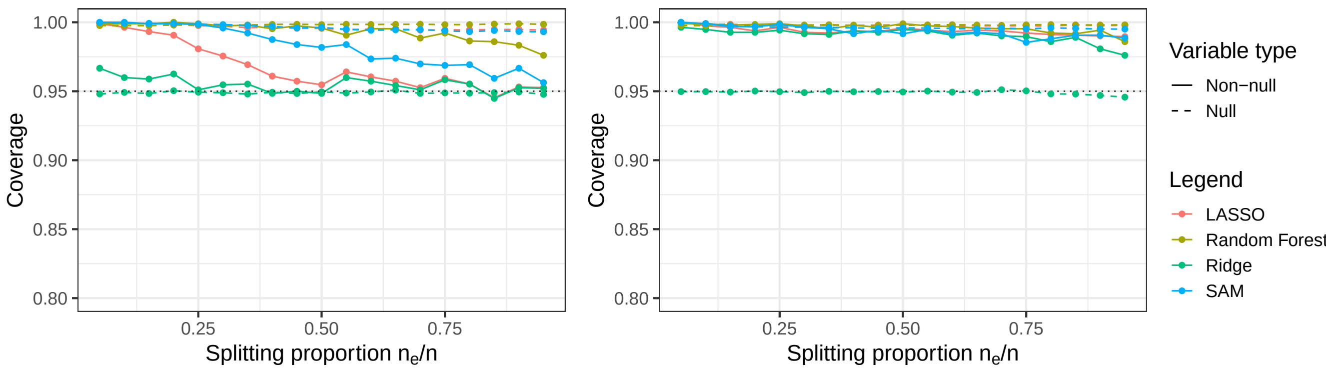

Figure 10: Coverage for the the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 4.2. The coefficient amplitude is chosen to be 5 for the left panel and the sample size n equals 1000 in the right panel; see Section 4.1 for remaining details. Standard errors are below 0.006 (left) and 0.004 (right).

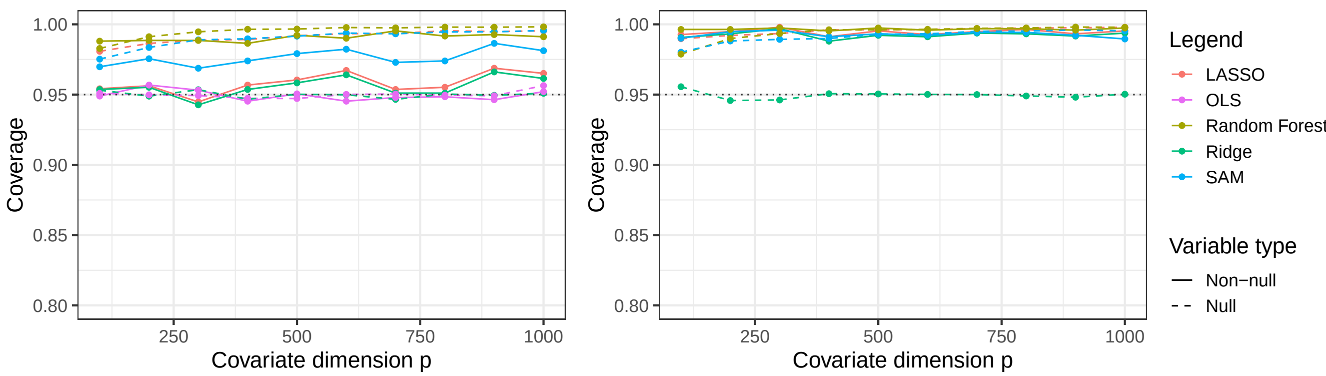

Figure 11: Coverage for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 4.3. OLS is run on the full sample. p is varied on the x-axis; see Section 4.1 for remaining details. Standard errors are below 0.006 (left) and 0.004 (right).

J.4.4 Robustness

Figure 16 studies the robustness of floodgate for a nonlinear $\mu^\star$. We see the coverage being rather conservative for the non-null variables, reflecting the coverage-protective gap between $f(\mu)$ and $f(\mu^\star) = \mathcal{I}$. Figure 17 shows that in the simulations of linear models and nonlinear models, the average half-width of floodgate is robust to estimation error in $P_{X|Z}$.

J.4.5 Co-sufficient floodgate

In this section, we demonstrate the performance of co-sufficient floodgate in a linear setting. Figure 18 tells a similar story as Figure 6 in Section 4.7. Note that despite the linearity of the true model in Figure 18, the LASSO performs poorly because the true model is quite dense (30 of the 50 covariates are non-null), which also explains why ridge regression performs so well.

J.4.6 Effect of covariate dependence

In Figure 19, we vary the covariate autocorrelation coefficient and plot the average half-widths of floodgate LCBs of non-null covariates under distributions with the linear (left panel) and the nonlinear (right panel) $\mu^\star$ described in Section 4.1, respectively. The left panel of Figure 19 also includes a curve for OLS. Since

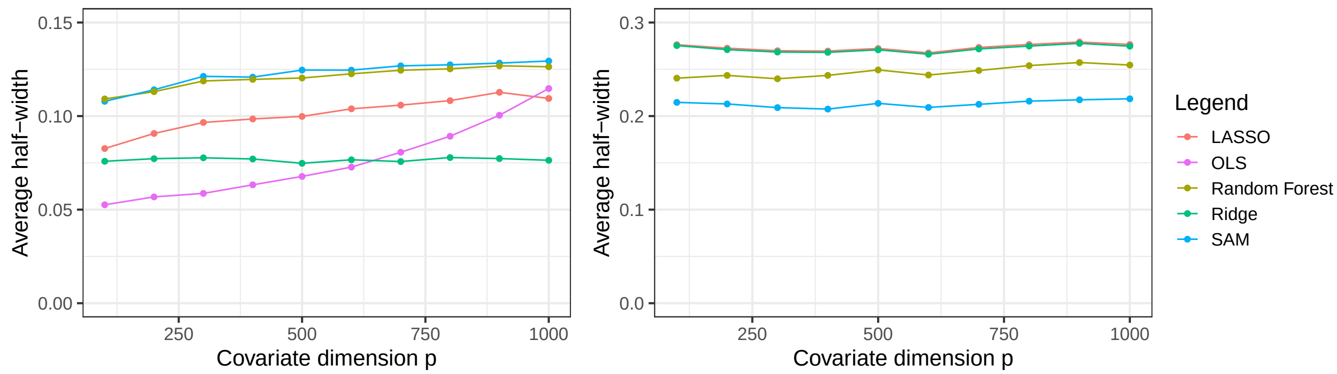

Figure 12: Average half-widths for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 4.3. The splitting proportion is chosen to be 0.25 for the left panel and the sample size n equals 3000 in the right panel. p is varied on the x-axis; see Section 4.1 for remaining details. Standard errors are below 0.002 (left) and 0.005 (right).

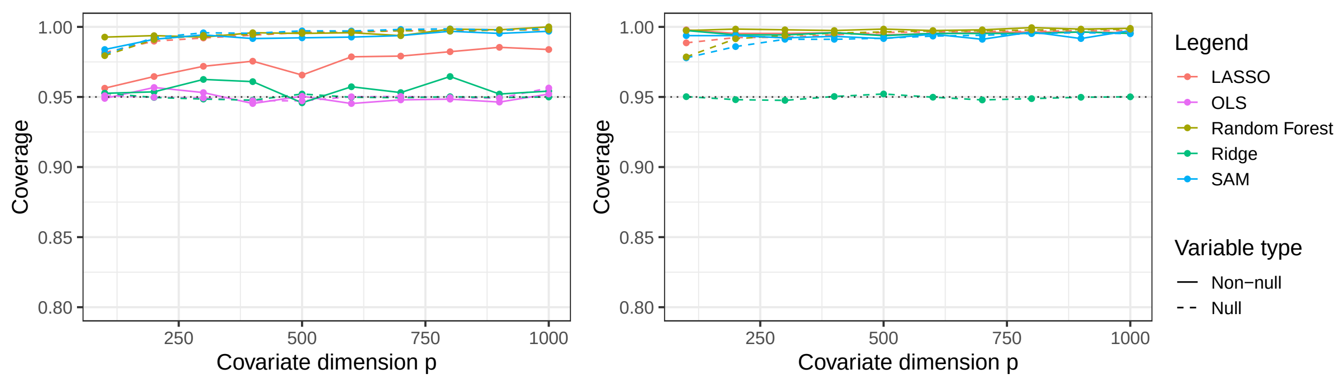

Figure 13: Coverage for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 4.3. The splitting proportion is chosen to be 0.25 for the left panel and the sample size n equals 3000 in the right panel. p is varied on the x-axis; see Section 4.1 for remaining details. Standard errors are below 0.006 (left) and 0.004 (right).

$\mathcal{I}$ in a linear model is proportional to $\sqrt{\mathbb{E}\left[\mathrm{Var}\left(X \mid Z\right)\right]}$ which varies with the autocorrelation coefficient, we divided the half-widths in Figure 19 by this quantity to make it easier to compare values across the x-axis. The main takeaway is that the effect of covariate dependence on floodgate is somewhat mild until the dependence gets very large (> 0.5 correlation). This behavior is intuitive, and indeed we see a parallel trend in the curves for OLS inference in Figure 19. The corresponding coverage plots of Figure 19 are given in Figure 20. Figures 21 and 22 are additional plots with a different covariate dimension specified in the captions. Figures 20 and 22 show that the coverage of floodgate is consistently at or above the nominal 95% level.

J.4.7 Effect of sample size

In Figures 23 and 24, we vary the sample size and plot the coverages and average half-widths of floodgate LCBs of non-null covariates under distributions with the linear and the nonlinear $\mu^\star$ described in Section 4.1, respectively. The main takeaway is that the accuracy of floodgate depends heavily on sample size. Note that in these plots, the signal size is scaled down by the square root of the sample size, so the *selection* problem is roughly getting no easier as the sample size increases, but we still see that floodgate

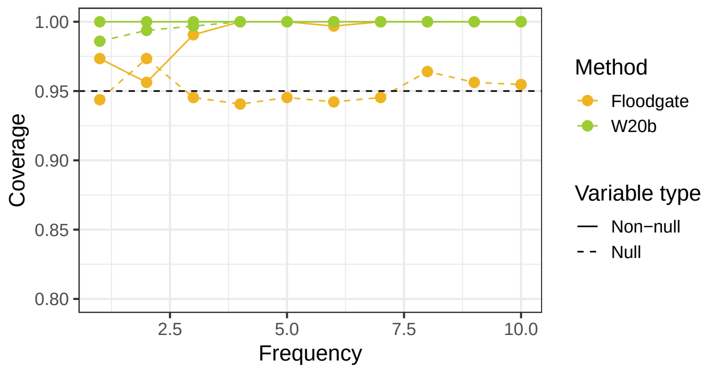

Figure 14: Coverage for floodgate and W20b in the sine function simulation of Section 4.4. The frequency λ is varied on the x-axis, and the dotted black line in the plot shows the nominal coverage level $1-\alpha$. The results are averaged over 640 independent replicates, and the standard errors are below 0.006.

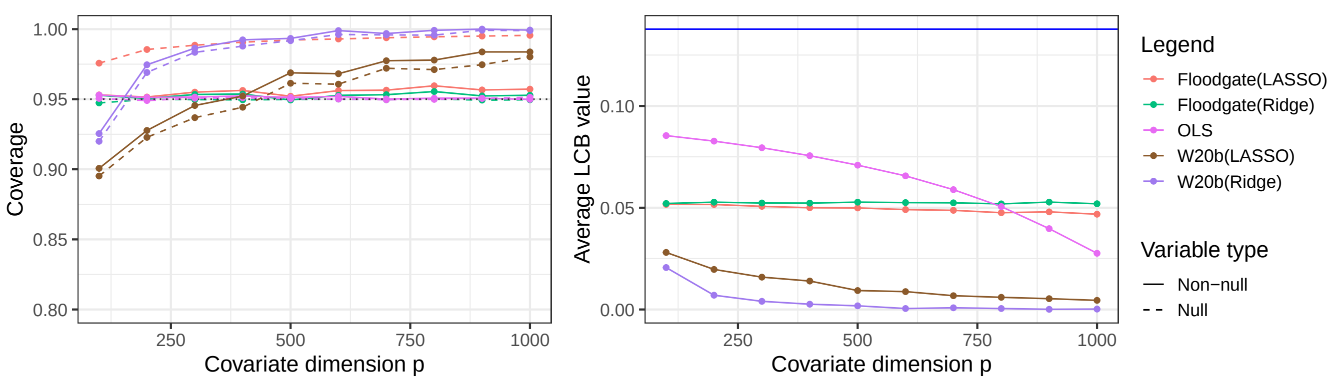

Figure 15: Coverage (left) and average LCB values (right) for floodgate, W20b, and OLS (run on the full sample) in the linear-$\mu^\star$ simulation of Section 4.4. p is varied on the x-axis, and the solid blue line in the right-hand plot shows the value of $\mathcal{I}$; see Section 4.1 for remaining details. The results are averaged over 640 independent replicates, and the standard errors are below 0.012 (left) and 0.004 (right).

can achieve much more accurate inference for larger sample sizes.

K Implementation details of genomics application

As mentioned in Section 2.6, the floodgate approach can be immediately generalized to conduct inference on the importance of a group of variables. This is practically useful in our application to the genomic data, where we group nearby SNPs whose effects are usually found challenging to be distinguished. Specifically, we use the exact same grouping at the same seven resolutions as Sesia et al. (2020b).

Regarding the genotype modelling, we consider the hidden Markov models (HMM) (Scheet and Stephens, 2006), as used in Sesia et al. (2019, 2020b), which provides a good description of the linkage disequilibrium (LD) structure. We obtain the fitted HMM parameters from Sesia et al. (2020b) on the UK Biobank data. Since HMM does not offer simple closed form expressions of the conditional quantities in Algorithm 1, we generate null copies of the genotypes and use them for the Monte Carlo analogue of floodgate. Below we simply describe the generating procedure. Under the HMM, we denote the covariates by W (genotypes or haplotypes) and the unobserved hidden states (local ancestries) by A, with the joint distribution over W denoted by P_W, the joint distribution over A denoted by P_A, which is the latent Markov chain model. For a given contiguous group of variables g_j, we can sample the null copy of W_{g_j} as follows:

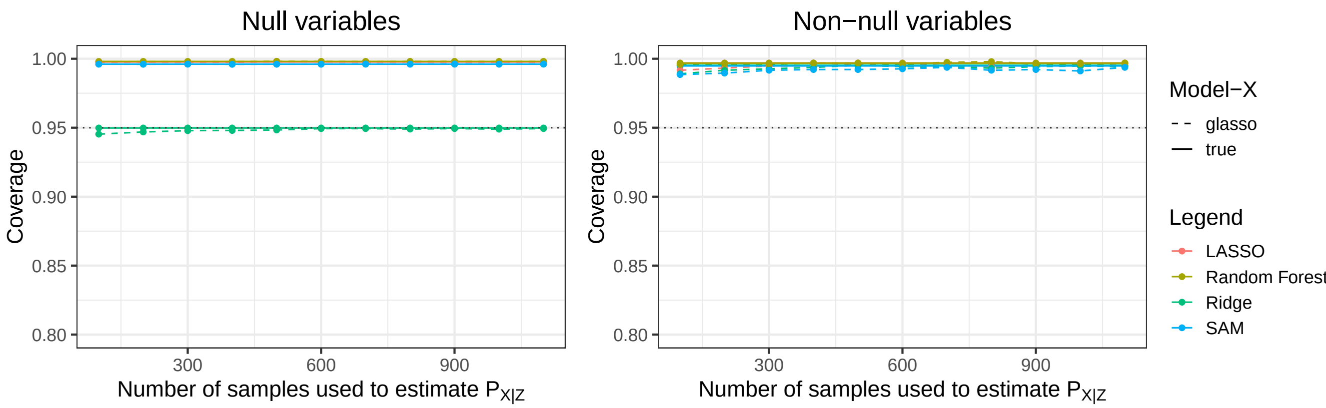

Figure 16: Coverage of null (left) and non-null (right) covariates when the covariate distribution is estimated in-sample for the nonlinear-$\mu^\star$ simulations of Section 4.5. See Section 4.1 for remaining details. Standard errors are below 0.001 (left) and 0.003 (right).

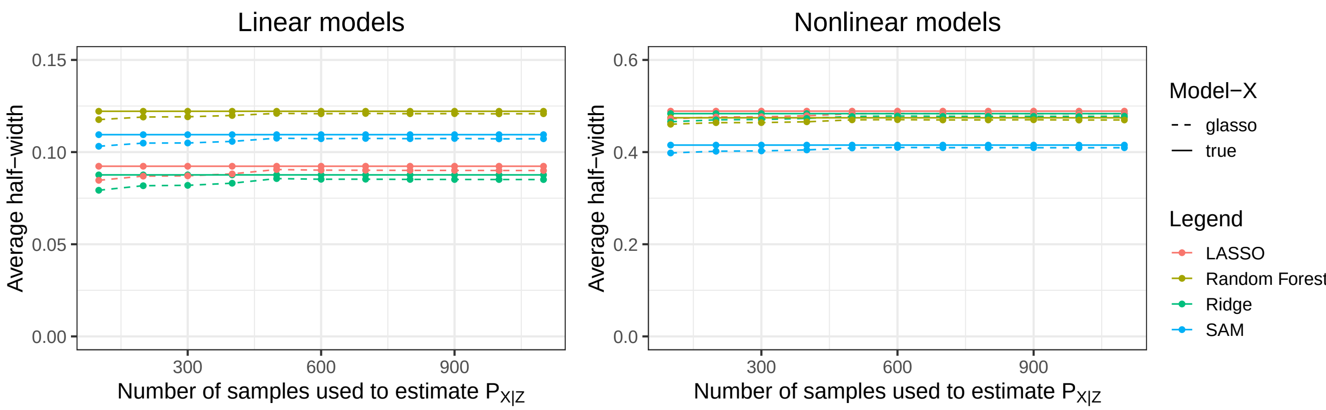

Figure 17: Half-width plot of non-null covariates when the covariate distribution is estimated in-sample for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section 4.5. See Section 4.1 for remaining details. Standard errors are below 0.002 (left) and 0.007 (right).

(1) Marginalize out W_{g_j} and recompute the parameters of the new HMM $P_{\text{-}g_j}$ over $W_{\text{-}g_j}$.

(2) Sample the hidden states $A_{\text{-}g_j}$ by applying the forward-backward algorithm to $W_{\text{-}g_j}$, with the new HMM $P_{\text{-}g_j}$.

(3) Given $A_{\text{-}g_j}$, sample A_{g_j} according to the latent Markov chain model P_A.

(4) Sample $\tilde{W}_{g_j}$ given A_{g_j} according to the emission distribution of the group g_j in the model of P_W.

To see why the above procedure produces a valid null copy of W_{g_j}, consider the following joint distribution, conditioning on $W_{\text{-}g_j}$

$$P_{\text{joint}} : (W_{g_j}, A_{g_j}, A_{\text{-}g_j}) \mid W_{\text{-}g_j}$$

If we sample $(\tilde{W}_{g_j}, A_{g_j}, A_{\text{-}g_j})$ from the above joint conditional distribution, without looking at W_{g_j} or Y, then $\tilde{W}_{g_j}$ has the same conditional distribution as W_{g_j}, given $W_{\text{-}g_j}$ and is conditionally independent from (W_{g_j}, Y), and thus is a valid null copy of W_{g_j}. Regarding how to sample from P_{joint}, we take advantage of the HMM structure and sample $A_{\text{-}g_j}, A_{g_j}, \tilde{W}_{g_j}$ sequentially since

$$A_{g_j} \mid A_{\text{-}g_j}, W_{\text{-}g_j} \overset{d}{=} A_{g_j} \mid A_{\text{-}g_j}, \tag{K.1}$$

$$W_{g_j} \mid A_{g_j}, A_{\text{-}g_j}, W_{\text{-}g_j} \overset{d}{=} W_{g_j} \mid A_{g_j}. \tag{K.2}$$

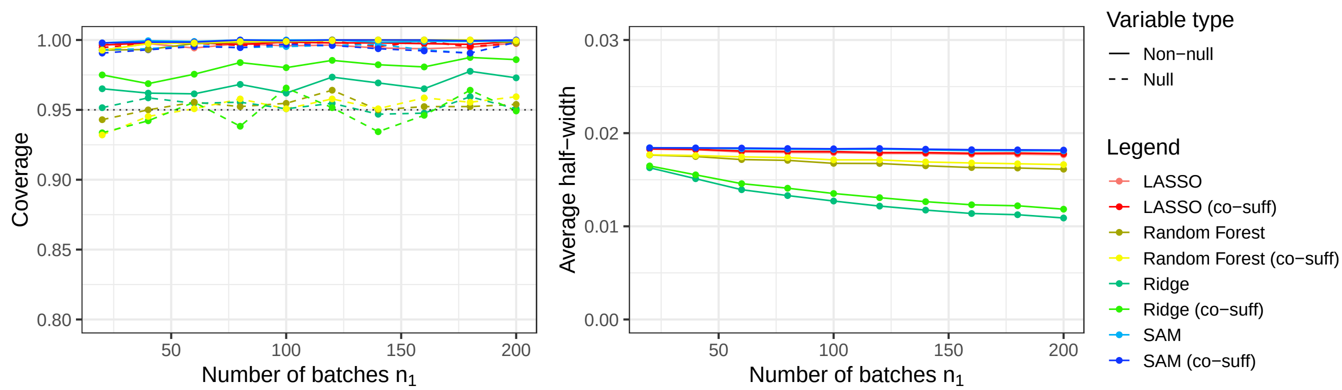

Figure 18: Coverage (left) and average half-widths (right) for co-sufficient floodgate and original floodgate in the linear-$\mu^\star$ simulations. The number of batches n_1 is varied over the x-axis. See Section 4.1 and 4.7 for remaining details. Standard errors are below 0.008 (left) and 0.001 (right).

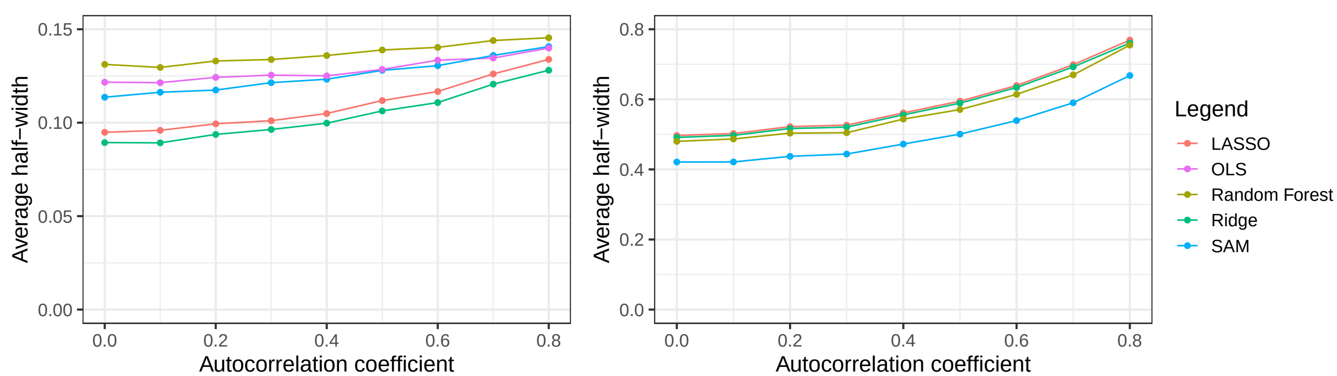

Figure 19: Average half-widths for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section J.4.6. The covariate dimension $p = 1000$ and the covariate autocorrelation coefficient is varied on the x-axis; see Section 4.1 for remaining details. Standard errors are below 0.002 (left) and 0.009 (right).

Sampling from $A_{\text{-}g_j} \mid W_{\text{-}g_j}$ is feasible since $P_{\text{-}g_j}$ is still a HMM whenever the group g_j is contiguous. Under the HMM with particular parameterization in Scheet and Stephens (2006), the cost of the forward-backward algorithm can be reduced, see Sesia et al. (2020b) for more details. We remark that marginalizing out W_{g_j} only changes the transition structure around the group g_j and the special parameterization over other variables is still beneficial in terms of the computation cost. Sampling of A_{g_j} and $\tilde{W}_{g_j}$ is computationally cheap due to (K.1) and (K.2). For a given number of null copies K, we will repeat the steps (2)-(4) for K times. But we remark the involving sampling probabilities only have to be computed once.

Regarding the quality control and data prepossessing of the UK Biobank data, we follow the Neale Lab GWAS with application 31063; details can be found on `http://www.nealelab.is/uk-biobank`. A few subjects withdrew consent and are removed from the analysis. Our final data set consisted of $361,128$ unrelated subjects and $591,513$ SNPs along 22 chromosomes.

For the platelet count phenotype, the analysis by Sesia et al. (2020b) makes several selections over the whole genome at seven different resolution levels. We focus on chromosome 12 and look at 248 selected groups from their analysis. For a given group of variables, we generate $K = 5$ null copies following the null copy generation procedure described above.

We applied floodgate with a 50-50 data split and fitted μ to the first half using the cross-validated LASSO as in Sesia et al. (2020b) and included both genotypes (SNPs from chromosomes 1–22) and the

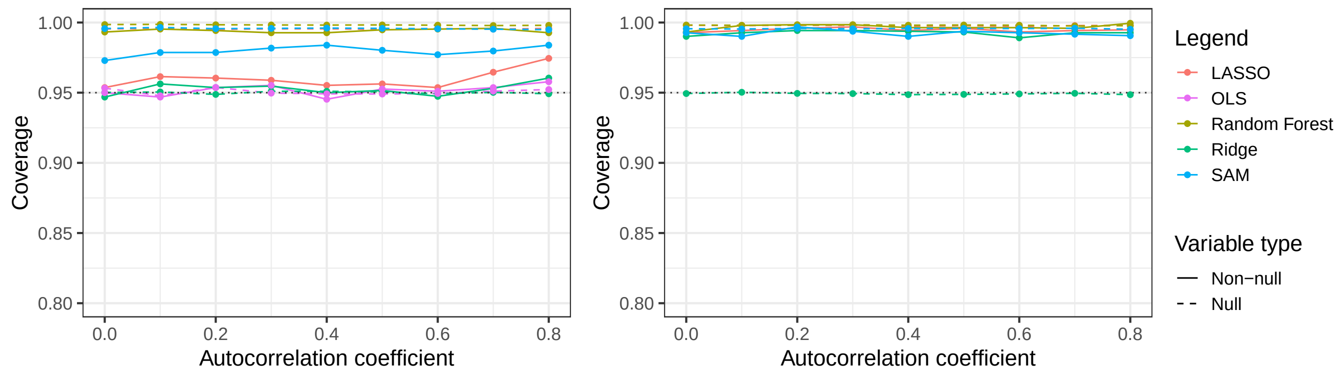

Figure 20: Coverage for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section J.4.6. The covariate dimension $p = 1000$ and the covariate autocorrelation coefficient is varied on the x-axis; see Section 4.1 for remaining details. Standard errors are below 0.006 (left) and 0.003 (right).

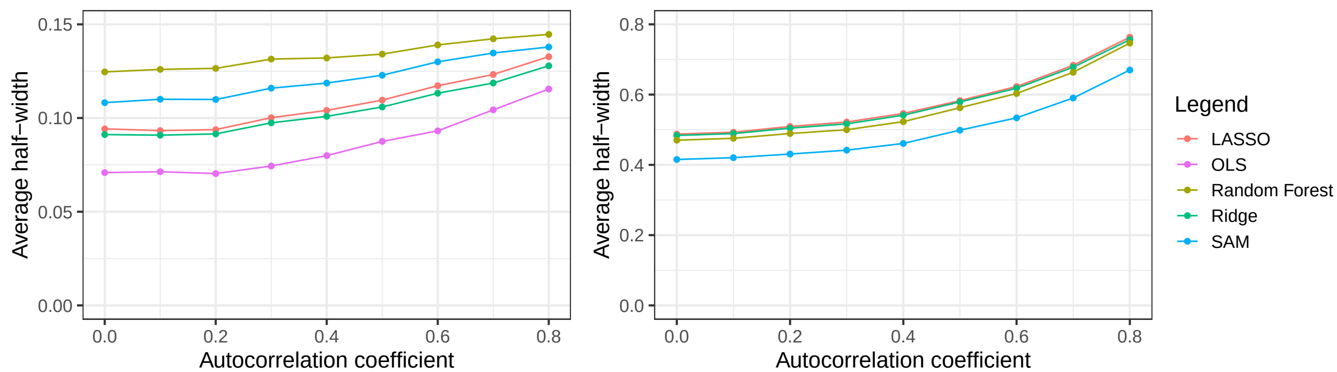

Figure 21: Average half-widths for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section J.4.6. The covariate dimension $p = 500$ and the covariate autocorrelation coefficient is varied on the x-axis; see Section 4.1 for remaining details. Standard errors are below 0.002 (left) and 0.01(right).

non-genetic variables sex, age and squared age. We centered Y by its sample mean from the first half of the data (the half used to fit μ) before applying floodgate. Although this changes nothing in theory, it does improve robustness as small biases in $\mu(X_i, Z_i) - \mathbb{E}\left[\mu(X_i, Z_i) \mid Z_i\right]$ would otherwise get multiplied by Y_i's mean in the computation of R_i in Algorithm 1.

Although our fitting of a linear model in no way changes the validity of floodgate's inference of the completely model-free mMSE gap, it does desensitize the LCB itself to the nonlinearities and interactions that partially motivated $\mathcal{I}$ as an object of inference in the first place. Our reasoning is purely pragmatic: as the universe of nonlinearities/interactions is exponentially larger than that of linear models, fitting such models requires either very strong nonlinear/interaction effects or prior knowledge of a curated set of likely nonlinearities/interactions. It is our understanding that nearly all genetic effects, linear and nonlinear/interaction alike, tend to be relatively weak, and the authors are not geneticists by training and thus lack the domain knowledge necessary to leverage the full flexibility of floodgate. Although we were already able to find substantial heritability for many blocks of SNPs with our default choice of the LASSO, it is our sincere hope and expectation that geneticists who specialize in the study of platelet count or similar traits would be able to find even more heritability using floodgate.

We report LCBs for all blocks simultaneously, although computationally we only actually run floodgate on those selected by Sesia et al. (2020b). Although their selection used all of the data (including the data

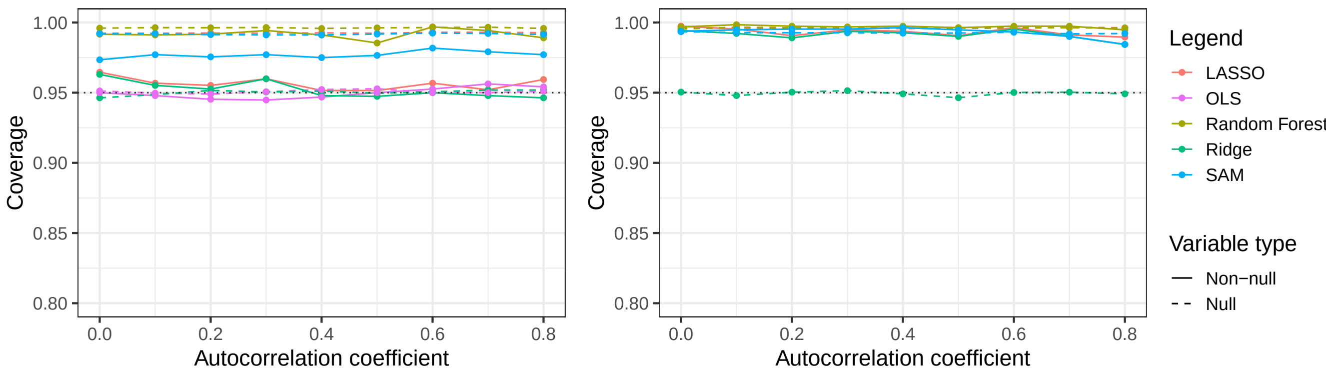

Figure 22: Coverage for the linear-$\mu^\star$ (left) and nonlinear-$\mu^\star$ (right) simulations of Section J.4.6. The covariate dimension $p = 500$ and the covariate autocorrelation coefficient is varied on the x-axis; see Section 4.1 for remaining details. Standard errors are below 0.007 (left) and 0.004 (right).

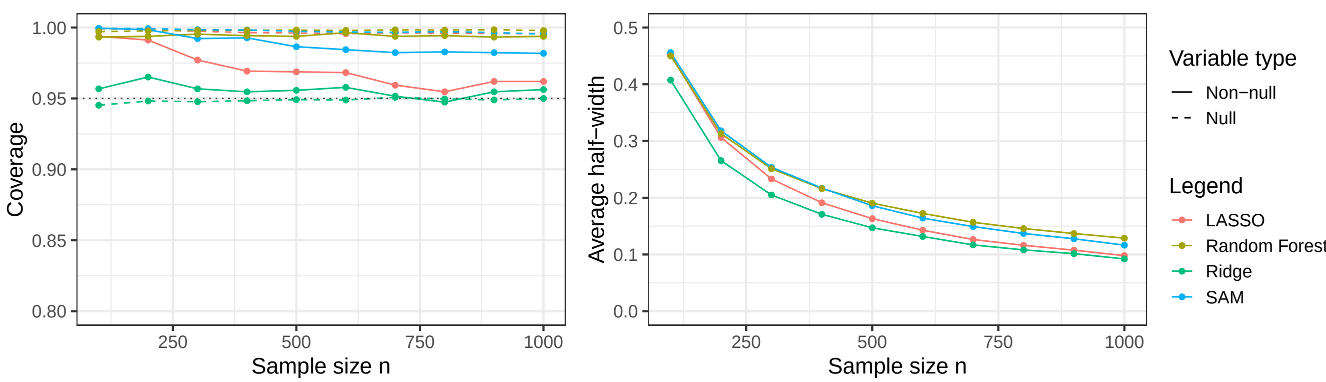

Figure 23: Coverage (left) and average half-widths (right) for the linear-$\mu^\star$ simulations of Section J.4.7. The sample size n is varied on the x-axis; see Section 4.1 for remaining details. Standard errors are below 0.007 (left) and 0.003 (right).

we used for floodgate), it does not affect the marginal validity of the LCBs we report, as explained in the last paragraph of Section 2.6.

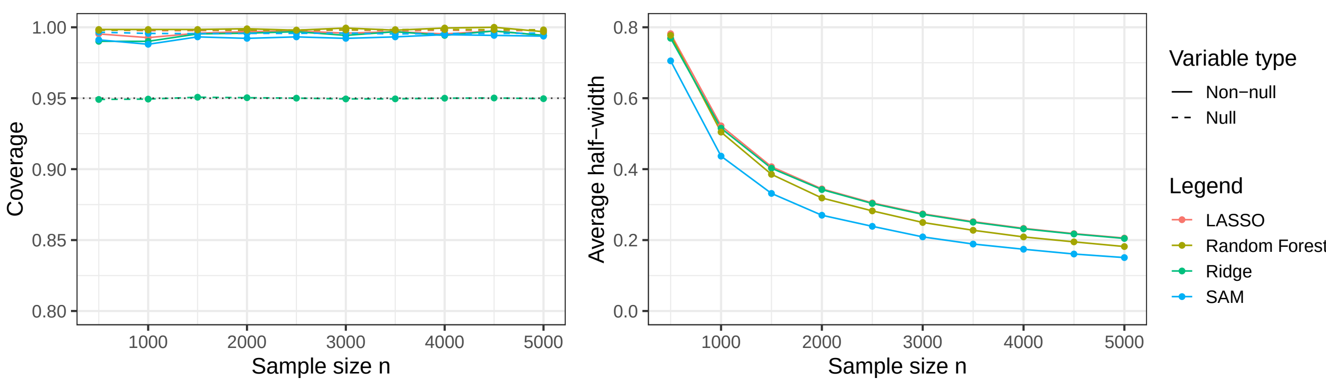

Figure 24: Coverage (left) and average half-widths (right) for the nonlinear-$\mu^\star$ simulations of Section J.4.7. The sample size n is varied on the x-axis; see Section 4.1 for remaining details. Standard errors are below 0.004 (left) and 0.011 (right).