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 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. In addition to proving floodgate's asymptotic validity, we rigorously quantify its accuracy (distance from confidence bound to estimand) and robustness. 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 $Z = (Z_1, \ldots, Z_{p-1})$ 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 $\mathbb{E}\left[Y \mid X, Z\right] \stackrel{a.s.}{=} \mathbb{E}\left[Y \mid Z\right]$, 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 function of *only* the joint distribution of (Y, X, Z), and is exactly zero when $Y \perp\!\!\!\perp X \mid Z$.

(Validity) Floodgate is asymptotically valid under extremely mild moment conditions, and in particular requires no smoothness, sparsity, or other constraints on $\mathbb{E}\left[Y \mid X, Z\right]$ that would ensure its learnability at *any* geometric rate. Floodgate requires the user to know the distribution of $X \mid Z$, although we prove this requirement can sometimes be relaxed to only knowing a model for $X \mid Z$, and we theoretically and numerically characterize floodgate's robustness to misspecification of this distribution.

(Accuracy) Floodgate derives accuracy from flexibility by allowing the user to estimate $\mathbb{E}\left[Y \mid X, Z\right]$ in whatever way they like, and we prove that the accuracy of inference is directly related 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 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 address computational considerations (Section 2.3), theoretically characterize the width of floodgate's confidence bounds (Section 2.4) and its robustness to model misspecification (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).

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

The standard approach to 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., 2016; 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 $\mathbb{E}\left[\mathrm{Cov}\left(Y, X \mid Z\right)\right]$ (see, for example, Robins et al. (2008, 2009); Li et al. (2011); Robins et al. (2017); Newey and Robins (2018); Shah and Peters (2018); Chernozhukov et al. (2018); Liu et al. (2019); Katsevich and Ramdas (2020)), which represents a generalization of the linear coefficient in a *partially* linear model. $\mathbb{E}\left[\mathrm{Cov}\left(Y, X \mid Z\right)\right]$ 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., 2018; 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.

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. However the authors only propose a consistent estimator for their MOVI and do not provide a method for inference (confidence lower- or upper-bounds).

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. To the best of our knowledge only Williamson et al. (2017) considers inference for the mMSE gap, but in order to guarantee asymptotic coverage of their confidence intervals, their theory assumes (a) the mMSE gap is strictly positive, and (b) a machine learning method is applied that converges to $\mathbb{E}\left[Y \mid X, Z\right]$ at a $o_p(n^{-1/4})$ rate. Our inference is valid for both zero and nonzero values of the mMSE gap, and does not assume anything that would ensure $\mathbb{E}\left[Y \mid X, Z\right]$ can be estimated at *any* geometric rate.

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 $[n]$ 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*

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

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:

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

- 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*:

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

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

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

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

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 $\mathbb{E}\left[Y \mid X, Z\right] \overset{a.s.}{=} \mathbb{E}\left[Y \mid Z\right]$, and in particular it is exactly zero when $Y \perp\!\!\!\perp X \mid Z$ and strictly positive if $\mathbb{E}\left[Y \mid X, Z\right]$ depends at all on X, allowing it to fully capture arbitrary nonlinearities and interactions in $\mathbb{E}\left[Y \mid X, Z\right]$.

2.2 Floodgate: asymptotic lower confidence bounds for the mMSE gap

As can be seen by Equation (2.4), the mMSE gap is a nonlinear functional of the true regression function $\mu^\star(x, z) := \mathbb{E}\left[Y \mid X = x, Z = z\right]$. 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 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 by keeping the water level (L) below a critical threshold ($\mathcal{I}$) under arbitrary weather/storm conditions (μ). 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

$$f(\mu) := \frac{\mathbb{E}\left[\text{Cov}(\mu^\star(X, Z), \mu(X, Z) \mid Z)\right]}{\sqrt{\mathbb{E}\left[\text{Var}(\mu(X, Z) \mid Z)\right]}}, \tag{2.5}$$

4

where by convention we define $0/0 = 0$ so that $f(\mu)$ remains well-defined when the denominator of (2.5) 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): we assume we know $P_{X|Z}$ but avoid assumptions on $Y \mid X, Z$. In practice $P_{X|Z}$ may be known due to experimental randomization or can be accurately estimated from a large unlabeled data set, but we also quantify the robustness of our inferences to this assumption in Section 2.5 and show it can sometimes be relaxed in Section 3.2. 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. And if we rewrite the numerator as

$$\mathbb{E}\left[\mathrm{Cov}(\mu^\star(X, Z), \mu(X, Z) \mid Z)\right] = \mathbb{E}\left[Y\left(\mu(X, Z) - \mathbb{E}\left[\mu(X, Z) \mid Z\right]\right)\right], \tag{2.6}$$

then we see we also know $\{R_i := Y_i\left(\mu(X_i, Z_i) - \mathbb{E}\left[\mu(X, Z_i) \mid Z_i\right]\right)\}_{i=1}^n$ 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 (2.5), 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.

Algorithm 1 Floodgate

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

Compute $R_i = Y_i\left(\mu(X_i, Z_i) - \mathbb{E}\left[\mu(X_i, Z_i) \mid Z_i\right]\right)$ 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 regression function $\mu : \mathbb{R}^p \to \mathbb{R}$ and i.i.d. data $\{(Y_i, X_i, Z_i)\}_{i=1}^n$, if $\mathbb{E}[Y^8]$, $\mathbb{E}[\mu^8(X, Z)] < \infty$, then $L_n^\alpha(\mu)$ from Algorithm 1 satisfies $\mathbb{P}\left(L_n^\alpha(\mu) \leq f(\mu)\right) \geq 1 - \alpha - O(n^{-1/2})$, which combined with Lemma 2.2 immediately establishes*

$$\mathbb{P}\left(L_n^\alpha(\mu) \leq \mathcal{I}\right) \geq 1 - \alpha - O(n^{-1/2}).$$

The proof can be found in Appendix A.1.2; 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 8th moments (lower-order moments would be needed for just an $o(1)$ rate). Beyond the pointwise $n^{-1/2}$ consistency of Theorem 2.3, a number of natural questions arise, such as floodgate's performance in high dimensions, that could benefit from a clearer exposition of the constant in the $O(n^{-1/2})$. Unfortunately however, that constant depends on μ and the data distribution in a somewhat complicated way and, although in principle that dependence can be deduced from careful review of the proof, we find it more illuminating to address these questions through the examination of invariances in floodgate.

In particular, floodgate (both f and Algorithm 1) is invariant to two aspects of μ:

(i) floodgate is invariant to any additive term in μ that depends only on Z,

(ii) floodgate is invariant to any positive global constant multiplying μ.

This means that 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 model 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 models. To be precise, floodgate using

5

$\mu(x, z) = cx + g(z)$ 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), and hence for a fixed data distribution, the convergence of floodgate's coverage is uniform over *all* partially-linear μ. Furthermore, it also turns out that when μ is partially linear, floodgate only depends on the data distribution through the *bivariate* distribution of (Y, X'), where $X' := \frac{X - \mathbb{E}[X \mid Z]}{\sqrt{\text{Var}(X - \mathbb{E}[X \mid Z])}}$ is the conditionally standardized version of X. Hence as the data-generating distribution varies, even if Z's dimension increases, as long as (Y, X') remains well-behaved (uniformly bounded higher moments and $\text{Var}(Y X')$ bounded below by a positive constant) the convergence of floodgate's coverage will still be uniform over partially-linear μ.

The final missing piece in our LCB procedure is the choice of μ, and this is where 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 2.4 that there is a direct relationship between the accuracy of μ and the accuracy of the resulting floodgate LCB.

Before continuing our study of floodgate LCBs, we first pause to address a natural question: what about an *upper* confidence bound (UCB)? Unfortunately it is impossible to do something analogous for a UCB, in the sense that to get a non-trivial UCB one needs to assume some sort of structure on $Y \mid X, Z$, such as smoothness or sparsity, that allow it to be estimated at a guaranteed rate; see Appendix C for a formal impossibility statement and proof. In particular, the assumed structure of $Y \mid X, Z$ must be incorporated into the UCB procedure itself to attain nontrivial results, in stark contrast to 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 2.4. Although it is disappointing that further assumptions would be needed for a UCB, we envision MOVI inference predominantly 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.

2.3 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 our assumed knowledge of $P_{X \mid Z}$, 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 $P_{X_i \mid Z_i}$ conditionally independently of X_i and Y_i and thus replace R_i and V_i, respectively, by the sample estimators

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

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

Theorem 2.4. *Under the conditions of Theorem 2.3 and* $\mathbb{E}\left[\text{Var}\left(Y\left(\mu(X, Z) - \mathbb{E}\left[\mu(X, Z) \mid Z\right]\right) \mid Z\right)\right] > 0$, *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

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

The proof can be found in Appendix A.2. Note that the additional assumption beyond Theorem 2.3 of $\mathbb{E}\left[\text{Var}\left(Y\left(\mu(X,Z) - \mathbb{E}\left[\mu(X,Z)\,|\,Z\right]\right)\,|\,Z\right)\right] > 0$ 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 $K > 1$. 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 and, of course, it will always be computationally faster to use smaller K.

2.4   Statistical accuracy

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, as formalized in the following theorem.

Theorem 2.5 (Floodgate accuracy)**.** *For i.i.d. data $\{(Y_i, X_i, Z_i)\}_{i=1}^n$ such that $\mathbb{E}[Y^8] < \infty$, $\text{Var}\left(Y\,|\,X,Z\right) \geq \tau$ a.s. for some $\tau > 0$, and a sequence of regression functions $\mu_n : \mathbb{R}^p \to \mathbb{R}$ such that for some C and all n either $\mathbb{E}\left[\text{Var}\left(\mu_n(X,Z)\,|\,Z\right)\right] = 0$ or $\frac{\mathbb{E}[\mu_n^8(X,Z)]}{\mathbb{E}[\text{Var}(\mu_n(X,Z)\,|\,Z)]^4} \leq C$, the output of Algorithm 1 satisfies*

$$\mathcal{I} - L_n^{\alpha}(\mu_n) = O_p\left(\inf_{\mu \in S_{\mu_n}} \mathbb{E}\left[(\mu(X,Z) - \mu^{\star}(X,Z))^2\right] + n^{-1/2}\right) \tag{2.7}$$

The proof can be found in Appendix A.3. We call the left-hand side of Equation (2.7) the *half-width* (by analogy with the *width* that would measure the accuracy of a two-sided confidence interval) and Theorem 2.5 shows it is directly related 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. We reiterate that the infimum in Equation (2.7) means that floodgate is *self-correcting* with respect to μ_n's conditional mean given Z (through invariance (i)) and global scale (through invariance (ii)).

2.5   Robustness

We now consider what happens when the distribution used in floodgate is not the true $P_{X|Z}$ but 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 $Y\,|\,X,Z$ 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 $\mathbb{E}_Q\left[\cdot\right]$); see Equation (A.68) 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., $\text{Var}_Q\left(\mu(X,Z)\,|\,Z\right) \overset{a.s.}{=} 0$, 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 $\mu : \mathbb{R}^p \to \mathbb{R}$,

$$\bar{\mu}(x,z) = \sqrt{\frac{\mathbb{E}\left[\text{Var}(\mu^{\star}(X,Z)\,|\,Z)\right]}{\mathbb{E}\left[\text{Var}(\mu(X,Z)\,|\,Z)\right]}}\left(\mu(x,z) - \mathbb{E}\left[\mu(X,Z)\,|\,Z = z\right]\right) + \mathbb{E}\left[\mu^{\star}(X,Z)\,|\,Z = z\right], \tag{2.8}$$

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 $\mathcal{I} = 0$. We can now state our formal robustness result.

Theorem 2.6 (Floodgate robustness). *For data $\{(Y_i, X_i, Z_i)\}_{i=1}^n$ i.i.d. draws from P satisfying $\mathbb{E}[Y^8] < \infty$ and $\mathrm{Var}\,(Y \mid X, Z) \geq \tau$ a.s. for some $\tau > 0$, a sequence of regression functions $\mu_n : \mathbb{R}^p \to \mathbb{R}$ such that for some C and all n either $\mathrm{Var}_{Q^{(n)}}\,(\mu_n(X, Z) \mid Z) \overset{a.s.}{=} 0$ or $\dfrac{\max\left\{\mathbb{E}[\mu_n^8(X,Z)], \mathbb{E}_{Q^{(n)}}[\mu_n^8(X,Z)]\right\}}{\mathbb{E}\left[\mathrm{Var}_{Q^{(n)}}\,(\mu_n(X,Z) \mid Z)\right]^4} \leq C$, and a sequence of conditional distributions $Q_{X|Z}^{(n)}$, the output of Algorithm 1 when $Q_{X|Z}^{(n)}$ is used as input satisfies*

$$\mathbb{P}\left(L_n^\alpha(\mu_n) \leq \mathcal{I} + \Delta_n\right) \geq 1 - \alpha - O(n^{-1/2}), \tag{2.9}$$

where

$$\Delta_n = f^{Q^{(n)}}(\mu_n) - \mathcal{I} \leq c_1 \sqrt{\mathbb{E}\left[\chi^2\left(P_{X|Z} \,\|\, Q_{X|Z}^{(n)}\right)\right]} - c_2\,\mathbb{E}\left[(\bar{\mu}_n(X, Z) - \mu^\star(X, Z))^2\right] \tag{2.10}$$

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 2.6 can be found in Appendix A.4. Equation (2.10) 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 (2.8). Given the $n^{-1/2}$-rate half-width lower-bound for floodgate, a sufficient condition for asymptotically-exact coverage is

$$\sqrt{\mathbb{E}\left[\chi^2\left(P_{X|Z} \,\|\, Q_{X|Z}^{(n)}\right)\right]} = o\left(n^{-1/2} + \mathbb{E}\left[(\bar{\mu}_n(X, Z) - \mu^\star(X, Z))^2\right]\right). \tag{2.11}$$

When $Q_{X|Z}^{(n)}$ is a standard well-specified 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 will show how to modify floodgate to attain asymptotically-exact inference without the need for estimation at all. We also note in passing that a weaker form of condition (2.11) that replaces the $n^{-1/2}$ with 1 is sufficient for a weaker guarantee of *asymptotic non-overestimation*, i.e., the property that $\liminf_{n \to \infty} \mathbb{P}\left(L_n^\alpha(\mu_n) \leq \mathcal{I} + \epsilon\right) \geq 1 - \alpha$ for any $\epsilon > 0$.

Theorem 2.6 treats the sequence $Q_{X|Z}^{(n)}$ as fixed, which of course means $Q_{X|Z}^{(n)}$ 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 $Q_{X|Z}^{(n)}$. For $Q_{X|Z}^{(n)}$ however, this strict separation may not be necessary in practice, and in our simulations we found floodgate to be quite robust to estimating $Q_{X|Z}^{(n)}$ on samples that included those used as input to floodgate; see Section 4.6.

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., 2019).

2.6 Straightforward generalizations

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

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 D.

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 and Janson (2020). 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 $\mathcal{I}^2 = \mathbb{E}\left[(\mathbb{E}[Y \mid Z] - \mathbb{E}[Y \mid X, Z])^2\right]$ 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

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

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, $f_{\ell_1}(\mu) \leq \mathcal{I}_{\ell_1}$, 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 2 formalizes the procedure and Theorem 3.3 establishes its asymptotic coverage.

Algorithm 2 Floodgate for the MACM gap

Input: Data $\{(Y_i, X_i, Z_i)\}_{i=1}^n$, $P_{X|Z}$, a 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 $L_n^\alpha(\mu) = 2 \max\left\{\bar{R} - \frac{z_\alpha s}{\sqrt{n}}, 0\right\}$.

Theorem 3.3 (MACM gap floodgate validity). *For any given 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 2 satisfies $\mathbb{P}\left(L_n^\alpha(\mu) \leq f_{\ell_1}(\mu)\right) \geq 1 - \alpha - O(n^{-1/2})$, which combined with Lemma 3.2 immediately establishes*

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

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 2. 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 2.3. We give precise sufficient conditions and a proof of their validity in Appendix E, and defer study of Algorithm 2's accuracy and robustness to future work.

3.2 Relaxing the assumptions by conditioning

In this section we show that we can relax the 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 (2019) 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 (2019).

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 identical) 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.

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

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. Also due to monotonicity of conditional variance, $f_n^{\mathcal{T}}(\mu) \leq f(\mu)$, which will be relevant for the relative accuracy of the proposed procedure compared to regular floodgate.

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 *batches* 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 3 (under the simplifying assumption that the number of batches, n_2, evenly divides the sample size n) 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).

Algorithm 3 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

$$(R_m, V_m) = \frac{1}{n_2} \left(\sum_{i=(m-1)n_2+1}^{mn_2} Y_i \left(\mu(X_i, Z_i) - \mathbb{E}\left[\mu(X_i, Z_i) \mid \boldsymbol{Z}_m, \boldsymbol{T}_m\right]\right), \sum_{i=(m-1)n_2+1}^{mn_2} \mathrm{Var}\left(\mu(X_i, Z_i) \mid \boldsymbol{Z}_m, \boldsymbol{T}_m\right) \right),$$

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 $0/0 = 0$.

Theorem 3.4 (Co-sufficient floodgate validity)**.** *For any given 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* $\mathbb{E}[Y^2] < \infty$ *and* $\mathbb{E}[\mu^4(X, Z)] < \infty$, *then* $L_n^{\alpha, \mathcal{T}}(\mu)$ *from Algorithm 3 satisfies* $\mathbb{P}\left(L_n^{\alpha, \mathcal{T}}(\mu) \leq f_n^{\mathcal{T}}(\mu)\right) \geq 1 - \alpha - o(1)$, *which*

immediately establishes

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

The proof can be found in Appendix A.6; the weaker moment conditions than Theorem 2.3 correspond to the weaker $o(1)$ term, and we defer to future work strengthening it to $O(n^{-1/2})$ following similar techniques as earlier results in the paper. Regarding computation, as in Section 2.3, we can replace the conditional expectations in the expressions for R_m and V_m with Monte Carlo estimates based on resampling $\boldsymbol{X}_m \mid \boldsymbol{Z}_m, \boldsymbol{T}_m$ conditionally independently of $\boldsymbol{X}$ and $\boldsymbol{y}$; see Appendix F.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 establish this for Gaussian and discrete Markov Chain covariate models in Appendices F.2 and F.3, respectively.

4 Simulations

4.1 Setup

In the following subsections of this section, we conduct simulation studies to complement the main theoretical claims of Section 2.2. We study the effects of the sample-splitting proportion (Section 4.2), covariate dimension (Section 4.3), covariate dependence (Section 4.4), sample size (Section 4.5), and model misspecification (Section 4.6). 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.

The covariates are sampled from a Gaussian autoregressive model of order 1 (AR(1)) with autocorrelation 0.3, except in Section 4.4 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 G.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. 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 Rorests (Breiman, 2001); see Appendix G.2 for implementation details of these algorithms. The Monte Carlo version of floodgate from Section 2.3 is not needed for the linear methods, and for the nonlinear methods, $K = 500$ is used.

As we provide the first inference results for the mMSE gap, it is challenging to compare floodgate to alternatives. One exception is in low-dimensional Gaussian linear models, where the mMSE gap is a simple function of the coefficient and thus ordinary least squares (OLS) inference can be compared as an alternative to floodgate; see Appendix G.3 for details of how it is made comparable. Thus, in the low-dimensional linear-$\mu^\star$ simulations of Sections 4.3–4.4, 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).

We always take the significance level $\alpha = 0.05$, and all results are averaged over 64 independent replicates.

4.2 Effect of sample splitting proportion

As mentioned in Section 2.3, we can split a fixed sample size n into a first part of size n_e for estimating $\hat{\mu}$ 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 Figures 1 and 2, we vary the sampling 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, respectively. Our main takeaway from these plots is that, while the optimal choice of splitting proportion varies between distributions and algorithms, the choice of 0.5 seems to frequently achieve a half-width close to the optimum. As one would expect, however, as the signal or sample size grows, there are diminishing returns to n_e, and the optimal sample split for some algorithms moves to the left. 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.
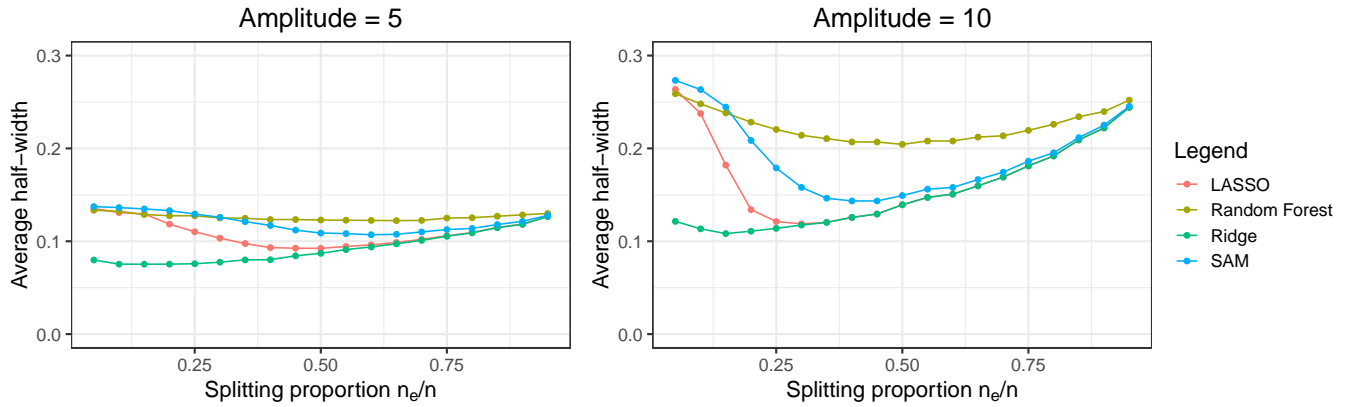
Figure 1: Average half-widths for the linear-$\mu^\star$ simulations of Section 4.2. The coefficient amplitude is given in the plot titles; see Section 4.1 for remaining details.

Figure 2: Average half-widths for the nonlinear-$\mu^\star$ simulations of Section 4.2. The sample size n is given in the plot titles; see Section 4.1 for remaining details.

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

13

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 3, 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 second panel of Figure 3 is similar except with a smaller n_e that is favorable for the linear algorithms in floodgate. 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 be narrower than OLS's when $p \approx n/2$. A similar message is found with nonlinear $\mu^\star$ in Figure 4, except OLS no longer applies and in this case the nonlinear algorithms outperform the linear ones in floodgate.
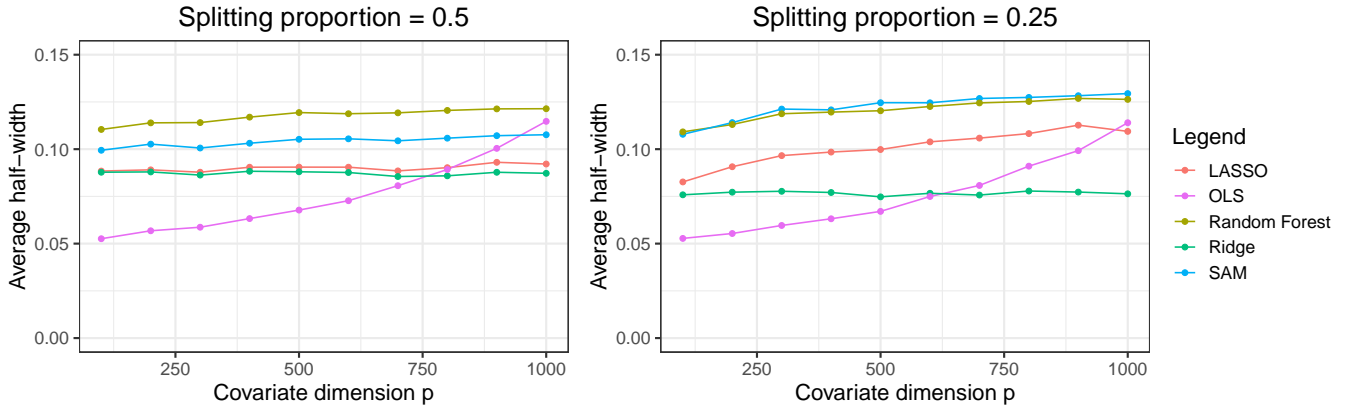
Figure 3: Average half-widths for the linear-$\mu^\star$ simulations of Section 4.3 with floodgate splitting proportion 0.5 (left) and 0.25 (right). OLS is run on the full sample. p is varied on the x-axis; see Section 4.1 for remaining details.
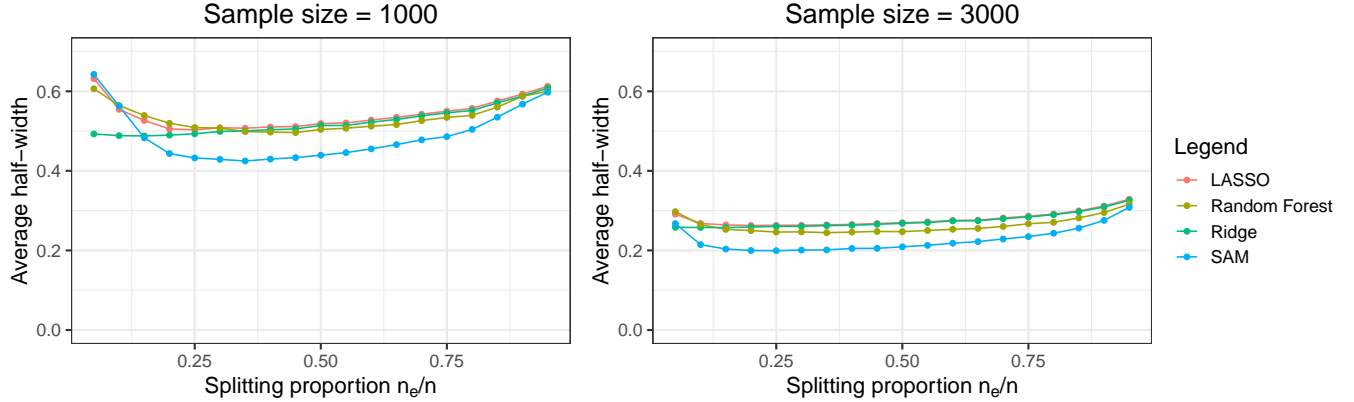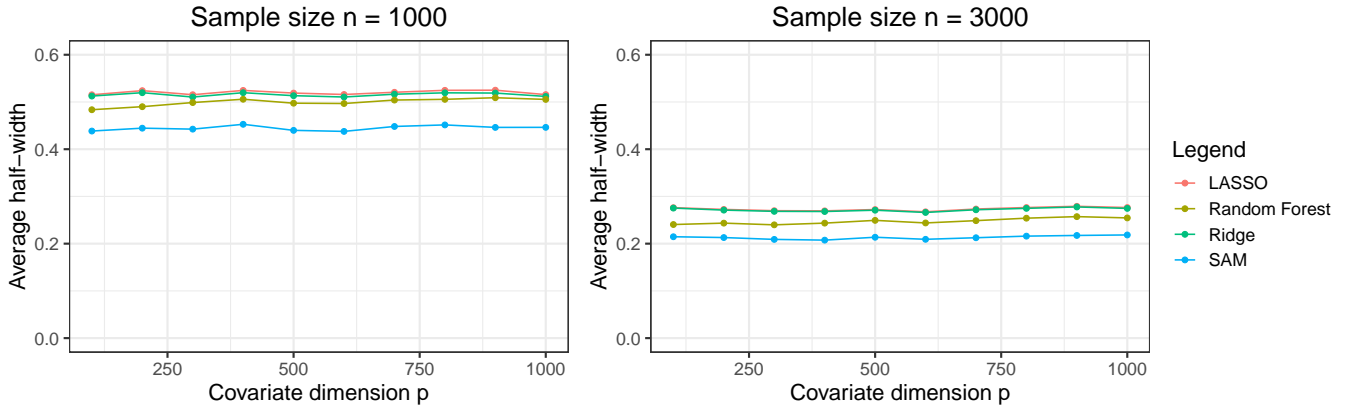

Figure 4: Average half-widths for the nonlinear-$\mu^\star$ simulations of Section 4.3. The sample size n is given in the plot titles and p is varied on the x-axis; see Section 4.1 for remaining details.

14

4.4 Effect of covariate dependence

In Figures 5 and 6, we vary the covariate autocorrelation coefficient 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, respectively. Figure 5 also includes a curve for OLS. Since $\mathcal{I}$ 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 Figures 5 and 6 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 5.
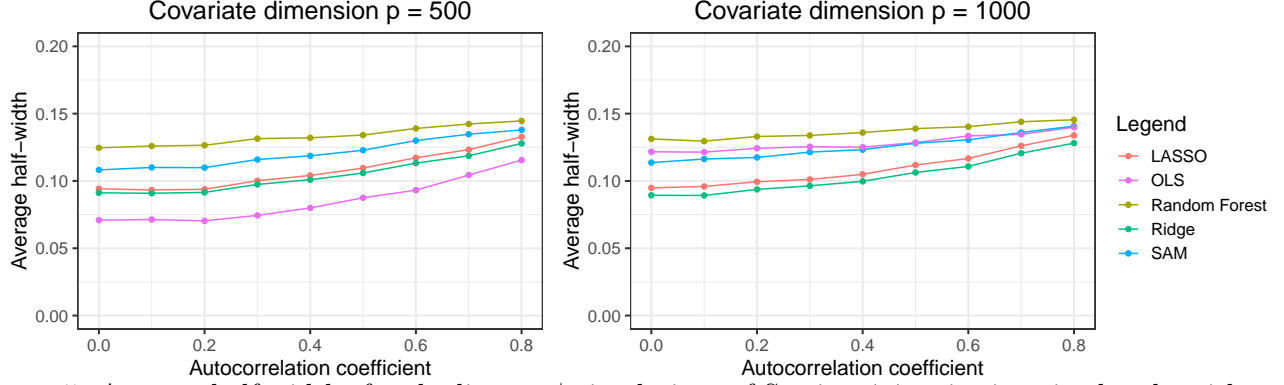
Figure 5: Average half-widths for the linear-$\mu^\star$ simulations of Section 4.4. p is given in the plot titles and the covariate autocorrelation coefficient is varied on the x-axis; see Section 4.1 for remaining details.
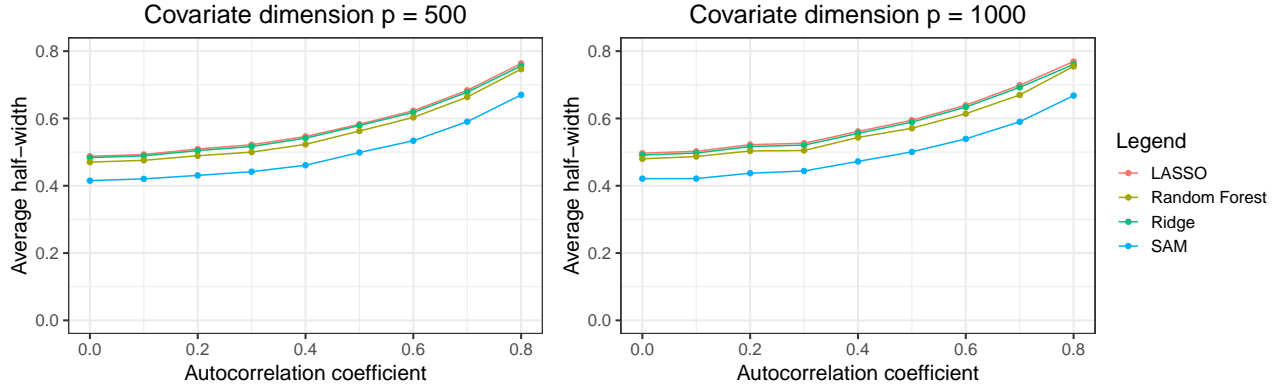
Figure 6: Average half-widths for the nonlinear-$\mu^\star$ simulations of Section 4.4. p is given in the plot titles and the covariate autocorrelation coefficient is varied on the x-axis; see Section 4.1 for remaining details.

4.5 Effect of sample size

In Figures 7 and 8, 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 can achieve much more accurate inference for larger sample sizes.
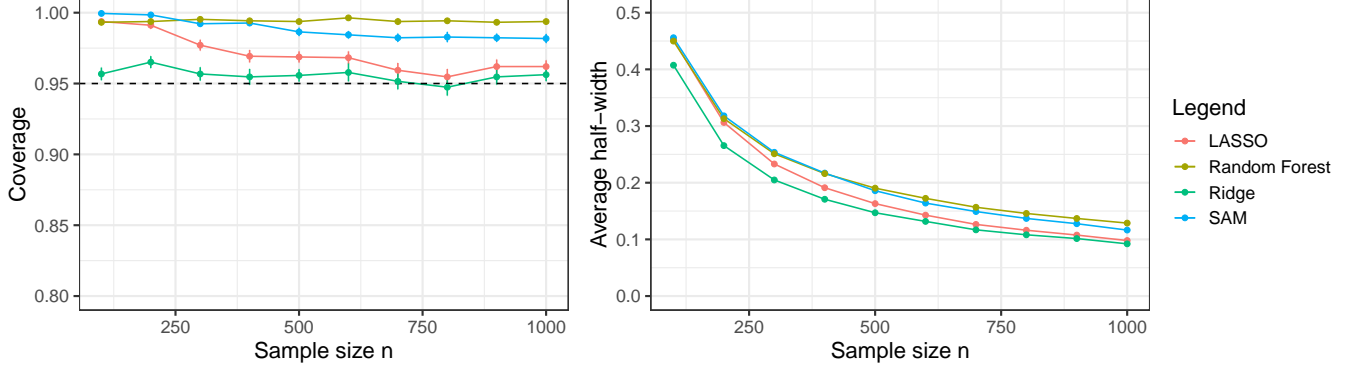
Figure 7: Coverage (left) and average half-widths (right) for the linear-$\mu^\star$ simulations of Section 4.5. The sample size n is varied on the x-axis; see Section 4.1 for remaining details.
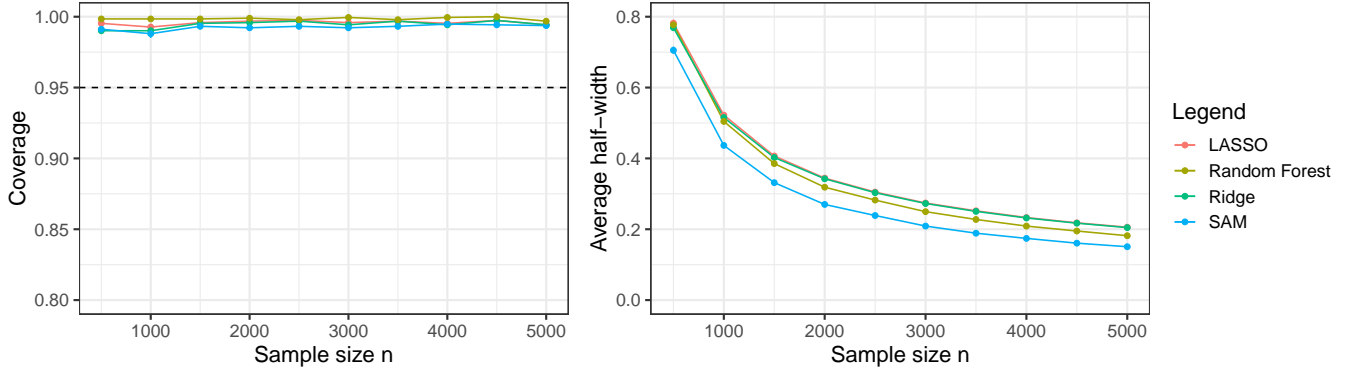
Figure 8: Coverage (left) and average half-widths (right) for the nonlinear-$\mu^\star$ simulations of Section 4.5. The sample size n is varied on the x-axis; see Section 4.1 for remaining details.

4.6 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 2.6 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.

Figures 9 and 10 vary how much in-sample data is used $P_{X|Z}$-estimation and show the coverage of floodgate for null and non-null variables when $\mu^\star$ is linear and nonlinear, respectively. 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 Theorem 2.6 stops applying. However, we see the coverage is consistently quite high, only dropping slightly for very low estimation sample sizes (i.e., very bad estimates of the covariance matrix. In the nonlinear model, we see the coverage being rather conservative for the non-null variables, reflecting the coverage-protective gap between $f(\mu)$ and $f(\mu^\star) = \mathcal{I}$.

16
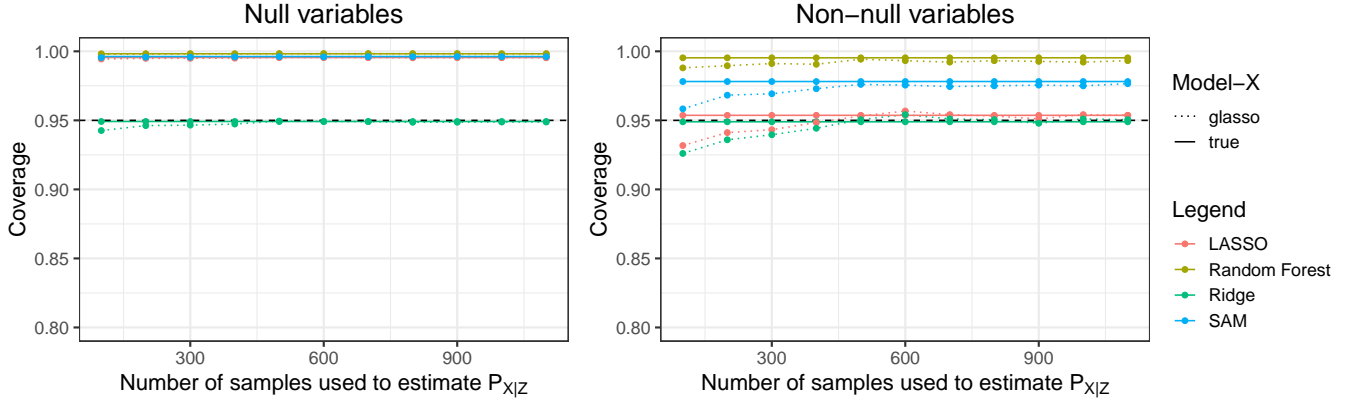
Figure 9: Coverage of null (left) and non-null (right) covariates when the covariate distribution is estimated in-sample for the linear-$\mu^\star$ simulations of Section 4.6. See Section 4.1 for remaining details.
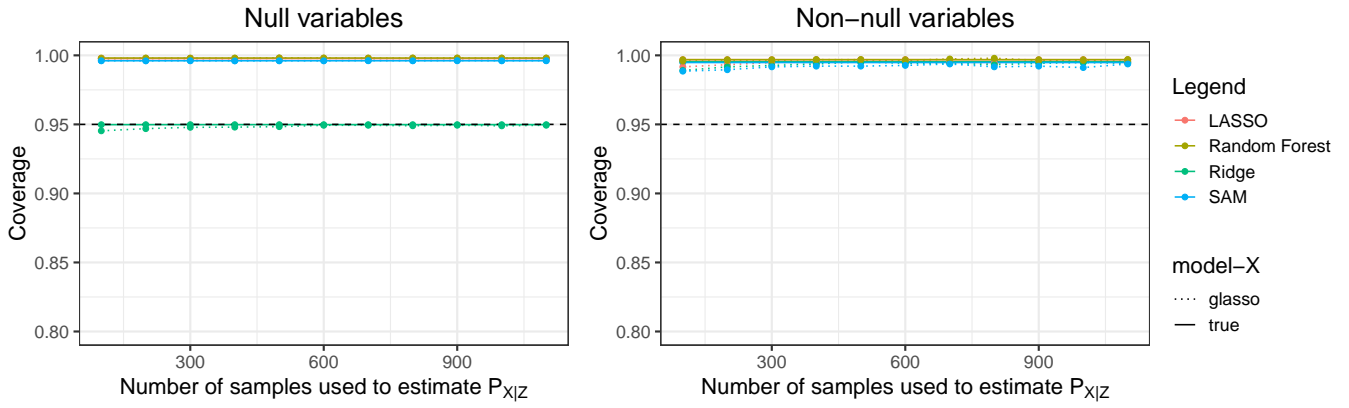
Figure 10: Coverage of null (left) and non-null (right) covariates when the covariate distribution is estimated in-sample for the nonlinear-$\mu^\star$ simulations of Section 4.6. See Section 4.1 for remaining details.

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 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 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., 2020), which carefully applied model-X knockoffs to the same data to perform multi-resolution *selection* of important SNPs, referred as KnockoffZoom. The output of their analysis 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 H for implementation details.
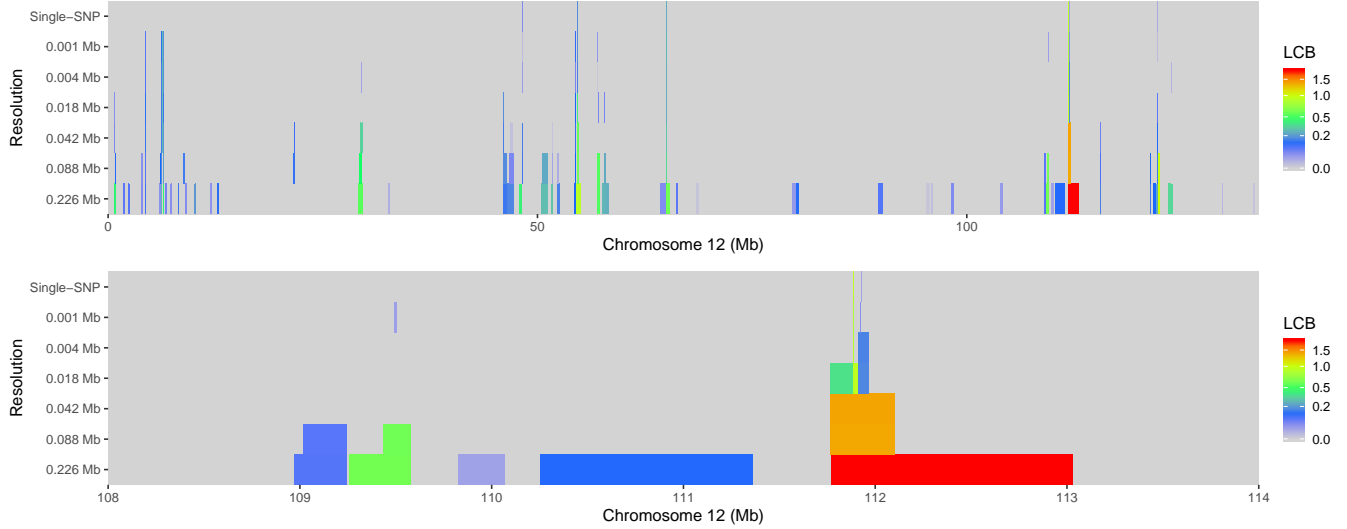
Figure 11: Colored Chicago plot analogous to Figure 1a of Sesia et al. (2020). The color of each point represents the floodgate LCB for the block that contains the SNP at the location indicated on the x-axis at the resolution (measured by average block width) indicated on the y-axis (note some blocks 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.

In particular, Figure 11 is a colored version of Figure 1a of (Sesia et al., 2020), 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. (2020). 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 11).

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 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

18

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

References

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

Bates, S., Sesia, M., Sabatti, C., and Candes, E. (2020). Causal inference in genetic trio studies. *arXiv preprint arXiv:2002.09644*.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 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.

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. (2019). The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

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., Zhao, L., et al. (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.

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

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.

Fisher, A., Rudin, C., and Dominici, F. (2018). Model class reliance: Variable importance measures for any machine learning model class, from the rashomon perspective. *arXiv preprint arXiv:1801.01489*, 68.

Huang, D. and Janson, L. (2019). Relaxing the assumptions of knockoffs by conditioning. *arXiv preprint arXiv:1903.02806*.

Hurt, J. (1976). Asymptotic expansions of functions of statistics. *Aplikace matematiky*, 21(6):444–456.

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*.

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., GSell, 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, E. T., 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.

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. and Janson, L. (2020). Fast and powerful conditional randomization testing via distillation. *arXiv preprint arXiv:2006.0398*.

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.

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.

Rinaldo, A., Wasserman, L., G'Sell, M., Lei, J., and Tibshirani, R. (2016). Bootstrapping and sample splitting for high-dimensional, assumption-free inference. *arXiv preprint arXiv:1611.05401*.

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

Sesia, M., Sabatti, C., and Candès, E. J. (2017). Gene hunting with knockoffs for hidden markov models. *arXiv preprint arXiv:1706.04677*.

Shah, R. D. and Peters, J. (2018). The hardness of conditional independence testing and the generalised covariance measure. *arXiv preprint arXiv:1804.07203*.

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

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. Cambridge university press.

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

Williamson, B. D., Gilbert, P. B., Simon, N., and Carone, M. (2017). Nonparametric variable importance assessment using machine learning techniques. *UW Biostatistics Working Paper Series. Working Paper 422*.

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, 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 proof, 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 $L_2(W) := L_2(\Omega, \sigma(W), P)$, where $\sigma(W)$ is the sub σ-algebra generated by $W = (X, Z)$. ($L_2(Z) := L_2(\Omega, \sigma(Z), P)$ is defined analogously). Then $\mu^\star(W), g^\star(Z)$ can be interpreted as the projection of Y onto the subspace $L_2(W), L_2(Z)$ respectively and $\mu^\star(W)$ admits a orthogonal decomposition $\mu^\star(W) = g^\star(Z) + h^\star(W)$. We remark that implies the following fact:

$$\mathbb{E}[\epsilon(Y, W) \mid W] = 0, \ \mathbb{E}[h(W) \mid Z] = 0. \tag{A.4}$$

Also mentioned in (2.6), we can formally derive the following equivalent expressions of $f(\mu)$,

$$
\begin{aligned}
f(\mu) \ &:= \ \frac{\mathbb{E}\left[\mathrm{Cov}(\mu^\star(X, Z), \mu(X, Z) \mid Z)\right]}{\sqrt{\mathbb{E}\left[\mathrm{Var}(\mu(X, Z) \mid Z)\right]}} \\
&= \ \frac{\mathbb{E}\left[\mathrm{Cov}(h^\star(W), h(W) \mid Z)\right]}{\sqrt{\mathbb{E}\left[h^2(W)\right]}} \\
&= \ \frac{\mathbb{E}\left[h^\star(W) h(W)\right]}{\sqrt{\mathbb{E}\left[h^2(W)\right]}} \\
&= \ \frac{\mathbb{E}\left[Y h(W)\right]}{\sqrt{\mathbb{E}\left[h^2(W)\right]}} - \frac{\mathbb{E}\left[\epsilon(Y, W) h(W)\right]}{\sqrt{\mathbb{E}\left[h^2(W)\right]}} - \frac{\mathbb{E}\left[g^\star(Z) h(W)\right]}{\sqrt{\mathbb{E}\left[h^2(W)\right]}} \\
&= \ \frac{\mathbb{E}\left[Y h(W)\right]}{\sqrt{\mathbb{E}\left[h^2(W)\right]}}
\end{aligned}
\tag{A.5}
$$

$$\tag{A.6}$$

where the second equality is by the definition 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.6) is very concise, we will work with this expression of $f(\mu)$ throughout the following proof. Also note we have a equivalent expression of $\mathcal{I}$.

$$\sqrt{\mathbb{E}\left[(h^\star)^2(W)\right]} = \sqrt{\mathbb{E}\left[\mathbb{E}\left[\left(\mu^\star(W) - \mathbb{E}\left[\mu^\star(W) \mid Z\right]\right)^2 \mid Z\right]\right]} = \sqrt{\mathbb{E}\left[\mathrm{Var}\left(\mathbb{E}\left[Y \mid X, Z\right] \mid Z\right)\right]} = \mathcal{I}. \tag{A.7}$$

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 $f(\mu^\star) = \mathcal{I}$.

$\square$

A.1.2 Theorem 2.3

Proof of Theorem 2.3. Under the stated moment conditions $\mathbb{E}\left[Y^8\right], \mathbb{E}\left[\mu^8(X, Z)\right] < \infty$, we have $\mathbb{E}\left[Yh(W)\right]$ and $\mathbb{E}\left[h^2(W)\right]$ exist, where recall $h(W) = \mu(W) - \mathbb{E}\left[\mu(W) \mid Z\right]$ is defined in Equation (A.2). This holds due to the following elementary facts

$$
\mathbb{E}\left[Yh(W)\right] \leq \sqrt{\mathbb{E}\left[Y^2\right]}\sqrt{\mathbb{E}\left[h^2(W)\right]}, \qquad \mathbb{E}\left[h^2(W)\right] \leq \mathbb{E}\left[(\mathbb{E}\left[\mu(W) \mid Z\right])^2\right] + \mathbb{E}\left[h^2(W)\right] = \mathbb{E}\left[\mu^2(W)\right] \tag{A.8}
$$

which comes from the Cauchy–Schwarz inequality and $\mathbb{E}\left[\mathbb{E}\left[\mu(W) \mid Z\right]h(W)\right] = \mathbb{E}\left[\mathbb{E}\left[\mu(W \mid Z)\mathbb{E}\left[h(W) \mid Z\right]\right]\right] = 0$, respectively. Note $f(\mu) = \mathbb{E}\left[Yh(W)\right]/\sqrt{\mathbb{E}\left[h^2(W)\right]}$ from Equation (A.6), thus $0 \leq f(\mu) < \infty$. First, when $\mathbb{E}\left[\mathrm{Var}(\mu(X, Z) \mid Z)\right] = 0$, we immediately have coverage since $L_n^\alpha(\mu) = 0$ by construction and $\mathcal{I} \geq 0$ by its definition.

Regarding the case where $\mathbb{E}\left[\mathrm{Var}(\mu(X, Z) \mid Z)\right] \neq 0$, we assume $\mathbb{E}\left[h^2(W)\right] = 1$ for the following proof without loss of generality (since floodgate is invariant to positive scaling). By Lemma 2.2, we have $\{L_n^\alpha(\mu) \leq f(\mu)\} \subset \{L_n^\alpha(\mu) \leq \mathcal{I}\}$, so it suffices to show that

$$
\mathbb{P}\left(L_n^\alpha(\mu) \leq f(\mu)\right) \geq 1 - \alpha - O(1/\sqrt{n}). \tag{A.9}
$$

Now we consider four different cases.

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

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

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

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

Note that assuming $\mathbb{E}\left[Y^4\right]$ and $\mathbb{E}\left[\mu^4(X, Z)\right] < \infty$ ensures all the above variances exist. The proof is omitted since later we will use the same strategy to show $\mathbb{E}\left[(\mathrm{Var}\left(h(W) \mid Z\right))^3\right] < \infty$ under the moment condition $\mathbb{E}\left[\mu^6(W)\right] < \infty$. Notice that, when $\mathrm{Var}\left(Yh(W)\right) = 0$, we have $R_i = \mathbb{E}\left[Yh(W)\right]$ for $i \in [n]$, and

thus $\bar{R} = \mathbb{E}\left[Yh(W)\right]$, $\hat{\Sigma}_{11} = \hat{\Sigma}_{12} = 0$; when $\mathrm{Var}\left(\mathrm{Var}\left(h(W) \mid Z\right)\right) = 0$, we have $V_i = \mathbb{E}\left[h^2(W)\right]$ for $i \in [n]$, and thus $\bar{V} = \mathbb{E}\left[h^2(W)\right]$, $\hat{\Sigma}_{22} = \hat{\Sigma}_{12} = 0$.

Case (I): $L_n^\alpha(\mu)$ simply equals $f(\mu)$ since $f(\mu) = \mathbb{E}\left[Yh(W)\right]/\sqrt{\mathbb{E}\left[h^2(W)\right]}$, hence (A.9) holds.

Case (II): we have

$$L_n^\alpha(\mu) = \max\left\{\frac{\bar{R}}{\sqrt{\mathbb{E}\left[h^2(W)\right]}} - \frac{z_\alpha s}{\sqrt{n}},\, 0\right\},$$

where $s^2 = \hat{\Sigma}_{11}/\bar{V}$, and can write down the following equivalence

$$\{L_n^\alpha(\mu) \leq f(\mu)\} = \left\{\bar{R} - \frac{z_\alpha \hat{\Sigma}_{11}}{\sqrt{n}} \leq \mathbb{E}\left[Yh(W)\right]\right\},$$

Now the problem has been reduced to show that

$$\mathbb{P}\left(\bar{R} - \frac{z_\alpha \hat{\Sigma}_{11}}{\sqrt{n}} \leq \mathbb{E}\left[Yh(W)\right]\right) \geq 1 - \alpha - O(1/\sqrt{n}). \tag{A.10}$$

Notice $\bar{R}$ is simply the sample mean estimator of the quantity $\mathbb{E}\left[Yh(W)\right]$ and $\hat{\Sigma}_{11}$ is the corresponding sample variance. Asymptotic coverage validity is a immediate result of the central limit theorem and Slutsky's argument. To establish the $1/\sqrt{n}$ rate, stronger results are needed. The classical Berry–Esseen bound serves as the main ingredient, which states that

Lemma A.1 (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

$$\sup_{x \in \mathbb{R}} |F_n(x) - \Phi(x)| \leq \frac{C\rho}{\sigma^3 \sqrt{n}} \tag{A.11}$$

And since σ in the above result is generally unknown and usually replaced by a consistent estimator, we need an additional Lemma.

Lemma A.2 (Coverage rate)**.** *Let $\mathrm{CI}(n, \alpha) = [T - \frac{s_\sigma z_{\alpha/2}}{\sqrt{n}}, T + \frac{s_\sigma z_{\alpha/2}}{\sqrt{n}}]$ for some $\alpha \in (0, 1)$, where $z_{\alpha/2}$ is the $1 - \alpha/2$ standard normal quantile, if*

- $\sup_{t \in \mathbb{R}} |\mathbb{P}\left(\sqrt{n}(T - \theta) \leq t\sigma\right) - \Phi(t)| \leq \frac{C}{\sqrt{n}}$, *where $\sigma > 0$, and*

- $\mathbb{E}\left[|s_\sigma - \sigma|\right] = O\left(\frac{1}{\sqrt{n}}\right)$,

then

$$\mathbb{P}\left(\theta \in \mathrm{CI}(n, \alpha)\right) \geq 1 - \alpha - O\left(\frac{1}{\sqrt{n}}\right), \tag{A.12}$$

where the constant in $O\left(\frac{1}{\sqrt{n}}\right)$ depends on C, σ and the constant of the rate $O\left(\frac{1}{\sqrt{n}}\right)$ in the assumption.

To apply Lemma A.1, since we are in case II where $\mathrm{Var}\,(Yh(W)) > 0, \mathrm{Var}\,(\mathrm{Var}\,(h(W)\,|\,Z)) = 0$, it suffices to verify the finiteness of the term "ρ" in our context:

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

where the first equality holds since we assume $\mathbb{E}\left[h^2(W)\right] = 1$, the second equality comes from 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 $r \geq 1$). For the last inequality, following the same procedure as (A.8) and using the fact that higher moments dominate lower moments, we obtain the finiteness when assuming $\mathbb{E}\left[Y^6\right], \mathbb{E}\left[\mu^6(W)\right] < \infty$, which holds under the assumed moment conditions.

The first condition of Lemma A.2 has been satisfied. In order to obtain (A.10), we only need to show that $s_\sigma^2 = \hat{\Sigma}_{11}$ (as a consistent estimator of $\sigma^2 := \mathrm{Var}\,(Yh(W))$ satisfies $\mathbb{E}\,[|s_\sigma - \sigma|] = O\left(\frac{1}{\sqrt{n}}\right)$. Notice that s_σ^2 is simply the (unbiased) sample variance estmator, hence $\mathbb{E}\left[s_\sigma^2\right] = \sigma^2$ and we have $\mathrm{Var}\left(s_\sigma^2\right) = O\left(\frac{1}{n}\right)$ when assuming $\mathbb{E}\left[Y^4 h^4(W)\right] < \infty$ (this holds under the stated moment conditions)

$$\begin{aligned}
\mathbb{E}\,[|s_\sigma - \sigma|] &= \mathbb{E}\left[\frac{|s_\sigma^2 - \sigma^2|}{s_\sigma + \sigma}\right] \\
&\leq \frac{\mathbb{E}\left[|s_\sigma^2 - \sigma^2|\right]}{\sigma} \\
&\leq \frac{\sqrt{\mathbb{E}\left[(s_\sigma^2 - \sigma^2)^2\right]}}{\sigma} \\
&= \frac{\sqrt{\mathrm{Var}\left(s_\sigma^2\right)}}{\sigma} = O\left(\frac{1}{\sqrt{n}}\right)
\end{aligned} \tag{A.13}$$

where the first inequality holds since s_σ is non-negative, the second inequality comes from the Cauchy–Schwarz inequality.

Case (III): since $\mathrm{Var}\,(Yh(W)) = 0, \mathrm{Var}\,(\mathrm{Var}\,(h(W)\,|\,Z)) > 0$, we have

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

Denote $T := \frac{\mathbb{E}[Yh(W)]}{\sqrt{\bar{V}}}$, which is a nonlinear function of the moment estimators, then asymptotic normality for T is a direct consequence of the multivariate delta method.

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

where $\tilde{\sigma}$ will be specified later 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 in the following we focus on case (IV) and omit the very similar proof for case (III).

Case (IV): we have

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

and denote $T := \frac{\bar{R}}{\sqrt{\bar{V}}}$. The proof consists of two parts.

(a) Under specific moment conditions, we will establish a Berry–Esseen-type bound for the nonlinear statistic T with the usual rate:

$$\sup_{t \in \mathbb{R}} \left|\mathbb{P}\left(\sqrt{n}(T - f(\mu)) \leq t\tilde{\sigma}\right) - \Phi(t)\right| = O\left(\frac{1}{\sqrt{n}}\right) \tag{A.14}$$

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

25

(b) By verifying that s satisfies Lemma A.2's consistency rate assumption and combining it with the above Berry–Esseen bound, we apply Lemma A.2 to establish (A.9).

Starting with (a), 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 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 A.3. *There exists $\varepsilon, M_\varepsilon > 0$ and a continuous linear functional $L : \mathcal{X} \to \mathbb{R}$ such that*

$$|H(x) - L(x)| \le M_\varepsilon ||x||^2 \quad \text{for all } x \in \mathcal{X} \text{ with } ||x|| \le \varepsilon \tag{A.15}$$

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 the notations to avoid conflict with the notations 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 $||U||_p = (\mathbb{E}[||U||^p])^{1/p}$ with $||u||^p := \sum_{j=1}^{d} |u_j|^p$.

Theorem A.4. *Pinelis et al. (2016, Theorem 2.11) Let $\mathcal{X}$ be a Hilbert space, let H satisfy Condition (A.3) for some real $\epsilon > 0$, and assume $\mathbb{E}[U] = 0$, $\tilde{\sigma} > 0$ and $\nu_3 < \infty$, then*

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{n} H(\bar{U})}{\tilde{\sigma}} \le t \right) - \Phi(t) \right| \le \frac{C}{\sqrt{n}} \tag{A.16}$$

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 (A.11). 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 of the the standard Berry–Esseen bound in Lemma A.1. Remark 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 A.4 to our problem, we first let $\mathcal{X} = \mathbb{R}^2$ and random vectors $\{U_i\}_{i=1}^{n} = \{(U_{i1}, U_{i2})\}_{i=1}^{n} \overset{i.i.d.}{\sim} U = (U_1, U_2)$ to be

$$U_{i1} = R_i - \mathbb{E}[Y h(W)], \quad U_{i2} = V_i - \mathbb{E}[h^2(W)] \tag{A.17}$$

Recall the definition $R_i = Y_i\big(\mu(X_i, Z_i) - \mathbb{E}[\mu(X, Z_i) \mid Z_i]\big)$ and $V_i = \mathrm{Var}\big(\mu(X_i, Z_i) \mid Z_i\big)$, hence we have $\mathbb{E}[U] = 0$. And $T - f(\mu)$ can be rewritten as

$$T - f(\mu) = \frac{\bar{U}_1 + \mathbb{E}[Y h(W)]}{\sqrt{\bar{U}_2 + \mathbb{E}[h^2(W)]}} - \frac{\mathbb{E}[Y h(W)]}{\sqrt{\mathbb{E}[h^2(W)]}} := H(\bar{U})$$

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

$$H(x) = H(x_1, x_2) := \frac{x_1 + \mathbb{E}[Y h(W)]}{\sqrt{x_2 + \mathbb{E}[h^2(W)]}} - \frac{\mathbb{E}[Y h(W)]}{\sqrt{\mathbb{E}[h^2(W)]}}$$

when $x_2 > -\mathbb{E}\left[h^2(W)\right]$ and is set to be $\frac{\mathbb{E}[Yh(W)]}{\mathbb{E}[h^2(W)]}$ otherwise. If we can verify its conditions, Theorem A.4 implies the following

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

First we need to verify Condition (A.3), i.e., there exists $\varepsilon, M_\varepsilon > 0$ and a continuous linear functional $L : \mathbb{R}^2 \to \mathbb{R}$ such that

$$|H(x) - L(x)| \le M_\varepsilon ||x||^2 \quad \text{for all } x \in \mathbb{R}^2 \text{ with } ||x|| \le \varepsilon \tag{A.18}$$

Second, we will show $\tilde{\sigma}$, ν_3, and ς_3 are finite under the stated moment conditions. Regarding the smoothness condition, consider the Taylor expansion of H at zero, we have $H(\mathbf{0}) = 0$ and

$$\frac{\partial H}{\partial x_1}(\mathbf{0})x_1 + \frac{\partial H}{\partial x_2}(\mathbf{0})x_2 = \frac{1}{\sqrt{\mathbb{E}\left[h^2(W)\right]}}x_1 - \frac{\mathbb{E}\left[Yh(W)\right]}{2(\sqrt{\mathbb{E}\left[h^2(W)\right]})^3}x_2.$$

Let $L(x) = L(x_1, x_2)$ be the above linear function, we have $L(\mathbf{0}) = 0$. Note that when $\epsilon = \mathbb{E}\left[h^2(W)\right]/2$,

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

Hence the second partial derivatives exist and are continuous over the compact set $||x|| \le \varepsilon$, thus are also bounded, which implies that there exists $M_\epsilon > 0$ such that (A.18) holds.

As for $\tilde{\sigma}$, ν_3, and ς_3, we verify the following moment bounds

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

Note that $\nu_3^3 = ||U||_3^3 = \mathbb{E}\left[|U_1|^3\right] + \mathbb{E}\left[|U_2|^3\right]$ and

$$
\begin{aligned}
(\varsigma_3\tilde{\sigma})^3 = \mathbb{E}\left[|L(U)|^3\right] &= \mathbb{E}\left[\left|\frac{1}{\sqrt{\mathbb{E}\left[h^2(W)\right]}}U_1 - \frac{\mathbb{E}\left[Yh(W)\right]}{2(\sqrt{\mathbb{E}\left[h^2(W)\right]})^3}U_2\right|^3\right] \\
&\le 2^{3-1}\left(\frac{1}{(\sqrt{\mathbb{E}\left[h^2(W)\right]})^3}\mathbb{E}\left[|U_1|^3\right] + \frac{(\mathbb{E}\left[Yh(W)\right])^3}{8(\sqrt{\mathbb{E}\left[h^2(W)\right]})^9}\mathbb{E}\left[|U_2|^3\right]\right) \quad \text{(A.19)}
\end{aligned}
$$

where 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 (A.19), we only need to show

(i) $\mathbb{E}\left[|U_1|^3\right] < \infty$.

(ii) $\mathbb{E}\left[|U_2|^3\right] < \infty$

(iii) $\tilde{\sigma}^2 = \mathbb{E}\left[L^2(U)\right] > 0$.

under the stated moment conditions. As for (ii), we have

$$
\begin{aligned}
\mathbb{E}\left[|U_2|^3\right] = \mathbb{E}\left[|U_{i2}|^3\right] &= \mathbb{E}\left[\left|V_i - \mathbb{E}\left[h^2(W)\right]\right|^3\right] \\
&\le 2^{3-1}\left(\mathbb{E}\left[\left|\text{Var}\left(\mu(W_i) \mid Z_i\right)\right|^3\right] + (\mathbb{E}\left[h^2(W)\right])^3\right) \\
&\le 2^{3-1}\left(\mathbb{E}\left[\mathbb{E}\left[\mu^6(W_i) \mid Z_i\right]\right] + (\mathbb{E}\left[h^2(W)\right])^3 < \infty\right)
\end{aligned}
$$

where the first inequality comes from to the C_r inequality, the second holds by Jensen's inequality, the third inequality holds due to the tower property of conditional expectation and the assumed moment condition $\mathbb{E}\left[\mu^6(W)\right] < \infty$. As for (i), we have

$$
\begin{aligned}
\mathbb{E}\left[|U_1|^3\right] = \mathbb{E}\left[|U_{1i}|^3\right] &= \mathbb{E}\left[|R_i - \mathbb{E}\left[Yh(W)\right]|^3\right] \\
&\leq 2^{3-1}\left(\mathbb{E}\left[|Y_i(\mu(W_i) - \mathbb{E}\left[\mu(W_i)\,|\,Z_i\right])|^3\right] + (\mathbb{E}\left[Yh(W)\right])^3\right) \\
&= 2^{3-1}\left(\mathbb{E}[|Y^3h^3(W)|] + (\mathbb{E}\left[Yh(W)\right])^3\right) < \infty
\end{aligned}
$$

where the first inequality holds due to the C_r inequality and the second inequality holds since we can upper-bound $\mathbb{E}[|Y^3h^3(W)|]$ as below and use the moment conditions $\mathbb{E}\left[Y^6\right] < \infty, \mathbb{E}\left[\mu^6(W)\right] < \infty$

$$
\begin{aligned}
\mathbb{E}[|Y^3h^3(W)|] &\leq \sqrt{\mathbb{E}[Y^6]\mathbb{E}[(\mu(W) - \mathbb{E}[\mu(W)\,|\,Z])^6]} \\
&\leq \sqrt{\mathbb{E}[Y^6]\cdot 2^5(\mathbb{E}[\mu^6(W)] + \mathbb{E}[(\mathbb{E}[\mu(W)\,|\,Z])^6])} \\
&\leq \sqrt{\mathbb{E}[Y^6]\cdot 2^5(\mathbb{E}[\mu^6(W)] + \mathbb{E}[\mathbb{E}[\mu^6(W)\,|\,Z]])} \\
&\leq 8\sqrt{\mathbb{E}[Y^6]\mathbb{E}[\mu^6(W)]}
\end{aligned}
$$

where for the first three inequalities, we apply the Cauchy–Schwarz inequality, C_r inequality and Jensen's inequality respectively. Now we are left with (iii): first we expand $\tilde{\sigma}^2$ as

$$
\tilde{\sigma}^2 = \mathbb{E}\left[L^2(U)\right] = \frac{1}{\mathbb{E}\left[h^2(W)\right]}\mathbb{E}\left[\left(-\frac{\mathbb{E}\left[Yh(W)\right]}{2\mathbb{E}\left[h^2(W)\right]}U_{i1} + U_{i2}\right)^2\right] \tag{A.20}
$$

According to the definition in (A.17), replacing $h(W)$ by the scaled version $h(W)/\sqrt{\mathbb{E}\left[h^2(W)\right]}$ will not change the value of $\tilde{\sigma}^2$, thus we can assume $\mathbb{E}\left[h^2(W)\right] = 1$ without loss of generality and concisely write down the following expression

$$
\begin{aligned}
\tilde{\sigma}^2 &= \mathbb{E}\left[\left(U_{i1} - \frac{\mathbb{E}\left[Yh(W)\right]}{2}U_{i2}\right)^2\right] \\
&= \mathbb{E}\left[\left(R_i - \mathbb{E}\left[Yh(W)\right] - \frac{\mathbb{E}\left[Yh(W)\right]}{2}\left(\operatorname{Var}\left(h(W_i)\,|\,Z_i\right) - 1\right)\right)^2\right] \\
&= \mathbb{E}\left[(A + B)^2\right] \tag{A.21}
\end{aligned}
$$

where the first and the second equality simply come from (A.17) and $\mathbb{E}\left[h^2(W)\right] = 1$ and the third equality is by rearranging and the terms A, B are defined as below:

$$
\begin{aligned}
A &:= Y_i h(W_i) - \mathbb{E}\left[Y_i h(W_i)\,|\,Z_i\right] \tag{A.22} \\
B &:= \mathbb{E}\left[Y_i h(W_i)\,|\,Z_i\right] - \mathbb{E}\left[Yh(W)\right] - \frac{\mathbb{E}\left[Yh(W)\right]}{2}\left(\operatorname{Var}\left(h(W_i)\,|\,Z_i\right) - 1\right), \tag{A.23}
\end{aligned}
$$

Now we can expand (A.21) as

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

where the first equality comes from the tower property of conditional expectation, the second equality holds since $B \in \sigma(Z_i)$ and the third equality holds due to $\mathbb{E}\left[A\,|\,Z_i\right] = 0$. (A.24) gives one lower bound for

$\tilde{\sigma}^2$. To proceed it in a different way, we equivalently write down

$$
\begin{aligned}
\tilde{\sigma}^2 &= \mathbb{E}\left[\left(U_{i1} - \frac{\mathbb{E}\left[Yh(W)\right]}{2}U_{i2}\right)^2\right] \\
&= \mathbb{E}\left[\left(\left(1, -\frac{\mathbb{E}\left[Yh(W)\right]}{2}\right)(U_{i1}, U_{i2})^\top\right)^2\right] \\
&= \boldsymbol{a}^\top \Sigma_U \boldsymbol{a} \tag{A.25}
\end{aligned}
$$

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

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

Since we are in the case where both $\mathrm{Var}\left(Yh(W)\right)$ and $\mathrm{Var}\left(\mathrm{Var}\left(h(W)\,|\,Z\right)\right)$ are positive, Σ_U will be positive definite if $Yh(W)$ is not a linear function of $\mathrm{Var}\left(h(W)\,|\,Z\right)$. Having (A.24) and (A.25) in hand, we prove $\tilde{\sigma} > 0$ as follows.

When $\mathbb{E}\left[\mathrm{Var}\left(Yh(W)\,|\,Z\right)\right] > 0$, we are done. If $\mathbb{E}\left[\mathrm{Var}\left(Yh(W)\,|\,Z\right)\right] = 0$ holds, then $\tilde{\sigma}^2 = \boldsymbol{a}^\top \Sigma_U \boldsymbol{a} = 0$ implies the degeneracy of Σ_U since the vector $\boldsymbol{a}$ is nonzero. It suffices to show it is impossible to have Σ_U degenerate when $\mathbb{E}\left[\mathrm{Var}\left(Yh(W)\,|\,Z\right)\right] = 0$. Note that in the degenerate case, $Yh(W)$ is a linear function of $\mathrm{Var}\left(h(W)\,|\,Z\right)$, i.e., $Yh(W) = c\mathrm{Var}\left(h(W)\,|\,Z\right) + d$ for some constants c, d, we then obtain

$$
\mathrm{Var}\left(Yh(W)\,|\,Z\right) = \mathrm{Var}\left(c\mathrm{Var}\left(h(W)\,|\,Z\right) + d\,|\,Z\right) = c^2\mathrm{Var}\left(\mathrm{Var}\left(h(W)\,|\,Z\right)\right) > 0
$$

where we make use of the fact $\mathrm{Var}\left(\mathrm{Var}\left(h(W)\,|\,Z\right)\right) > 0$ and $\mathrm{Var}\left(Yh(W)\right) > 0$ (thus $c^2 > 0$). The above result contradicts the assumption $\mathbb{E}\left[\mathrm{Var}\left(Yh(W)\,|\,Z\right)\right] = 0$. This finishes showing the positiveness of $\tilde{\sigma}$, thus verifying (iii). Therefore, the Berry–Esseen-type bounds in (A.14) is established, which completes the proof for part (a).

Regarding part (b), we need to show that s satisfies the requirement of Lemma A.2, i.e., $\mathbb{E}\left[|s - \tilde{\sigma}|\right] = O\left(\frac{1}{\sqrt{n}}\right)$. Again $\mathbb{E}\left[h^2(W)\right] = 1$ without loss of generality and $\tilde{\sigma}^2 = \mathbb{E}\left[\left(V_{i1} - \frac{\mathbb{E}[Yh(W)]}{2}V_{i2}\right)^2\right]$. According to Algorithm 1, s^2 is a estimator of $\tilde{\sigma}^2$, constructed in a way that the population quantities $\mathbb{E}\left[Yh(W)\right], \mathrm{Var}\left(Yh(W)\right), \mathrm{Cov}\left(Yh(W), \mathrm{Var}\left(h(W)\,|\,Z\right)\right), \mathrm{Var}\left(\mathrm{Var}\left(h(W)\,|\,Z\right)\right)$ are replaced by their sample mean counterparts. Similarly as dealing with $\mathbb{E}\left[|s_\sigma - \sigma|\right]$, we can show $\mathbb{E}\left[|s - \tilde{\sigma}|\right] = O\left(\frac{1}{\sqrt{n}}\right)$ when assuming

$$
\mathbb{E}\left[Y^2h^2(W)\right], \mathbb{E}\left[Y^4h^4(W)\right], \mathbb{E}\left[Y^2h^2(W)(\mathrm{Var}\left(h(W)\,|\,Z\right))^2\right], \mathbb{E}\left[(\mathrm{Var}\left(h(W)\,|\,Z\right))^4\right] < \infty
$$

The above requirements are satisfied under the stated moment conditions $\mathbb{E}\left[Y^8\right], \mathbb{E}\left[\mu^8(W)\right] < \infty$, which is based on similar bounding strategy for dealing with $\mathbb{E}\left[Y^3h^3(W)\right]$ in previous steps. Now, combining (a) and (b) yields (A.9), which completes the proof for case (IV). Thus, the asymptotic coverage validity with a rate of $1/\sqrt{n}$ for the lower confidence bounds produced by Algorithm 1 has been established. $\qquad\square$

A.1.3   Ancillary proof

Proof of Lemma A.2. First denote $Z \sim \mathcal{N}\left(0, \sigma^2\right)$ and define $\hat{Z}$ to be $\hat{Z}\,|\,s_\sigma \sim \mathcal{N}\left(0, s_\sigma^2\right)$, then we have

$$
\begin{aligned}
\mathbb{P}\left(\theta \in \mathrm{CI}(n, \alpha)\right) &= \mathbb{P}\left(\sqrt{n}|T - \theta| \leq z_\alpha s_\sigma\right) \\
&\geq \mathbb{P}\left(|Z| \leq z_\alpha s_\sigma\right) - \sup_{t \geq 0}\left|\mathbb{P}\left(\sqrt{n}|T - \theta| \leq t\right) - \mathbb{P}\left(|Z| \leq t\right)\right| \\
&\geq \mathbb{P}\left(|\hat{Z}| \leq z_\alpha s_\sigma\right) + \mathbb{P}\left(|Z| \leq z_\alpha s_\sigma\right) - \mathbb{P}\left(|\hat{Z}| \leq z_\alpha s_\sigma\right) - \frac{C}{\sqrt{n}} \\
&\geq \mathbb{P}\left(|\hat{Z}| \leq z_\alpha s_\sigma\right) - \sup_{t \geq 0}\left|\mathbb{P}\left(|\hat{Z}| \leq t\right) - \mathbb{P}\left(|Z| \leq t\right)\right| - \frac{C}{\sqrt{n}} \\
&= 1 - \alpha - \sup_{t \geq 0}\left|\mathbb{P}\left(|\hat{Z}| \leq t\right) - \mathbb{P}\left(|Z| \leq t\right)\right| - \frac{C}{\sqrt{n}} \tag{A.26}
\end{aligned}
$$

29

where the first inequality holds by the definition of sup, the second equality holds as a result of the rearranging and the assumed Berry–Esseen type bound for T and the last equality comes from

$$\mathbb{P}\left(|\hat{Z}| \leq z_\alpha s_\sigma\right) = \mathbb{E}\left[\mathbb{P}\left(|\hat{Z}| \leq z_\alpha s_\sigma \,|\, s_\sigma\right)\right] = \mathbb{E}\left[1 - \alpha\right] = 1 - \alpha.$$

Now it suffices to bound $\Delta(t) := \left|\mathbb{P}\left(\hat{Z} \leq \sigma t\right) - \mathbb{P}\left(Z \leq \sigma t\right)\right|$, we have

$$
\begin{aligned}
\Delta(t) = \left|\mathbb{P}\left(\hat{Z} \leq \sigma t\right) - \mathbb{P}\left(Z \leq \sigma t\right)\right| &= \left|\mathbb{E}\left[\mathbb{P}\left(\hat{Z} \leq \sigma t \,|\, s_\sigma\right)\right] - \Phi(t)\right| \\
&= \left|\mathbb{E}\left[\Phi(\sigma t / s_\sigma)\right] - \Phi(t)\right|
\end{aligned}
\tag{A.27}
$$

where the first equality holds due to the law of total probability and the second equality is by the definition of $\hat{Z}$. Consider the Taylor expansion of $\Phi(\sigma t / s)$ (as a function of s when treating t fixed) at σ, we have

$$\Phi(\sigma t/s) = \Phi(t) + (s - \sigma)\left.\frac{d\Phi(\sigma t/s)}{ds}\right|_{s=\sigma} + o_p(|s - \sigma|) = \Phi(t) - (s - \sigma)\phi(\frac{\sigma t}{\sigma})\frac{\sigma t}{\sigma^2} + o_p(|s - \sigma|)$$

where $\phi(t)$ is the standard normal density function. Intuitively, since $\Phi(\sigma t/s)$ is smooth enough over a small region around σ, we expect that the term $|\mathbb{E}\left[\Phi(\sigma t/s_\sigma)\right] - \Phi(t)|$ to keep the rate of $\mathbb{E}\left[|s_\sigma - \sigma|\right]$. Formally, notice that $\Phi(\sigma t/s) - \Phi(t)$ is bounded, and $\lim_{x\to\infty} \phi(x)x = 0$ thus $\sup_{x\geq 0} \phi(x)x \leq C_0$ for some C_0, we have $\phi(t)\frac{t}{\sigma}$ bounded over a small region around σ, for any $t \geq 0$. Similarly, we can show the boundedness of the third derivative (uniformly for t). Then under the condition $\mathbb{E}\left[|s_\sigma - \sigma|\right] = O\left(\frac{1}{\sqrt{n}}\right)$, we can apply Theorem 1 in Hurt (1976) (with q chosen to be 1) to (A.27) thus obtain the following

$$\sup_{t\geq 0}\Delta(t) \;\leq\; C'\mathbb{E}\left[|s_\sigma - \tilde{\sigma}|\right] \leq \frac{C''}{\sqrt{n}} \tag{A.28}$$

for some constant $C', C'' > 0$ which will depend on σ, C_0, and the constant in the $\frac{1}{\sqrt{n}}$ rate of the term $\mathbb{E}\left[|s_\sigma - \sigma|\right]$. Combining (A.28) with (A.26) establishes (A.12). $\qquad\square$

A.2 Proofs in Section 2.3

Proof of Theorem 2.4. Similarly as in the proof of 2.3, we immediately have the coverage validity when $\mu(X) \in \sigma(Z)$. Otherwise, it suffices to show

$$\inf_{K>1} \mathbb{P}\left(L_n^\alpha(\mu) \leq f(\mu)\right) \geq 1 - \alpha - O(1/\sqrt{n}). \tag{A.29}$$

Recall that the proof A.1.2 considers 4 different cases then deals with them separately. Now the conditional quantities in Algorithm 1 are replaced by their Monte Carlo estimators R_i^K, V_i^K as defined below.

$$R_i^K = Y_i\left(\mu(X_i, Z_i) - \frac{1}{K}\sum_{k=1}^{K}\mu(X_i^{(k)}, Z_i)\right), \qquad V_i^K = \frac{1}{K-1}\sum_{k=1}^{K}\left(\mu(X_i^{(k)}, Z_i) - \frac{1}{K}\sum_{k=1}^{K}\mu(X_i^{(k)}, Z_i)\right)^2,$$
$$\tag{A.30}$$

for fixed $K > 1$. Essentially we can conduct similar analysis, but to avoid lengthy derivations, we assume the moment condition $\mathbb{E}\left[\mathrm{Var}\left(Y(\mu(X, Z) - \mathbb{E}\left[\mu(X, Z)\,|\,Z\right])\,|\,Z\right)\right] > 0$ thus focus on this specific case.

First let

$$T := \frac{\frac{1}{n}\sum_{i=1}^{n} R_i^K}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} V_i^K}},$$

asymptotic normality for the above quantity is a direct consequence of the multivariate delta method:

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

30

And the unknown variance $\tilde{\sigma}^2$ can be replaced by a consistent estimator. To establish (A.29), we follow the proof strategy of Theorem 2.3. Specifically, we apply the Berry-Esseen bound for nonlinear statistics (Theorem A.4) and Lemma A.2.

Starting with the Berry Esseen bound, we first let random vectors $\{U_i\}_{i=1}^n = \{(U_{i1}, U_{i2})\}_{i=1}^n \overset{i.i.d.}{\sim} U = (U_1, U_2)$ to be

$$U_{i1} = R_i^K - \mathbb{E}\left[Yh(W)\right], \quad U_{i2} = V_i^K - \mathbb{E}\left[h^2(W)\right]. \tag{A.31}$$

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

$$\{X_i^{(k)}\}_{k=1}^K \perp\!\!\!\perp (X_i, Y_i) \mid Z_i, \tag{A.32}$$

$$\{X_i^{(k)}\}_{k=1}^K \mid Z_i \overset{i.i.d.}{\sim} X_i \mid Z_i \tag{A.33}$$

thus we have

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^K \mu(\tilde{X}_i^{(k)}, Z_i) \,\middle|\, Z_i\right] = \mathbb{E}\left[\mu(X_i, Z_i) \mid Z_i\right], \tag{A.34}$$

$$\mathbb{E}\left[\frac{1}{K-1}\sum_{k=1}^K \left(\mu(\tilde{X}_i^{(k)}, Z_i) - \frac{1}{K}\sum_{k=1}^K \mu(\tilde{X}_i^{(k)}, Z_i)\right)^2 \,\middle|\, Z_i\right] = \mathrm{Var}\left(\mu(X_i, Z_i) \mid Z_i\right) \tag{A.35}$$

which further implies $\mathbb{E}\left[U\right] = 0$. Specifically, we have the following derivation

$$
\begin{aligned}
\mathbb{E}\left[U_{i2}\right] &= \mathbb{E}\left[V_i^K - \mathbb{E}\left[h^2(W)\right]\right] \\
&= \mathbb{E}\left[\mathrm{Var}\left(\mu(X_i, Z_i) \mid Z_i\right)\right] - \mathbb{E}\left[h^2(W)\right] \\
&= \mathbb{E}\left[\mathrm{Var}\left(h(X_i, Z_i) \mid Z_i\right)\right] - \mathbb{E}\left[h^2(W)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[h^2(X_i, Z_i) \mid Z_i\right] - \left(\mathbb{E}\left[h(X_i, Z_i) \mid Z_i\right]\right)^2\right] - \mathbb{E}\left[h^2(W)\right] = 0
\end{aligned}
$$

where the second equality holds due to (A.35), the third equality holds by the definition of $h(W) = h(X, Z)$, and the last equality holds as a result of $\mathbb{E}\left[h(X_i, Z_i) \mid Z_i\right] = 0$ in (A.4) and the tower property of conditional expectation, and

$$
\begin{aligned}
\mathbb{E}\left[U_{i2}\right] &= \mathbb{E}\left[R_i^K - \mathbb{E}\left[Yh(W)\right]\right] \\
&= \mathbb{E}\left[Y_i\left(\mu(X_i, Z_i) - \frac{1}{K}\sum_{k=1}^K \mu(\tilde{X}_i^{(k)}, Z_i)\right)\right] - \mathbb{E}\left[Yh(W)\right] \\
&= \mathbb{E}\left[Y_i\mu(W_i)\right] - \mathbb{E}\left[\mathbb{E}\left[Y_i \mid Z_i\right]\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^K \mu(\tilde{X}_i^{(k)}, Z_i) \,\middle|\, Z_i\right]\right] - \mathbb{E}\left[Yh(W)\right] \\
&= \mathbb{E}\left[Y_i\mu(W_i)\right] - \mathbb{E}\left[\mathbb{E}\left[Y_i \mid Z_i\right]\mathbb{E}\left[\mu(X_i, Z_i) \mid Z_i\right]\right] - \mathbb{E}\left[Yh(W)\right] \\
&= \mathbb{E}\left[Y_i\mu(W_i)\right] - \mathbb{E}\left[Y_i\mathbb{E}\left[\mu(X_i, Z_i) \mid Z_i\right]\right] - \mathbb{E}\left[Yh(W)\right] = 0
\end{aligned}
$$

where the first and second equality follow by the definition, the third equality holds due to (A.32), the fourth equality holds due to (A.34), the fifth equality comes from the tower property of total expectation and the last one is by the definition of $h(W)$. And $T - f(\mu)$ can be rewritten as

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

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

$$H(x) = f(x_1, x_2) := \frac{x_1 + \mathbb{E}\left[Yh(W)\right]}{\sqrt{x_2 + \mathbb{E}\left[h^2(W)\right]}} - \frac{\mathbb{E}\left[Yh(W)\right]}{\sqrt{\mathbb{E}\left[h^2(W\right]}}$$

when $x_1 > -\mathbb{E}\left[h^2(W)\right]$ and is set to be $\frac{\mathbb{E}[Yh(W)]}{\mathbb{E}[h^2(W)]}$ otherwise. This is the same nonlinear function as that in the proof of Theorem 2.3, where we already verify the smoothness condition, i.e. Condition (A.3).

Now it is left to verify the following moment bounds

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

Note that $\nu_3^3 = ||U||_3^3 = \mathbb{E}\left[|U_1|^3\right] + \mathbb{E}\left[|U_2|^3\right]$ and

$$
\begin{aligned}
(\varsigma_3\tilde{\sigma})^3 = \mathbb{E}\left[|L(U)|^3\right] &= \mathbb{E}\left[\left|-\frac{\mathbb{E}\left[Yh(W)\right]}{2(\sqrt{\mathbb{E}\left[h^2(W)\right]})^3}U_1 + \frac{1}{\sqrt{\mathbb{E}\left[h^2(W)\right]}}U_2\right|^3\right] \\
&\leq 2^{3-1}\left(\frac{(\mathbb{E}\left[Yh(W)\right])^3}{8(\sqrt{\mathbb{E}\left[h^2(W)\right]})^9}\mathbb{E}\left[|U_1|^3\right] + \frac{1}{(\sqrt{\mathbb{E}\left[h^2(W)\right]})^3}\mathbb{E}\left[|U_2|^3\right]\right) \quad \text{(A.36)}
\end{aligned}
$$

where the inequality holds as a result of the C_r inequality. Due to the fact that the finite of higher moments implies that of lower moments and (A.36), we only need to show

(i) $\mathbb{E}\left[|U_1|^3\right] < \infty$.

(ii) $\mathbb{E}\left[|U_2|^3\right] < \infty$

(iii) $\tilde{\sigma}^2 = \mathbb{E}\left[L^2(U)\right] > 0$.

under the stated moment conditions. Remark the definition of $U = (U_1, U_2)$ depends on K, and we are going to verify (i), (ii) and (iii) for arbitrary $K > 1$. Recall the definition of U_2 in (A.31), we have

$$
\begin{aligned}
\mathbb{E}\left[|U_2|^3\right] = \mathbb{E}\left[|U_{i2}|^3\right] &= \mathbb{E}\left[|V_i^K - \mathbb{E}\left[h^2(W)\right]|^3\right] \\
&\leq 2^{3-1}\left(\mathbb{E}\left[|V_i^K|^3\right] + (\mathbb{E}\left[h^2(W)\right])^3\right)
\end{aligned}
$$

where the inequality holds due to the C_r inequality. Expanding V_i^K, we obtain

$$
\begin{aligned}
\mathbb{E}\left[|V_i^K|^3\right] &= \mathbb{E}\left[\left|\frac{1}{K-1}\sum_{k=1}^{K}\left(\mu(\tilde{X}_i^{(k)}, Z_i) - \frac{1}{K}\sum_{k=1}^{K}\mu(\tilde{X}_i^{(k)}, Z_i)\right)^2\right|^3\right] \quad \text{(A.37)} \\
&= \frac{1}{(K-1)^3}\mathbb{E}\left[\left|\sum_{k=1}^{K}\mu^2(\tilde{X}_i^{(k)}, Z_i) - K\left(\frac{\sum_{k=1}^{K}\mu(\tilde{X}_i^{(k)}, Z_i)}{K}\right)^2\right|^3\right] \\
&\leq \frac{2^{3-1}}{(K-1)^3}\mathbb{E}\left[\left|\sum_{k=1}^{K}\mu^2(\tilde{X}_i^{(k)}, Z_i)\right|^3\right] + \frac{2^{3-1}K^3}{(K-1)^3}\mathbb{E}\left[\left(\frac{\sum_{k=1}^{K}\mu(\tilde{X}_i^{(k)}, Z_i)}{K}\right)^6\right] \\
&\leq 2^5\,\mathbb{E}\left[\left|\frac{\sum_{k=1}^{K}\mu^2(\tilde{X}_i^{(k)}, Z_i)}{K}\right|^3\right] + 2^5\,\mathbb{E}\left[\left(\frac{\sum_{k=1}^{K}\mu(\tilde{X}_i^{(k)}, Z_i)}{K}\right)^6\right] \\
&= 2^5\,(\mathrm{II}_1 + \mathrm{II}_2) \quad \text{(A.38)}
\end{aligned}
$$

where the second equality is simply by expanding and rearranging and the first inequality comes from the C_r inequality. For the last inequality, we use the fact $K > 1$ thus $K \leq 2(K-1)$. Now the problem is reduced to bounding the two terms in (A.38). And by (A.37) and the fact

$$\mathbb{E}\left[\mu(\tilde{X}_i^{(k)}, Z_i)\,\Big|\,Z_i\right] = \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\mu(\tilde{X}_i^{(k)}, Z_i)\,\Big|\,Z_i\right], \quad \forall\, k = 1, \cdots, K.$$

32

we can assume $\mathbb{E}\left[\mu(\tilde{X}_i^{(k)}, Z_i)\,\middle|\, Z_i\right] = 0, \ k = 1, \cdots, K$ without loss of generality. We further write $\mathrm{II}_1, \mathrm{II}_2$ as below

$$\mathrm{II}_1 = \mathbb{E}\left[\mathbb{E}\left[\left|\frac{\sum_{k=1}^{K}\mu^2(\tilde{X}_i^{(k)}, Z_i)}{K}\right|^3 \middle| Z_i\right]\right], \quad \mathrm{II}_2 = \mathbb{E}\left[\mathbb{E}\left[\left(\frac{\sum_{k=1}^{K}\mu(\tilde{X}_i^{(k)}, Z_i)}{K}\right)^6 \middle| Z_i\right]\right] \tag{A.39}$$

Conditional on Z_i, $\mu(\tilde{X}_i^{(k)}, Z_i), \ k = 1, \cdots, K$ 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

$$\mathbb{E}\left[\left|\sum_{k=1}^{K}\mu^2(\tilde{X}_i^{(k)}, Z_i)\right|^3 \middle| Z_i\right] \leq c_{3,K}\ \sum_{k=1}^{K}\mathbb{E}\left[\mu^6(\tilde{X}_i^{(k)}, Z_i)\,\middle|\, Z_i\right], \tag{A.40}$$

$$\mathbb{E}\left[\left(\sum_{k=1}^{K}\mu(\tilde{X}_i^{(k)}, Z_i)\right)^6 \middle| Z_i\right] \leq c_{6,K}\ \sum_{k=1}^{K}\mathbb{E}\left[\mu^6(\tilde{X}_i^{(k)}, Z_i)\,\middle|\, Z_i\right] \tag{A.41}$$

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

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

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

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

We then can simply bound $c_{3,K}$ and $c_{6,K}$ by $C'K^{1/2}$ and $C''K^2$ for some universal constants C', C'' which do not depend on K. Combining these with (A.39), (A.40) and (A.41) yields the following

$$\mathrm{II}_1 \leq \frac{C'}{K^{3/2}}\mathbb{E}\left[\mu^6(\tilde{X}_i^{(k)}, Z_i)\right], \quad \mathrm{II}_2 \leq \frac{C''}{K^3}\mathbb{E}\left[\mu^6(\tilde{X}_i^{(k)}, Z_i)\right].$$

Under the moment condition $\mathbb{E}\left[\mu^6(X, Z)\right] < \infty$, we finally obtain $\mathbb{E}\left[|U_2|^3\right] < \infty$ for arbitrary $K > 1$. As for (i), we apply the same bounding strategy to $\mathbb{E}\left[|U_1|^3\right]$:

$$\begin{aligned}
\mathbb{E}\left[|U_1|^3\right] = \mathbb{E}\left[|U_{1i}|^3\right] &= \mathbb{E}\left[|R_i^K - \mathbb{E}\left[Yh(W)\right]|^3\right] \\
&\leq 2^{3-1}\left(\mathbb{E}\left[|R_i^K|^3\right] + (\mathbb{E}\left[Yh(W)\right])^3\right) \\
&\leq 2^{3-1}\left(\sqrt{\mathbb{E}\left[Y_i^6\right]}\sqrt{\mathbb{E}\left[(G_i^K)^6\right]} + (\mathbb{E}\left[Yh(W)\right])^3\right)
\end{aligned}$$

where the equality is by the definition of U_2 in (A.31), the first inequality holds due to the C_r inequality and the second inequality is a result of applying the Cauchy–Schwarz inequality to Y_i^3 and $(G_i^K)^3$, where $G_i^K = \mu(X_i, Z_i) - \frac{1}{K}\sum_{k=1}^{K}\mu(X_i^{(k)}, Z_i)$. Under the moment condition $\mathbb{E}\left[Y^6\right], \mathbb{E}\left[\mu^6(X, Z)\right] < \infty$, it suffices to bound $\mathbb{E}\left[(G_i^K)^6\right]$. Simple expansion gives

$$\mathbb{E}\left[(G_i^K)^6\right] = \mathbb{E}\left[\left|\mu(X_i, Z_i) - \frac{1}{K}\sum_{k=1}^{K}\mu(\tilde{X}_i^{(k)}, Z_i)\right|^6\right] \tag{A.42}$$

$$\leq 2^{3-1}\left(\mathbb{E}\left[\mu^6(X_i, Z_i)\right] + \mathbb{E}\left[\left|\frac{1}{K}\sum_{k=1}^{K}\mu(\tilde{X}_i^{(k)}, Z_i)\right|^6\right]\right) \tag{A.43}$$

where the inequality holds due to the C_r inequality. Then using the similar strategy as bounding the term II_2, i.e. applying the extension of the Bahr-Esseen inequality, we have

$$\mathbb{E}\left[(G_i^K)^6\right] \le \frac{C'''}{K^3}\mathbb{E}\left[\mu^6(\tilde{X}_i^{(k)}, Z_i)\right],$$

for some universal constant C'''.

Regarding (iii), first rewrite G_i^K and V_i^K as below:

$$G_i^K = h(W_i) - \frac{1}{K}\sum_{k=1}^{K} h(\tilde{X}_i^{(k)}, Z_i), \tag{A.44}$$

$$V_i^K = \mathrm{Var}\left(h(W_i)\,|\,Z_i\right) + \frac{1}{K-1}\sum_{k=1}^{K}\left(h(\tilde{X}_i^{(k)}, Z_i) - \frac{1}{K}\sum_{k=1}^{K}h(\tilde{X}_i^{(k)}, Z_i)\right)^2 - \mathrm{Var}\left(h(W_i)\,|\,Z_i\right) \tag{A.45}$$

where we make use of the fact $\mathbb{E}\left[\mu(\tilde{X}_i^{(k)}, Z_i)\,|\,Z_i\right] = \mathbb{E}\left[\mu(X_i, Z_i)\,|\,Z_i\right], k = 1,\cdots,K$ and $h(\tilde{X}_i^{(k)}, Z_i) = \mu(\tilde{X}_i^{(k)}, Z_i) - \mathbb{E}\left[\mu(\tilde{X}_i^{(k)}, Z_i)\,|\,Z_i\right]$, then replace U_1, U_2 by U_{i1}, U_{i2} and expand $\tilde{\sigma}^2$ as

$$\tilde{\sigma}^2 = \mathbb{E}\left[L^2(U)\right] = \frac{1}{\mathbb{E}\left[h^2(W)\right]}\mathbb{E}\left[\left(-\frac{\mathbb{E}\left[Yh(W)\right]}{2(\mathbb{E}\left[h^2(W)\right])}U_{i1} + U_{i2}\right)^2\right] \tag{A.46}$$

According to the definition (A.31) and the expressions in (A.44) and (A.45), replacing $h(W)$ by the scaled version $h(W)/\sqrt{\mathbb{E}\left[h^2(W)\right]}$ will not change the value of $\tilde{\sigma}^2$, thus we can assume $\mathbb{E}\left[h^2(W)\right] = 1$ without loss of generality and concisely write down the following expression

$$\begin{aligned}\tilde{\sigma}^2 &= \mathbb{E}\left[\left(-\frac{\mathbb{E}\left[Yh(W)\right]}{2}U_{i1} + U_{i2}\right)^2\right] \\ &= \mathbb{E}\left[\left(-\frac{\mathbb{E}\left[Yh(W)\right]}{2}(V_i^K - 1) + Y_iG_i^K - \mathbb{E}\left[Yh(W)\right]\right)^2\right] \\ &= \mathbb{E}\left[(\mathrm{III}_1 - \mathrm{III}_2)^2\right]\end{aligned} \tag{A.47}$$

where the first and the second equality simply come from (A.31) and $\mathbb{E}\left[h^2(W)\right] = 1$ and the third equality is by rearranging and the terms $\mathrm{III}_1, \mathrm{III}_2$ are defined as below:

$$\mathrm{III}_1 := -\frac{\mathbb{E}\left[Yh(W)\right]}{2}\left(\mathrm{Var}\left(h(W_i)\,|\,Z_i\right) - 1\right) + Y_ih(W_i) - \mathbb{E}\left[Yh(W)\right]$$

$$\mathrm{III}_2 := Y_i\left(\frac{1}{K}\sum_{k=1}^{K}h(\tilde{X}_i^{(k)}, Z_i)\right) + \frac{\mathbb{E}\left[Yh(W\right]}{2}\left(\frac{1}{K-1}\sum_{k=1}^{K}\left(h(\tilde{X}_i^{(k)}, Z_i) - \frac{1}{K}\sum_{k=1}^{K}h(\tilde{X}_i^{(k)}, Z_i)\right)^2 - \mathrm{Var}\left(h(W_i)\,|\,Z_i\right)\right)$$

Notice that the definition of $h(\tilde{X}_i^{(k)}, Z_i)$ and (A.35) together imply

$$\mathbb{E}\left[\mathrm{III}_2\,|\,X_i, Y_i\right] = 0. \tag{A.48}$$

Applying the tower property of conditional expectation to (A.47) then expanding yield the following expression:

$$\begin{aligned}\tilde{\sigma}^2 &= \mathbb{E}\left[\mathbb{E}\left[(\mathrm{III}_1^2 + \mathrm{III}_2^2 - 2\mathrm{III}_1\mathrm{III}_2)\,|\,X_i, Y_i\right]\right] \\ &= \mathbb{E}\left[\mathrm{III}_1^2 + \mathbb{E}\left[\mathrm{III}_2^2\,|\,X_i, Y_i\right] - 2\mathrm{III}_1\mathbb{E}\left[\mathrm{III}_2\,|\,X_i, Y_i\right]\right] \\ &= \mathbb{E}\left[\mathrm{III}_1^2 + \mathbb{E}\left[\mathrm{III}_2^2\,|\,X_i, Y_i\right]\right] \\ &\ge \mathbb{E}\left[\mathrm{III}_1^2\right]\end{aligned}$$

where the second equality holds since $\mathrm{III}_1 \in \sigma(X_i, Y_i)$, the third equality comes from (A.48). Remark that III_1 equals $A + B$, where A, B are defined as (A.22) and (A.23) in the proof of Theorem 2.3. Then according to those derivations, we have $\mathbb{E}\left[\mathrm{III}_1^2\right] > 0$ 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, the Berry-Esseen bound for the nonlinear statistics T is obtained. In view of Lemma A.2, it suffices to verify its condition (b) in order to establish (A.29), i.e. proving s is a consistent estimator of $\tilde{\sigma}$ satisfying $\mathbb{E}\left[|s - \tilde{\sigma}|\right] = O\left(\frac{1}{\sqrt{n}}\right)$. Note that s^2 is a function of the sample mean estimator $(\bar{R}, \bar{V}, \hat{\Sigma}_{11}, \hat{\Sigma}_{12}, \hat{\Sigma}_{22})$, by applying multivariate delta method. Similarly as the proof of Theorem 2.3, we have $\mathbb{E}\left[(s - \tilde{\sigma})^2\right] = O\left(\frac{1}{n}\right)$ when the following higher moments exist, i.e.

$$\mathbb{E}\left[(V_i^K)^4\right], \mathbb{E}\left[(R_i^K)^4\right], \mathbb{E}\left[(V_i^K R_i^K)^2\right] < \infty.$$

The above holds for arbitrary $K > 1$ when applying the similar strategy as showing $\mathbb{E}\left[|U_1|^3\right], \mathbb{E}\left[|U_2|^3\right] < \infty$ in previous derivations, under the stated moment conditions $\mathbb{E}\left[Y^8\right], \mathbb{E}\left[\mu^8(X,Z)\right] < \infty$. Finally, we conclude the asymptotic coverage with a rate of $n^{-1/2}$ for any given $K > 1$. $\qquad\square$

A.3 Proofs in Section 2.4

Proof of Theorem 2.5. 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

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

Then it suffices to separately show

$$\mathcal{I} - f(\mu_n) = O_p\left(\inf_{\mu' \in S_{\mu_n}} \mathbb{E}\left[(\mu_n'(X,Z) - \mu^\star(X,Z))^2\right]\right) \tag{A.49}$$

$$f(\mu_n) - L_\alpha^n(\mu_n) = O_P\left(n^{-1/2}\right) \tag{A.50}$$

Recall the definitions in Algorithm 1, when $\mu(X,Z) \in \sigma(Z)$, we have $f(\mu_n) = L_\alpha^n(\mu_n) = 0$, hence in the following we focus on the case where $\mu(X,Z) \notin \sigma(Z)$. Note we have

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

then since $f(\mu_n) - L_\alpha^n(\mu_n) \leq |T - f(\mu_n)| + \frac{z_\alpha s}{\sqrt{n}}$, it suffices to show

$$|T - f(\mu_n)| = O_P\left(n^{-1/2}\right), \quad s = O_P(1).$$

When conditioning on μ_n, showing the above is quite straightforward: in the proof of Theorem 2.3, we establish the asymptotic normality of T and show s converges in probability to $\tilde{\sigma}$ (which is the variance of the asymptotic normal distribution, as defined in (A.20). Unconditionally, we need slightly more work and the stated uniform moment conditions. The proof proceeds through verifying the following: note that by definition of bounded in probability, $|T - f(\mu_n)| = O_P\left(n^{-1/2}\right)$ says for any $\epsilon > 0$, there exists M for which

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

It suffices to prove for any $\mu_n \in \mathcal{U} := \{\mu : \mathbb{E}\left[\mu^8(X,Z)\right] / (\mathbb{E}\left[\mathrm{Var}\left(\mu(X,Z)\,|\,Z\right)\right])^4 \leq C\}$,

$$\sup_n \mathbb{P}\left(\sqrt{n}|T - f(\mu_n)| > M\right) \leq \epsilon, \tag{A.51}$$

and the choice of M (when fixing ϵ) is uniform over $\mu_n \in \mathcal{U}$. Define the random variable G_{μ_n} by $G_{\mu_n} \overset{d}{\sim} \mathcal{N}\left(0, \tilde{\sigma}^2(\mu_n)\right)$, where $\tilde{\sigma}^2(\mu_n)$ denotes the variance $\tilde{\sigma}^2$ with the input of Algorithm 1 being μ_n, then we have

$$\mathbb{P}\left(\sqrt{n}|T - f(\mu_n)| > M\right) \le \mathbb{P}\left(|G_{\mu_n}| > M\right) + \Delta \tag{A.52}$$

where Δ is defined as

$$\Delta := \sup_{\mu_n \in \mathcal{U}} \sup_{M > 0} \left|\mathbb{P}\left(\sqrt{n}|T - f(\mu_n)| > M\right) - \mathbb{P}\left(|G_{\mu_n}| > M\right)\right| \tag{A.53}$$

Recall the derivations in the proof of Theorem A.1.2 where we assume $\mathbb{E}\left[h^2(W)\right] = 1$ without loss of generality (since we can always scale h by $\sqrt{\mathbb{E}\left[h^2(W)\right]}$), here we have a sequence of regression function estimators μ_n which does not admit the same scaling. But the stated moment conditions $\mathbb{E}[Y^8] < \infty$ and $\mathbb{E}\left[\mu_n^8(X, Z)\right] / (\mathbb{E}\left[\operatorname{Var}\left(\mu_n(X, Z) \mid Z\right)\right])^4 \le C$ ensure a uniform moment bound after scaling, hence for the following we can assume $\mathbb{E}\left[h_n^2(W)\right] = 1$. Based on the definition of $\tilde{\sigma}^2$ in (A.20) and the derivations in the proof of Theorem A.1.2, we have $\tilde{\sigma}^2(\mu_n)$ uniformly bounded. Denote this upper bound by $\tilde{\sigma}_0^2$, we then obtain

$$\sup_{\mu_n \in \mathcal{U}} \mathbb{P}\left(|G_{\mu_n}| > M\right) \le \mathbb{P}\left(|G_0| > M\right) \tag{A.54}$$

where $G_0 \overset{d}{\sim} \mathcal{N}\left(0, \tilde{\sigma}_0^2\right)$. According to the proof of Theorem 2.3, we have the following Berry-Esseen bound

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

To show the constant in the above rate of $\frac{1}{\sqrt{n}}$ is uniformly bounded, we first notice that

$$\inf_{\mu_n \in \mathcal{U}} \tilde{\sigma}^2(\mu_n) \ge \inf_{\mu_n \in \mathcal{U}} \mathbb{E}\left[\operatorname{Var}\left(Y h_n(W) \mid Z\right)\right] \tag{A.55}$$

$$\ge \inf_{\mu_n \in \mathcal{U}} \mathbb{E}\left[\operatorname{Var}\left(Y h_n(W) \mid X, Z\right)\right] \tag{A.56}$$

$$= \inf_{\mu_n \in \mathcal{U}} \mathbb{E}\left[h_n^2(W)\operatorname{Var}\left(Y \mid X, Z\right)\right] \tag{A.57}$$

$$\ge \tau > 0 \tag{A.58}$$

where the first inequality holds due to (A.24), the second inequality holds as a result of the total law of conditional variance, the last equality holds by the assumption $\mathbb{E}\left[h_n^2(W)\right] = 1$ and the moment lower bound condition $\operatorname{Var}\left(Y|X, Z\right) \ge \tau > 0$. Assuming $\mathbb{E}[Y^8] < \infty$ and $\mathbb{E}\left[\mu_n^8(X, Z)\right] / (\mathbb{E}\left[\operatorname{Var}\left(\mu_n(X, Z) \mid Z\right)\right])^4 \le C$, we can uniformly control the higher moments involved in (A.4) (i.e. $\nu_2, \nu_3, \varsigma_3$), therefore establish the rate of $\frac{1}{\sqrt{n}}$ in (A.53):

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

Combining this with (A.52) and (A.54), we have

$$\sup_{\mu_n \in \mathcal{U}} \mathbb{P}\left(\sqrt{n}|T - f(\mu_n)| > M\right) \le \mathbb{P}\left(|G_0| > M\right) + \frac{C'}{\sqrt{n}}$$

for some constant C' depending on C, τ and $\mathbb{E}\left[Y^8\right]$. Therefore we obtain (A.51) and the choice of M can be universally chosen over $\mu_n \in \mathcal{U}$, which finally establishes $|T - f(\mu_n)| = O_P\left(n^{-1/2}\right)$. Using similar strategies, we can prove $s = O_P(1)$. Now we proceed to prove (A.49), first it can be simplified into the following form due to (A.5) and (A.7),

$$\mathcal{I} - f(\mu_n) = \sqrt{\mathbb{E}[(h^\star)^2(W)]} - \frac{\mathbb{E}\left[h_n(W)h^\star(W)\right]}{\sqrt{\mathbb{E}\left[h_n^2(W)\right]}} \tag{A.59}$$

36

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.59). 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 $\bar{\mu}_n$:

$$\bar{\mu}_n(x, z) := \sqrt{\frac{\mathcal{I}}{\mathbb{E}\left[h_n^2(W)\right]}} \left(\mu_n(x, z) - \mathbb{E}\left[\mu_n(X, Z) \mid Z = z\right]\right) + \mathbb{E}\left[\mu^\star(X, Z) \mid Z = z\right], \tag{A.60}$$

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 $\mu(X, Z) \in \sigma(Z)$, we have $\bar{\mu}_n(x, z) = \mathbb{E}\left[\mu^\star(X, Z) \mid Z = z\right], \bar{h}_n(w) = 0$, thus

$$\mathcal{I} - f(\mu_n) = \mathcal{I} = \frac{\mathbb{E}\left[(\bar{h}_n(W) - h^\star(W))^2\right]}{\sqrt{\mathbb{E}[(h^\star)^2(W)]}} \tag{A.61}$$

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.59) 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

$$\mathcal{I} - f(\mu_n) = \frac{\mathbb{E}\left[(\bar{h}_n(W) - h^\star(W))^2\right]}{2\sqrt{\mathbb{E}[(h^\star)^2(W)]}} \tag{A.62}$$

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

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

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

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

37

we thus establish (A.63). Regarding case (b), we have

$$
\begin{aligned}
\inf_{\mu' \in S_{\mu_n}} \mathbb{E}\left[(\mu'(X,Z) - \mu^\star(X,Z))^2\right] &= \inf_{c>0} \mathbb{E}\left[(ch_n(W) - h^\star(W))^2\right] \\
&= \inf_{c>0} \mathbb{E}\left[(ch_n(W) - h_0(W) + h_0(W) - h^\star(W))^2\right] \\
&= \mathbb{E}\left[(h_0(W) - h^\star(W))^2\right] + \inf_{c>0} \mathbb{E}\left[(ch_n(W) - h_0(W))^2\right] \\
&= \mathbb{E}\left[(h_0(W) - h^\star(W))^2\right] \\
&= \mathbb{E}\left[(h^\star)^2(W)\right] - \mathbb{E}\left[(h_0(W))^2\right] \quad\quad\quad\quad\quad\quad (\text{A.64})
\end{aligned}
$$

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

$$
\begin{aligned}
\mathbb{E}\left[(\bar{h}_n(W) - h^\star(W))^2\right] &= 2\mathbb{E}\left[(h^\star)^2(W)\right] - 2\mathbb{E}\left[\bar{h}_n(W)h^\star(W)\right] \\
&= 2\mathbb{E}\left[(h^\star)^2(W)\right] - 2\mathbb{E}\left[(h_0(W))^2\right]\rho \quad\quad\quad\quad (\text{A.65})
\end{aligned}
$$

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

$$
\begin{aligned}
\rho &:= \frac{\mathbb{E}\left[\bar{h}_n(W)h^\star(W)\right]}{\mathbb{E}\left[(h_0(W))^2\right]} \\
&= \frac{\mathcal{I}\sqrt{\mathbb{E}\left[h_n^2(W)\right]}}{\mathbb{E}\left[h_n(W)h^\star(W)\right]} \quad\quad\quad\quad\quad\quad\quad (\text{A.66})
\end{aligned}
$$

thus we have $\rho > 0$ in case (b) and $\rho \geq 1$ by the Cauchy-Schwarz inequality. Combining this with (A.64) and (A.65) yields (A.63). Finally we establish the bound in (2.7).

$\square$

A.4   Proofs in Section 2.5

In the true distribution case, the output $L_\alpha^n(\mu)$ from Algorithm 1 is a asymptotic lower confidence bound for $f(\mu)$. 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 output from Algorithm 1 as denoted by $L_\alpha^{n,Q}(\mu)$. Note that $f(\mu)$ can be rewritten with explicit subscripts as below (here we use the equivalent expression of $f(\mu)$ in (A.6) and expand $h(W)$).

$$
f(\mu) = \frac{\mathbb{E}_P\left[Y\left(\mu(X,Z) - \mathbb{E}_P\left[\mu(X,Z)\,|\,Z\right]\right)\right]}{\sqrt{\mathbb{E}_{P_Z}\left[\mathrm{Var}_P\left(\mu(X,Z)\,|\,Z\right)\right]}} \quad\quad\quad\quad (\text{A.67})
$$

Clearly, $L_\alpha^{n,Q}(\mu)$ is a lower confidence bound for the following quantity:

$$
f^Q(\mu) := \frac{\mathbb{E}_P\left[Y\left(\mu(X,Z) - \mathbb{E}_Q\left[\mu(X,Z)\,|\,Z\right]\right)\right]}{\sqrt{\mathbb{E}_{P_Z}\left[\mathrm{Var}_Q\left(\mu(X,Z)\,|\,Z\right)\right]}}. \quad\quad\quad\quad (\text{A.68})
$$

Denote $\omega(x,z) := \frac{dP_{X|Z}(x|z)}{dQ_{X|Z}(x|z)}$ (further abbreviated as $\frac{dP}{dQ}$ without causing confusion). Remark 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 if we consider discrete case. We also enforce the support of Q must contain the support of P. Then we can quantify the difference between $f(\mu)$ and $f^Q(\mu)$ as in Lemma A.5.

Lemma A.5. *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:*

- $\mu(X, Z) \in \sigma(Z)$;

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

for some constants c_0, *then we have the following bounds*

$$\Delta(P, Q) := \sup_{\mu \in \mathcal{U}} |\theta^Q(\mu) - f(\mu)| \leq C\sqrt{\mathbb{E}_{P_Z}\left[\chi^2\left(P_{X|Z}||Q_{X|Z}\right)\right]} \tag{A.69}$$

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 A.5 gives a quantitative characterization about how much we need to adjust.

Proof of Lemma A.5. When $\mu(X, Z) \in \sigma(Z)$, $f(\mu) = f^Q(\mu) = 0$, 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

$$\mathbb{E}_P\left[\mu^4(X, Z)\right], \mathbb{E}_Q\left[\mu^4(X, Z)\right] \leq c_0. \tag{A.70}$$

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

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

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

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

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

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 $1/d$.

$$a/\sqrt{b} = f(\mu) \leq \mathcal{I} \leq (\mathbb{E}_P\left[Y^4\right])^{1/4} \leq (c_1)^{1/4}, \quad 1/d = 1/\mathbb{E}_{P_Z}\left[\mathrm{Var}_Q\left(\mu(X, Z) \mid Z\right)\right] = 1 \tag{A.71}$$

where the first equality is by Lemma 2.2 and the second equality is by applying Jensen's inequality $(\mathbb{E}_{P_Z}\left[\mathrm{Var}_P\left(\mathbb{E}\left[Y \mid X, Z\right] \mid Z\right)\right] \leq \mathbb{E}_{P_Z}\left[\mathbb{E}_P\left[(\mathbb{E}\left[Y \mid X, Z\right])^2 \mid Z\right]\right] \leq \mathbb{E}\left[Y^2\right] \leq \sqrt{\mathbb{E}\left[Y^4\right]})$. 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 $|a - c|$:

$$\begin{aligned}|a - c| &= |\mathbb{E}_P\left[\mu^\star(W)\left(\mu(W) - \mathbb{E}_P\left[\mu(W) \mid Z\right]\right)\right] - \mathbb{E}_P\left[\mu^\star(W)\left(\mu(W) - \mathbb{E}_Q\left[\mu(W) \mid Z\right]\right)\right]| \\ &= |\mathbb{E}_P\left[\mu^\star(W)\left(\mathbb{E}_P\left[\mu(W) \mid Z\right] - \mathbb{E}_Q\left[\mu(W) \mid Z\right]\right)\right]| \\ &= |\mathbb{E}_{P_Z}\left[\mathbb{E}_P\left[\mu^\star(W) \mid Z\right]\left(\mathbb{E}_P\left[\mu(W) \mid Z\right] - \mathbb{E}_Q\left[\mu(W) \mid Z\right]\right)\right]| \end{aligned} \tag{A.72}$$

we can rewrite $|\mathbb{E}_P[\mu(W)\,|\,Z] - \mathbb{E}_Q[\mu(W)\,|\,Z]|$ in the form of integral then derive the following bound

$$\left|\int \mu(x,Z)(1-\omega(x,Z))dQ_{X|Z}(x\,|\,Z)\right| \leq \sqrt{\mathbb{E}_{Q_{X|Z}}[\mu^2(X,Z)\,|\,Z]}\sqrt{\int(1-w(x,Z))^2 dQ_{X|Z}(x\,|\,Z)}$$

$$= \sqrt{\mathbb{E}_{Q_{X|Z}}[\mu^2(W)\,|\,Z]}\sqrt{\chi^2\left(P_{X|Z}||Q_{X|Z}\right)} \tag{A.73}$$

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 (A.73) into (A.72) and further bound $|a-c|$ by

$$|a-c| \leq \mathbb{E}_{P_Z}\left[\mathbb{E}_{P_{X|Z}}[\mu^\star(W)\,|\,Z]\sqrt{\mathbb{E}_{Q_{X|Z}}[\mu^2(W)\,|\,Z]}\sqrt{\chi^2\left(P_{X|Z}||Q_{X|Z}\right)}\right]$$

$$\leq \sqrt{\mathbb{E}_{P_Z}\left[(\mathbb{E}_{P_{X|Z}}[\mu^\star(W)\,|\,Z])^2\mathbb{E}_{Q_{X|Z}}[\mu^2(W)\,|\,Z]\right]}\cdot\sqrt{\mathbb{E}_{P_Z}\left[\chi^2\left(P_{X|Z}||Q_{X|Z}\right)\right]} \tag{A.74}$$

For the first part of the product in (A.74), we can apply the Cauchy–Schwarz inequality and Jensen's inequality and bound it by $(\mathbb{E}_P[(\mu^\star)^4(W)]\,\mathbb{E}_Q[\mu^4(W)])^{1/4}$, which is upper bounded by some constant under the stated condition $\mathbb{E}[Y^4] < \infty$ and $\mathbb{E}_Q[\mu^4(X,Z)] \leq c_0$ (from (A.70)). Now we write down $|b-d|$ below

$$\begin{aligned}|b-d| &= |\mathbb{E}_{P_Z}[\mathrm{Var}_P(\mu(W)\,|\,Z)] - \mathbb{E}_{P_Z}[\mathrm{Var}_Q(\mu(X,Z)\,|\,Z)]| \\ &\leq |\mathbb{E}_{P_Z}[(\mathbb{E}_P[\mu(W)\,|\,Z])^2 - (\mathbb{E}_Q[\mu(W)\,|\,Z])^2]| \\ &\quad + |\mathbb{E}_{P_Z}[\mathbb{E}_P[\mu^2(W)\,|\,Z] - \mathbb{E}_Q[\mu^2(W)\,|\,Z]]|\end{aligned} \tag{A.75}$$

Similarly as (A.73), we have

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

then under the moment bounds $\mathbb{E}_Q[\mu^4(X,Z)] \leq c_0$ in (A.70) we can show the second term in (A.75) is upper bounded by $\sqrt{c_0\mathbb{E}_{P_Z}[\chi^2(P_{X|Z}||Q_{X|Z})]}$. Regarding the first term in (A.75), we can write

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

then apply similar strategies in (A.72) and (A.74) to control it under $C\sqrt{\mathbb{E}_{P_Z}[\chi^2(P_{X|Z}||Q_{X|Z})]}$ for some constant C. And this will make use of the moment bound conditions $\mathbb{E}_P[\mu^4(X,Z)], \mathbb{E}_Q[\mu^4(X,Z)] \leq c_0$ in (A.70). Finally we establish the bound in (A.69). $\qquad\square$

Proof of Theorem 2.6. When $\mu_n(X,Z) \in \sigma(Z)$, we simply have $L^\alpha_{n,Q^{(n)}}(\mu_n) = 0$, thus

$$\mathbb{P}\left(L^\alpha_{n,Q^{(n)}}(\mu_n) \leq \mathcal{I}\right) = 1 \geq 1 - \alpha - O(n^{-1/2}).$$

Otherwise we consider the nontrivial case where $\mathbb{E}_{P_Z}\left[\mathrm{Var}_{Q^{(n)}}(\mu(X,Z)\,|\,Z)\right] > 0$. Similarly as in the proof of Theorem 2.5, when assuming $\mathbb{E}[Y^8] < \infty$, the moment lower bound condition $\mathrm{Var}(Y\,|\,X,Z) \geq \tau > 0$ and a uniform moment conditions $\max\left\{\mathbb{E}[\mu^8_n(X,Z)], \mathbb{E}_{Q^{(n)}}[\mu^8_n(X,Z)]\right\}/(\mathbb{E}[\mathrm{Var}_{Q^{(n)}}(\mu_n(X,Z)\,|\,Z)])^4 \leq C$, we have

$$\mathbb{P}\left(L^\alpha_{n,Q^{(n)}}(\mu_n) \leq \mathcal{I} + \Delta_n\right) \geq 1 - \alpha - O(n^{-1/2}).$$

where $\Delta_n = f^{Q^{(n)}}(\mu_n) - \mathcal{I}$. Note that the constant in the rate of $n^{-1/2}$ depends on τ and C. It is worth mentioning that when the specified conditional distribution is $Q^{(n)}$, in the proof of establishing the

coverage rate of $n^{-1/2}$, bounding those higher moments actually involves the term $\mathbb{E}\left[\mu_n^8(X,Z)\right]$, in addition to $\mathbb{E}_{Q^{(n)}}\left[\mu_n^8(X,Z)\right]$.

Now it suffices to characterize the term Δ_n, first notice that

$$\Delta_n = f^{Q^{(n)}}(\mu_n) - \mathcal{I} = (f^{Q^{(n)}}(\mu_n) - f(\mu_n)) - (\mathcal{I} - f(\mu_n)). \tag{A.76}$$

Then we can apply Lemma A.5 to P, $Q^{(n)}$ and μ_n under the stated conditions, which will give the following bound

$$(f^{Q^{(n)}}(\mu_n) - f(\mu_n)) \leq C' \sqrt{\mathbb{E}\left[\chi^2\left(P_{X|Z} \,\|\, Q_{X|Z}^{(n)}\right)\right]} \tag{A.77}$$

for some constant depending on $\mathbb{E}\left[Y^8\right]$ and C. Regarding the term $\mathcal{I} - f(\mu_n)$, we recall the derivations in the proof of Theorem 2.5, specifically (A.61) and (A.62), then the following holds

$$\mathcal{I} - f(\mu_n) \geq \frac{\mathbb{E}\left[(\bar{h}_n(W) - h^\star(W))^2\right]}{2\mathcal{I}} = \frac{\mathbb{E}\left[(\bar{\mu}_n(W) - \mu^\star(W))^2\right]}{2\mathcal{I}} \tag{A.78}$$

where the equality holds by the definition of $h^\star$, $\bar{\mu}_n$ and $\bar{h}_n$. Combining (A.76), (A.77) and (A.78) yields (2.10). $\qquad\square$

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.32), X's null copy $\tilde{X}$ is constructed such that

$$\tilde{X} \perp\!\!\!\perp (X,Y) \mid Z, \quad \text{and} \quad \tilde{X} \mid Z \overset{d}{=} X \mid Z. \tag{A.79}$$

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

$$\tilde{Y} \perp\!\!\!\perp (X,Y) \mid Z, \quad \tilde{Y} \mid Z \overset{d}{=} Y \mid Z \tag{A.80}$$

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.80)). Now by the symmetry argument, we have

$$\mathbb{E}\left[\mathbb{1}_{\{Y \cdot [\mu(\tilde{X},Z) - \mathbb{E}[\mu(X,Z)\,|\,Z]] < 0\}}\right] = \mathbb{E}\left[\mathbb{1}_{\{\tilde{Y} \cdot [\mu(X,Z) - \mathbb{E}[\mu(X,Z)\,|\,Z]] < 0\}}\right]. \tag{A.81}$$

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

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

where the second equality is by (A.81), 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

$$\mathbb{E}\left[\left(\mathbb{1}_{\{\tilde{Y} \cdot h(W) < 0\}} - \mathbb{1}_{\{Y \cdot h(W) < 0\}}\right) \Big| W = w\right] \tag{A.82}$$

41

for each $w = (x, z)$. Due to the property (A.80), we have

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

hence we can simplify the conditional expectation of the first indicator function in (A.82) into the following

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}_{\{\tilde{Y} \cdot h(W) < 0\}} \mid W = w\right] &= \mathbb{P}\left(\tilde{Y} = 1,\ h(W) < 0 \mid W = w\right) + \mathbb{P}\left(\tilde{Y} = -1,\ h(W) > 0 \mid W = w\right) \\
&= \mathbb{P}\left(Y = 1 \mid Z = z\right) \mathbb{1}_{\{h(w) < 0\}} + \mathbb{P}\left(Y = -1 \mid Z = z\right) \mathbb{1}_{\{h(w) > 0\}} \quad (A.83)
\end{aligned}$$

Similarly we have

$$\mathbb{E}\left[\mathbb{1}_{\{Y \cdot h(W) < 0\}} \mid W = w\right] = \mathbb{P}\left(Y = 1 \mid W = w\right) \mathbb{1}_{\{h(w) < 0\}} + \mathbb{P}\left(Y = -1 \mid W = w\right) \mathbb{1}_{\{h(w) > 0\}} \quad (A.84)$$

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

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

hence in this case, by comparing (A.83) and (A.84) we know $h(w) > 0$ will maximize (A.82) with maximum value

$$\begin{aligned}
\mathbb{P}\left(Y = -1 \mid Z = z\right) - \mathbb{P}\left(Y = -1 \mid W = w\right) &= (1 - \mathbb{E}\left[Y \mid Z = z\right])/2 - (1 - \mathbb{E}\left[Y \mid W = w\right])/2 \\
&= (\mathbb{E}\left[Y \mid W = w\right] - \mathbb{E}\left[Y \mid Z = z\right])/2 \quad (A.85)
\end{aligned}$$

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

$$h(w) \begin{cases} > 0, & \text{when } \mathbb{E}\left[Y \mid W = w\right] > \mathbb{E}\left[Y \mid Z = z\right] \\ < 0, & \text{when } \mathbb{E}\left[Y \mid W = w\right] < \mathbb{E}\left[Y \mid Z = z\right] \\ \text{can be any choice}, & \text{when } \mathbb{E}\left[Y \mid W = w\right] = \mathbb{E}\left[Y \mid Z = z\right] \end{cases} \quad (A.86)$$

will maximize (A.82) with the maximum value $|\mathbb{E}\left[Y \mid W = w\right] - \mathbb{E}\left[Y \mid Z = z\right]|/2$. Remark the definition of $h(w) = \mu(w) - g(z)$, we can restate (A.86) as

$$\begin{cases} \mu(x, z) = \mu(w) > g(z), & \text{when } \mathbb{E}\left[Y \mid W = w\right] > \mathbb{E}\left[Y \mid Z = z\right] \\ \mu(x, z) = \mu(w) < g(z), & \text{when } \mathbb{E}\left[Y \mid W = w\right] < \mathbb{E}\left[Y \mid Z = z\right] \\ \text{can be any choice}, & \text{when } \mathbb{E}\left[Y \mid W = w\right] = \mathbb{E}\left[Y \mid Z = z\right] \end{cases} \quad (A.87)$$

where again $g(z) = \mathbb{E}\left[\mu(X, Z) \mid Z = z\right]$. Apparently, choosing $\mu(x, z)$ to be the true regression function $\mu^\star(x, z)$ will satisfy (A.87). Hence we show $f_{\ell_1}(\mu)$ is maximized at $\mu^\star$ with maximum value

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

which equals $\mathcal{I}_{\ell_1}$. Clearly from (A.87), $\mu^\star(x, z)$ is not the unique maximizer and any function in the set described in the following set can attains the maximum.

$$\{\mu : \mathbb{R}^p \to \mathbb{R} \mid \text{sign}\left(\mu(x, z) - \mathbb{E}\left[\mu(X, Z) \mid Z = z\right]\right) = \text{sign}\left(\mathbb{E}\left[Y \mid X = x\right] - \mathbb{E}\left[Y \mid Z = z\right]\right)\}. \quad (A.88)$$

$\square$

Proof of Theorem 3.3. According to Algorithm 2, we first denote

$$\begin{aligned}
U &:= \mu(X, Z),\quad g(z) := \mathbb{E}[\mu(X, Z) \mid Z = z], \\
G_z(u) &:= \mathbb{P}\left(U < u \mid Z = z\right),\quad F_z(u) := \mathbb{P}\left(U \leq u \mid Z = z\right). \quad (A.89)
\end{aligned}$$

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),

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

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

$$\mathbb{E}\left[G_Z(g(Z))\mathbb{1}_{\{Y=1\}} + (1 - F_Z(g(Z)))\mathbb{1}_{\{Y=-1\}}\right] = \mathbb{E}\left[\mathbb{1}_{\{Y\cdot[\mu(\tilde{X},Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]]<0\}}\right]. \tag{A.90}$$

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

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

Due to the property (A.79), 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.90). In light of the deterministic relationship in Lemma 3.2, we have $\{L_n^\alpha(\mu) \leq f_{\ell_1}(\mu)\} \subset \{L_n^\alpha(\mu) \leq \mathcal{I}_{\ell_1}\}$, hence it suffices to prove

$$\mathbb{P}(L_n^\alpha(\mu) \leq f_{\ell_1}(\mu)) \geq 1 - \alpha - O(n^{-1/2}). \tag{A.91}$$

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.91) trivially holds. Remark this includes the case when $\mu(X,Z) \in \sigma(Z)$. Otherwise, applying the classical Berry-Esseen bound (Lemma (A.1)) and Lemma (A.2) to i.i.d. bounded random variables R_i will yield (A.91), where the constant will depend on $\mathrm{Var}(R_i)$. $\qquad\square$

Proof of Theorem E.2. Similar to the proof of Theorem 3.3 in Appendix A.5, it suffices to deal with the case where $\mu(X,Z) \notin \sigma(Z)$ and prove

$$\mathbb{P}(L_n^\alpha(\mu) \leq f_{\ell_1}(\mu)) \geq 1 - \alpha + o(1). \tag{A.92}$$

Note unlike in Algorithm 2, when $g(Z_i)$ and R_i are replaced by $g^M(Z_i)$ and $R_i^{M,K}$, respectively, in Algorithm 2, we do not have $\mathbb{E}\left[R_i^{M,K}\right]$ equal to $f_{\ell_1}(\mu)/2$ anymore, where

$$f_{\ell_1}(\mu)/2 = \mathbb{E}\left[\mathbb{1}_{\{Y\cdot[\mu(\tilde{X},Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]]<0\}}\right] - \mathbb{E}\left[\mathbb{1}_{\{Y\cdot[\mu(X,Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]]<0\}}\right], \tag{A.93}$$

and $R_i^{M,K}$ is defined as

$$R_i^{M,K} = \frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{1}_{\left\{Y_i(\mu(\tilde{X}_i^{(k)},Z_i)-g^M(Z_i))<0\right\}}\right) - \mathbb{1}_{\{Y_i(\mu(X_i,Z_i))-g^M(Z_i))<0\}} \tag{A.94}$$

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 (A.92). 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 E.2. In either cases, it is reduced to prove

$$\left|\mathbb{E}\left[R_i^M\right] - \frac{f_{\ell_1}(\mu)}{2}\right| = O\left(\frac{1}{\sqrt{M}}\right) \tag{A.95}$$

43

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

$$\mathbb{E}\left[\mathbb{1}_{\{Y(\mu((\tilde{X},Z))-g^M(Z))<0\}} - \mathbb{1}_{\{Y(\mu((X,Z))-g^M(Z))<0\}}\right] \tag{A.96}$$

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 (A.93) and separately bound

$$\mathrm{II}_1 \quad := \quad \left|\mathbb{E}\left[\mathbb{1}_{\{Y(\mu(\tilde{X},Z)-g^M(Z))<0\}} - \mathbb{1}_{\{Y\cdot[\mu(\tilde{X},Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]]<0\}}\right]\right|,$$

$$\mathrm{II}_2 \quad := \quad \left|\mathbb{E}\left[\mathbb{1}_{\{Y(\mu(X,Z))-g^M(Z)<0\}} - \mathbb{1}_{\{Y\cdot[\mu(X,Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]]<0\}}\right]\right|.$$

Starting from the second term above, we rewrite it as

$$\begin{aligned}
\mathrm{II}_2 \;=\;& \left|\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{Y(\mu(X,Z))-g^M(Z))<0\}} - \mathbb{1}_{\{Y\cdot[\mu(X,Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]]<0\}}\,\big|\,Z,Y,\{\tilde{X}^{(m)}\}_{m=1}^M\right]\right]\right| \\
\leq\;& \left|\mathbb{E}\left[\mathbb{1}_{\{Y=1\}}\mathbb{E}\left[\mathbb{1}_{\{\mu(X,Z))<g^M(Z)\}} - \mathbb{1}_{\{\mu(X,Z)<\mathbb{E}[\mu(X,Z)\,|\,Z]\}}\,\big|\,Z,Y,\{\tilde{X}^{(m)}\}_{m=1}^M\right]\right]\right| \\
&+ \left|\mathbb{E}\left[\mathbb{1}_{\{Y=-1\}}\mathbb{E}\left[\mathbb{1}_{\{\mu(X,Z))>g^M(Z)\}} - \mathbb{1}_{\{\mu(X,Z)>\mathbb{E}[\mu(X,Z)\,|\,Z]\}}\,\big|\,Z,Y,\{\tilde{X}^{(m)}\}_{m=1}^M\right]\right]\right| \\
\leq\;& \mathbb{E}\left[\max\{|G_{Z,Y}(g^M(Z)) - G_{Z,Y}(g(Z))|, |F_{Z,Y}(g^M(Z)) - F_{Z,Y}(g(Z))|\}\right] \\
:=\;& \mathbb{E}[A] \tag{A.97}
\end{aligned}$$

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 A.1:

$$\left|\mathbb{P}\left(\left|\frac{\sqrt{M}|g^M(Z) - g(Z)|}{\varsigma(Z)}\right| > \sqrt{M}\delta_{Z,Y}\,\Big|\,Z,Y\right) - \bar{\Phi}(|\sqrt{M}\delta_{Z,Y}|)\right| \leq \frac{C}{\sqrt{M}} \cdot \frac{\mathbb{E}\left[|\mu^3(X,Z)|\,|\,Z\right]}{\varsigma^3(Z)} \tag{A.98}$$

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 (A.97), by considering the event $B := \{|g^M(Z) - g(Z)|/\varsigma(Z) \leq \delta_{Z,Y}\}$, we can decompose (A.97) into

$$\mathbb{E}[A] = \mathbb{E}\left[A\mathbb{1}_{\{B\}}\right] + \mathbb{E}\left[A\mathbb{1}_{\{B^c\}}\right] \tag{A.99}$$

For the first term, we have

$$\begin{aligned}
\mathbb{E}\left[A\mathbb{1}_{\{B\}}\right] \;\leq\;& \mathbb{E}\left[C_{g^M(Z),Z,Y}\,|g^M(Z) - g(Z)|\,\mathbb{1}_{\{B\}}\right] \\
=\;& \mathbb{E}\left[\mathbb{E}\left[C_{g^M(Z),Z,Y}\,|g^M(Z) - g(Z)|\,\mathbb{1}_{\{B\}}\,\Big|\,Z,Y\right]\right] \\
\leq\;& \mathbb{E}\left[C_{Z,Y}\mathbb{E}\left[|g^M(Z) - g(Z)|\,|\,Z,Y\right]\right] \\
\leq\;& \mathbb{E}\left[C_{Z,Y}\sqrt{\mathbb{E}\left[|g^M(Z) - g(Z)|^2\,\Big|\,Z,Y\right]}\right] \tag{A.100}
\end{aligned}$$

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 E.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 (A.100) 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 E.1. Now we have showed

$$\mathbb{E}\left[A\mathbb{1}_{\{B\}}\right] = O\left(\frac{1}{\sqrt{M}}\right), \tag{A.101}$$

it suffices to prove the same rate for $\mathbb{E}\left[A\mathbb{1}_{\{B^c\}}\right]$:

$$
\begin{aligned}
\mathbb{E}\left[A\mathbb{1}_{\{B^c\}}\right] &\leq 2\,\mathbb{P}\left(B^c\right) \\
&= 2\,\mathbb{E}\left[\mathbb{P}\left(B^c \mid Z\right)\right] \\
&= 2\,\mathbb{E}\left[\mathbb{P}\left(\sqrt{M}|g^M(Z) - g(Z)|/\varsigma(Z) > \sqrt{M}\delta_{Z,Y} \mid Z\right)\right] \\
&\leq 2\mathbb{E}\left[\bar{\Phi}(\left|\sqrt{M}\delta_{Z,Y}\right|) + \frac{C}{\sqrt{M}} \cdot \frac{\mathbb{E}\left[|\mu^3(X,Z)| \mid Z\right]}{\varsigma^3(Z)}\right] \\
&\leq 2\mathbb{E}\left[\frac{2}{\sqrt{2\pi}}\frac{\exp\{-M\delta_{Z,Y}^2\}}{\sqrt{M}\delta_{Z,Y}} + \frac{C}{\sqrt{M}} \cdot \frac{\mathbb{E}\left[|\mu^3(X,Z)| \mid Z\right]}{\varsigma^3(Z)}\right]
\end{aligned}
$$

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 (A.98) 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 E.1, the following holds

$$\mathbb{E}\left[A\mathbb{1}_{\{B^c\}}\right] = O\left(\frac{1}{\sqrt{M}}\right). \tag{A.102}$$

Finally we prove

$$\left|\mathbb{E}\left[\mathbb{1}_{\{Y(\mu(X,Z))-g^M(Z))<0\}} - \mathbb{1}_{\{Y\cdot[\mu(X,Z)-\mathbb{E}[\mu(X,Z)\mid Z]]<0\}}\right]\right| = O\left(\frac{1}{\sqrt{M}}\right).$$

Regarding the term

$$\mathrm{II}_1 = \left|\mathbb{E}\left[\mathbb{1}_{\{Y(\mu((\tilde{X},Z))-g^M(Z))<0\}} - \mathbb{1}_{\{Y\cdot[\mu(\tilde{X},Z)-\mathbb{E}[\mu(X,Z)\mid Z]]<0\}}\right]\right|$$

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 (E.1). Hence it suffices to notice the following derivations for $F_z(u)$:

$$
\begin{aligned}
F_z(u) = \mathbb{P}\left(U \leq u \mid Z = z\right) &= \mathbb{E}_{Y|Z=z}\left[\mathbb{P}\left(U \leq u \mid Z = z, Y\right) \mid Z = z\right] \\
&= \mathbb{E}_{Y|Z=z}\left[F_{z,Y}(u) \mid Z = z\right],
\end{aligned}
$$

and similarly for $G_z(u)$. Together with the definition of $C_{u,z,y}$ and (a) in Assumption E.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 (A.92) i.e.

$$\mathbb{P}\left(L_n^\alpha(\mu) \leq f_{\ell_1}(\mu)\right) \geq 1 - \alpha - O\left(\frac{1}{\sqrt{n}}\right), \tag{A.103}$$

when assuming $n^2/M = O(1)$. Similarly as the proof of Theorem 2.3, the above holds when applying Lemma A.2 and checking the two conditions. Since R_i are bounded by definition, the second condition can

45

be easily verified using the similar strategy as in the derivations for (A.13). Then it suffices to establish the following Berry-Esseen bound:

$$\Delta := \sup_{t \in \mathbb{R}} \left| \mathbb{P} \left(\sqrt{n} \left(\frac{\bar{R} - f_{\ell_1}(\mu)/2}{\sqrt{\mathrm{Var}\left(R_i^M\right)}} \right) \leq t \right) - \Phi(t) \right| = O\left(\frac{1}{\sqrt{n}} \right).$$

Notice that

$$
\begin{aligned}
\Delta &= \sup_{t \in \mathbb{R}} \left| \mathbb{P} \left(\sqrt{n} \left(\frac{\bar{R} - \mathbb{E}\left[R_i^M\right]}{\sqrt{\mathrm{Var}\left(R_i^M\right)}} \right) \leq t + \sqrt{n} \frac{\left(\mathbb{E}\left[R_i^M\right] - f_{\ell_1}(\mu)/2\right)}{\sqrt{\mathrm{Var}\left(R_i^M\right)}} \right) - \Phi(t) \right| \\
&\leq \sup_{t \in \mathbb{R}} \left| \mathbb{P} \left(\sqrt{n} \left(\frac{\bar{R} - \mathbb{E}\left[R_i^M\right]}{\sqrt{\mathrm{Var}\left(R_i^M\right)}} \right) \leq t \right) - \Phi(t) \right| + \sup_{t \in \mathbb{R}} \left| \Phi \left(t + \sqrt{n} \frac{\left(\mathbb{E}\left[R_i^M\right] - f_{\ell_1}(\mu)/2\right)}{\sqrt{\mathrm{Var}\left(R_i^M\right)}} \right) - \Phi(t) \right| \\
&:= \Delta_1 + \Delta_2
\end{aligned}
$$

Since the first derivative of $\Phi(t)$ is bounded by $1/\sqrt{2\pi}$ over $\mathbb{R}$, we have

$$\Delta_2 \leq \frac{\sqrt{n}}{\sqrt{2\pi}} \frac{|f_{\ell_1}(\mu)/2 - \mathbb{E}\left[R_i^M\right]|}{\sqrt{\mathrm{Var}\left(R_i^M\right)}}$$

by Taylor expansion. Note that as a result of (A.95), we have

$$\sqrt{n} |\mathbb{E}\left[R_i^M\right] - f_{\ell_1}(\mu)/2| = O(1/\sqrt{n}). \tag{A.104}$$

Then it suffices to prove $\Delta_1 = O(1/\sqrt{n})$ and $\mathrm{Var}\left(R_i^M\right) > 0$. $\Delta_1 = O(1/\sqrt{n})$ holds when applying the triangular array version of the Berry Esseen bound in Lemma A.1 (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

$$
\begin{aligned}
&\mathrm{Var}\left(R_i^{M,K}\right) \\
&= \mathbb{E}\left[\mathrm{Var}\left(R_i^{M,K} \mid Z_i, Y_i, \{\tilde{X}_i^{(m)}\}_{m=1}^M\right)\right] + \mathrm{Var}\left(\mathbb{E}\left[R_i^{M,K} \mid Z_i, Y_i, \{\tilde{X}_i^{(m)}\}_{m=1}^M\right]\right) \\
&\geq \mathbb{E}\left[\mathrm{Var}\left(R_i^{M,K} \mid Z_i, Y_i, \{\tilde{X}_i^{(m)}\}_{m=1}^M\right)\right] \\
&= \mathbb{E}\left[\mathrm{Var}\left(\frac{1}{K} \sum_{k=1}^K \left(\mathbb{1}_{\left\{Y_i(\mu(\tilde{X}_i^{(k)}, Z_i)) - g^M(Z_i)) < 0\right\}} \right) - \mathbb{1}_{\left\{Y_i(\mu(X_i, Z_i)) - g^M(Z_i)) < 0\right\}} \,\middle|\, Z_i, Y_i, \{\tilde{X}_i^{(m)}\}_{m=1}^M \right)\right] \\
&\geq \mathbb{E}\left[\mathrm{Var}\left(\mathbb{1}_{\left\{Y_i(\mu(X_i, Z_i)) - g^M(Z_i)) < 0\right\}} \,\middle|\, Z_i, Y_i, \{\tilde{X}_i^{(m)}\}_{m=1}^M \right)\right] := \sigma_M^2 \tag{A.105}
\end{aligned}
$$

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 (A.105), 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 (A.97) for bounding the term II_2, we can similarly bound $|\sigma_M^2 - \sigma_0^2|$ by the following quantity:

$$\begin{aligned}
|\sigma_M^2 - \sigma_0^2| &\leq \mathbb{E}\left[3\max\{|G_{Z,Y}(g^M(Z)) - G_{Z,Y}(g(Z))|\,,|F_{Z,Y}(g^M(Z)) - F_{Z,Y}(g(Z))|\}\right] \\
&= 3\mathbb{E}[A] = 3(\mathbb{E}\left[A\mathbb{1}_{\{B\}}\right] + \mathbb{E}\left[A\mathbb{1}_{\{B^c\}}\right]) = O\left(\frac{1}{\sqrt{M}}\right).
\end{aligned}$$

where the last equality holds due to the results (A.101) and (A.102) from previous derivations for the term II_2. Finally we conclude (A.103), which immediately implies a weaker version of the result, i.e. the statement of Theorem E.2. $\qquad\square$

A.6 Proofs in Section 3.2

Proof of Theorem 3.4. When $\mathcal{T}$ is degenerate or $\mu(X) \in \sigma(Z)$, we immediately have $L_n^{\alpha,\mathcal{T}}(\mu) = 0$ according to Algorithm 3, which implies the coverage validity. Below we focus on the non-trivial case. Due to the deterministic relationship

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

it suffices to prove

$$\mathbb{P}_P\left(L_n^{\alpha,\mathcal{T}}(\mu) \leq f_n^{\mathcal{T}}(\mu)\right) \geq 1 - \alpha - o(1). \tag{A.106}$$

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

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

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

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

Using similar strategies in the proof of Theorem 2.3, we can show the above holds under the moment conditions $\mathbb{E}\left[Y^2\right], \mathbb{E}\left[\mu^4(X)\right] < \infty$ 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.106) holds immediately. When $\mathrm{Var}\left(V_m\right) = 0$, 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 $\mathrm{Var}\left(R_m\right), \mathrm{Var}\left(V_m\right) > 0$. 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

$$\begin{aligned}
U_{m1} &= R_m - \mathbb{E}\left[Y_i\left(\mu(X_i,Z_i) - \mathbb{E}\left[\mu(X_i,Z_i)\,|\,\boldsymbol{Z}_m,\boldsymbol{T}_m\right]\right)\right], &\tag{A.107}\\
U_{m2} &= V_m - \mathbb{E}\left[\mathrm{Var}\left(\mu(X_i,Z_i)\,|\,\boldsymbol{Z}_m,\boldsymbol{T}_m\right)\right] &\tag{A.108}
\end{aligned}$$

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

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

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 $\mathbb{E}[h^{\mathcal{T}}(W_i) \mid \boldsymbol{Z}_m, \boldsymbol{T}_m] = 0$ and the tower property of conditional expectation. The last inequality holds by the tower property of conditional expectation and the fact that $h^{\mathcal{T}}(W_i) \in \sigma(\boldsymbol{X}_m, \boldsymbol{Z}_m)$. Let $T = \bar{R}/\bar{V}$, then $T - f_n^{\mathcal{T}}(\mu)$ can be rewritten as

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

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:

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

when $x_2 > -\mathbb{E}[(h^{\mathcal{T}}(W_i)^2)]$ and is set to be $\frac{\mathbb{E}[Y_i h^{\mathcal{T}}(W_i)]}{\sqrt{\mathbb{E}[(h^{\mathcal{T}}(W_i)^2)]}}$ 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

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

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 $\text{Var}(R_m), \text{Var}(V_m) > 0$. Also notice s^2 is a consistent estimator of $\tilde{\sigma}^2$, then by the argument of Slutsky's Theorem, (A.106) is established.

$\square$

A.7 Proofs in Section F

Lemma A.6. *Under the moment conditions $\mathbb{E}[\mu^2(X, Z)], \mathbb{E}[(\mu^\star)^2(X, Z)] < \infty$, we can quantify the gap between $f(\mu)$ and $f_n^{\mathcal{T}}(\mu)$ as below.*

$$
f(\mu) - f_n^{\mathcal{T}}(\mu) = O(\max\{\text{II}(\mu), \text{II}(\mu^\star)\}) \tag{A.109}
$$

where $\text{II}(\mu) = \mathbb{E}_{\boldsymbol{Z}}[\text{Var}_{\boldsymbol{T}|\boldsymbol{Z}}(\mathbb{E}[\mu(X_i, Z_i) \mid \boldsymbol{Z}, \boldsymbol{T}])]$.

When this lemma is used in the proof of Proposition F.1 and F.2, the nature 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 A.6. Recall the definition of $f(\mu)$ and $f_n^{\mathcal{T}}(\mu)$,

$$f(\mu) = \frac{\mathbb{E}\left[\text{Cov}(\mu^\star(X,Z),\mu(X,Z)\,|\,Z)\right]}{\sqrt{\mathbb{E}\left[\text{Var}(\mu(X,Z)\,|\,Z)\right]}}, \tag{A.110}$$

$$f_n^{\mathcal{T}}(\mu) = \frac{\mathbb{E}\left[\text{Cov}(\mu^\star(X_i,Z_i),\mu(X_i,Z_i)\,|\,\boldsymbol{Z},\boldsymbol{T})\right]}{\sqrt{\mathbb{E}\left[\text{Var}(\mu(X_i,Z_i)\,|\,\boldsymbol{Z},\boldsymbol{T})\right]}}, \tag{A.111}$$

then denote $W_i = (X_i, Z_i)$, $h(W_i) := \mu(W_i) - \mathbb{E}\left[\mu(W_i)\,|\,Z_i\right]$, $h^{\mathcal{T}}(W_i) := \mu^\star(W_i) - \mathbb{E}\left[\mu^\star(W_i)\,|\,\boldsymbol{Z},\boldsymbol{T}\right]$ and assume $\mathbb{E}\left[h^2(W_i)\right] = 1$ without loss of generality. First notice a simple fact $|\frac{a}{b} - \frac{c}{d}| = \frac{|ad-bc|}{bd} = \frac{|ad-cd+cd-bc|}{bd} \leq \frac{|a-c|}{b} + \frac{c|b-d|}{bd}$ for $a,b,c,d > 0$, then let the numerator and denominator of $f(\mu)$ in (A.110) to be a, b respectively (similarly denote c, d for $f_n^{\mathcal{T}}(\mu)$ in (A.111)). And we have

$$\max\{\frac{1}{b},\ \frac{c}{bd}\} \leq 1 + f_n^{\mathcal{T}}(\mu) \leq 1 + f_n^{\mathcal{T}}(\mu^\star) \leq 1 + f(\mu^\star) \leq 1 + \mathbb{E}\left[(\mu^\star)^2(X,Z)\right] < \infty,$$

hence it suffices to bound $|a - c|$ and $|b - d|$. First we have the following

$$\begin{aligned}
a - c &= \mathbb{E}\left[\text{Cov}\left(\mu^\star(W_i),\mu(W_i)\,|\,\boldsymbol{Z}\right)\right] - \mathbb{E}\left[\text{Cov}\left(\mu^\star(W_i),\mu(W_i)\,|\,\boldsymbol{Z},\boldsymbol{T}\right)\right] \\
&= \mathbb{E}\left[\text{Cov}\left(\mathbb{E}\left[\mu^\star(W_i)\,|\,\boldsymbol{Z},\boldsymbol{T}\right],\mathbb{E}\left[\mu(W_i)\,|\,\boldsymbol{Z},\boldsymbol{T}\right]\,|\,\boldsymbol{Z}\right)\right] \\
&= \mathbb{E}_{\boldsymbol{Z}}\left[\text{Cov}_{\boldsymbol{T}|\boldsymbol{Z}}\left(\mathbb{E}\left[\mu^\star(W_i)\,|\,\boldsymbol{Z},\boldsymbol{T}\right],\mathbb{E}\left[\mu(W_i)\,|\,\boldsymbol{Z},\boldsymbol{T}\right]\right)\right].
\end{aligned} \tag{A.112}$$

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 $\text{Cov}\left(\mu^\star(W_i),\mu(W_i)\,|\,\boldsymbol{Z}\right)$ 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

$$|a - c| \leq \sqrt{\mathbb{E}_{\boldsymbol{Z}}\left[\text{Var}_{\boldsymbol{T}|\boldsymbol{Z}}\left(\mathbb{E}\left[\mu^\star(W_i)\,|\,\boldsymbol{Z},\boldsymbol{T}\right]\right)\right]}\sqrt{\mathbb{E}_{\boldsymbol{Z}}\left[\text{Var}_{\boldsymbol{T}|\boldsymbol{Z}}\left(\mathbb{E}\left[\mu(W_i)\,|\,\boldsymbol{Z},\boldsymbol{T}\right]\right)\right]} \tag{A.113}$$

Regarding the term $|b - d|$, we have

$$\begin{aligned}
|b - d| &= \left|\sqrt{\mathbb{E}\left[h^2(W_i)\right]} - \sqrt{\mathbb{E}\left[(h^{\mathcal{T}})^2(W_i)\right]}\right| \\
&= \frac{\left|\mathbb{E}\left[h^2(W_i)\right] - \mathbb{E}\left[(h^{\mathcal{T}})^2(W_i)\right]\right|}{\sqrt{\mathbb{E}\left[h^2(W_i)\right]} + \sqrt{\mathbb{E}\left[(h^{\mathcal{T}})^2(W_i)\right]}} \\
&\leq \frac{\left|\mathbb{E}\left[h^2(W_i)\right] - \mathbb{E}\left[(h^{\mathcal{T}})^2(W_i)\right]\right|}{\sqrt{\mathbb{E}\left[h^2(W_i)\right]}} \\
&\leq \mathbb{E}\left[\text{Var}\left(\mu(W_i)\,|\,\boldsymbol{Z}\right)\right] - \mathbb{E}\left[\text{Var}\left(\mu(W_i)\,|\,\boldsymbol{Z},\boldsymbol{T}\right)\right] \\
&= \mathbb{E}_{\boldsymbol{Z}}\left[\text{Var}_{\boldsymbol{T}|\boldsymbol{Z}}\left(\mathbb{E}\left[\mu(W_i)\,|\,\boldsymbol{Z},\boldsymbol{T}\right]\right)\right]
\end{aligned} \tag{A.114}$$

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 $\text{Var}\left(\mu(W_i)\,|\,\boldsymbol{Z}\right)$ then getting the second term of line 4 cancelled out. Finally, combining (A.113) and (A.114) establishes the bound in (A.109). $\square$

A.7.1 Proofs in Section F.2

Proof of Proposition F.1. Throughout the proof, the nature 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 A.6, it suffices to bound the term $\text{II}(\mu)$, $\text{II}(\mu^\star)$ in (A.109).

(ii) we bound $\text{II}(\mu)$, $\text{II}(\mu^\star)$ with the same strategy. Specifically, we will show

$$\text{II}(\mu) = O\left(\mathbb{E}_{Z_i}\left[\mathbb{E}_F\left[\mu^2(W_i)\right]\mathbb{E}\left[h_{ii}\mid Z_i\right]\right]\right)$$

and similarly for $\text{II}(\mu^\star)$ 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 $\mathbb{E}\left[h_{ii}\mid Z_i\right]$ above, we can carefully bound it by $1/(n-1) + \mathbb{E}\left[\boldsymbol{\Xi}\mid Z_i\right]$, where $\boldsymbol{\Xi}$ is defined in (A.124).

(iv) Simply expanding $\mathbb{E}\left[\boldsymbol{\Xi}\mid Z_i\right]$ into three terms: $\text{III}_1, \text{III}_2, \text{III}_3$, which are defined in (A.125), (A.126) and (A.126), we will show $\text{III}_2 = 0$ and figure out the stochastic representation of $\text{III}_1, \text{III}_3$, 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 (F.1).

Having proved Lemma A.6, now we directly start with step (ii). Notice the following

$$
\begin{aligned}
\text{II}(\mu) &= \mathbb{E}_{\boldsymbol{Z}}\left[\text{Var}_{\boldsymbol{T}|\boldsymbol{Z}}\left(\mathbb{E}\left[\mu(W_i)\mid \boldsymbol{Z}, \boldsymbol{T}\right]\right)\right] \\
&= \mathbb{E}_{\boldsymbol{Z}}\left[\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\left(\mathbb{E}_F\left[\mu(W_i)\right] - \mathbb{E}_{F_{\boldsymbol{T}}}\left[\mu(W_i)\right]\right)^2\right]\right] \\
&= \mathbb{E}_{\boldsymbol{Z}}\left[\text{Var}_F\left(\mu(W_i)\right)\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\frac{\left(\mathbb{E}_F\left[\mu(W_i)\right] - \mathbb{E}_{F_{\boldsymbol{T}}}\left[\mu(W_i)\right]\right)^2}{\text{Var}_F\left(\mu(W_i)\right)}\right]\right] \\
&\leq \mathbb{E}_{\boldsymbol{Z}}\left[\text{Var}_F\left(\mu(W_i)\right)\min\left\{\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\chi^2(F_{\boldsymbol{T}}||F)\right], 2\right\}\right] \quad\quad\quad (A.115)
\end{aligned}
$$

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 $\text{Var}_F\left(\mu(W_i)\right) \in \sigma(\boldsymbol{Z})$. Regarding the last inequality, we make use of the variational representation of χ^2-divergence:

$$\chi^2(P||Q) = \sup_{\mu}\frac{(\mathbb{E}_P(\mu) - \mathbb{E}_Q(\mu))^2}{\text{Var}_Q(\mu)}$$

and the fact that

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\frac{\left(\mathbb{E}_F\left[\mu(W_i)\right] - \mathbb{E}_{F_{\boldsymbol{T}}}\left[\mu(W_i)\right]\right)^2}{\text{Var}_F\left(\mu(W_i)\right)}\right] \\
&\leq \frac{\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\mathbb{E}_F\left[\mu^2(W_i)\right]\right] + \mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\mathbb{E}_{F_{\boldsymbol{T}}}\left[\mu^2(W_i)\right]\right] - 2\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\mathbb{E}_{F_{\boldsymbol{T}}}\left[\mu(W_i)\right]\mathbb{E}_F\left[\mu(W_i)\right]\right]}{\text{Var}_F\left(\mu(W_i)\right)} \\
&= \frac{\mathbb{E}_F\left[\mu^2(W_i)\right] + \mathbb{E}_F\left[\mu^2(W_i)\right] - 2(\mathbb{E}_F\left[\mu(W_i)\right])^2}{\text{Var}_F\left(\mu(W_i)\right)} \\
&= \frac{2\text{Var}_F\left(\mu(W_i)\right)}{\text{Var}_F\left(\mu(W_i)\right)} = 2
\end{aligned}
$$

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 $\mathbb{E}_F\left[\mu(W_i)\right] \in \sigma(\boldsymbol{Z})$. Denote $u_i = (1, Z_i)^\top$ and the following n by p matrix by $\boldsymbol{U}$:

$$\boldsymbol{U} = \begin{pmatrix} u_1^\top \\ \vdots \\ u_n^\top \end{pmatrix} = (\boldsymbol{1}, \boldsymbol{Z}) \quad\quad\quad (A.116)$$

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

$$\left(\begin{array}{c} X_i \\ \boldsymbol{T} \end{array} \right) \sim \mathcal{N}\left(\left[\begin{array}{c} (1, Z_i)\gamma \\ \boldsymbol{U}^\top \boldsymbol{U}\gamma \end{array} \right], \sigma^2 \left[\begin{array}{cc} 1 & e_i^\top \boldsymbol{U} \\ \boldsymbol{U}^\top e_i^\top & \boldsymbol{U}^\top \boldsymbol{U} \end{array} \right] \right) \tag{A.117}$$

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:

$$F_{\boldsymbol{T}} \quad : X_i \mid \boldsymbol{Z}, \boldsymbol{T} \sim \mathcal{N}\left(e_i^\top \boldsymbol{H}\boldsymbol{X}, \sigma^2(1 - h_{ii}) \right)$$
$$F \quad : X_i \mid \boldsymbol{Z} \sim \mathcal{N}\left((1, Z_i)\gamma, \sigma^2 \right)$$

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,

$$\begin{aligned} e_i^\top \boldsymbol{H}\boldsymbol{X} - (1, Z_i)\gamma &= e_i^\top \boldsymbol{U}\hat{\gamma}^{OLS} - (1, Z_i)\gamma \\ &= e_i^\top \boldsymbol{U}(\hat{\gamma}^{OLS} - \gamma) \sim \mathcal{N}(0, \sigma^2 h_{ii}) \end{aligned} \tag{A.118}$$

Since $2\sigma^2 > \sigma^2(1 - h_{ii})$, applying Lemma A.7 yields the following

$$\begin{aligned} \chi^2(F_{\boldsymbol{T}}\|F) &= \frac{1}{2}\left[\frac{1}{\sqrt{1 - h_{ii}^2}} \exp\left\{ \frac{(e_i^\top \boldsymbol{H}\boldsymbol{X} - (1, Z_i)\gamma)^2}{\sigma^2(1 + h_{ii})} \right\} - 1 \right] \\ &\leq \frac{1}{\sqrt{1 - h_{ii}}} \exp\left\{ \frac{(e_i^\top \boldsymbol{H}\boldsymbol{X} - (1, Z_i)\gamma)^2}{\sigma^2(1 + h_{ii})} \right\} - 1 \\ &= \frac{1}{\sqrt{1 - h_{ii}}} \exp\left\{ \frac{h_{ii}G^2}{1 + h_{ii}} \right\} - 1 \end{aligned} \tag{A.119}$$

where $G \sim \mathcal{N}(0, 1)$ is independent from $\boldsymbol{X}$ and the last equality holds due to (A.118). Plugin (A.119) back to (A.115), we have

$$\begin{aligned} \mathrm{II}(\mu) &\leq \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_F\left(\mu(W_i)\right) \min\left\{ \mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\chi^2(F_{\boldsymbol{T}}\|F)\right], 2 \right\} \right] \\ &\leq \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_F\left(\mu(W_i)\right) \min\left\{ \mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\frac{1}{\sqrt{1 - h_{ii}}} \exp\left\{ \frac{h_{ii}G^2}{1 + h_{ii}} \right\} - 1 \right], 2 \right\} \right] \end{aligned}$$

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

$$
\begin{aligned}
\mathbb{E}_{T|Z}\left[\frac{1}{\sqrt{1-h_{ii}}}\exp\left\{\frac{h_{ii}G^2}{1+h_{ii}}\right\}\right] - 1 &= \frac{1}{\sqrt{1-h_{ii}}\sqrt{1-2h_{ii}/(1+h_{ii})}} - 1 \\
&= \frac{\sqrt{1+h_{ii}}}{1-h_{ii}} - 1 \\
&\leq \frac{1+h_{ii}}{1-h_{ii}} - 1 \\
&\leq 4h_{ii}
\end{aligned}
$$

hence we can bound $\mathrm{II}(\mu)$ by the summation of the following two terms:

$$
\mathrm{II}_1 := \mathbb{E}_Z\left[\mathrm{Var}_F\left(\mu(W_i)\right)\mathbb{1}_{\{E\}}\cdot 4h_{ii}\right], \quad \mathrm{II}_2 := \mathbb{E}_Z\left[\mathrm{Var}_F\left(\mu(W_i)\right)\mathbb{1}_{\{E^c\}}\cdot 2\right]
$$

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 $\mathrm{Var}_F\left(\mu(W_i)\right) \leq \mathbb{E}_F\left[\mu^2(W_i)\right] \in \sigma(Z_i)$ Regarding II_2, we have

$$
\begin{aligned}
\mathrm{II}_2 &= 2\,\mathbb{E}_Z\left[\mathrm{Var}_F\left(\mu(W_i)\right)\mathbb{1}_{\{E^c\}}\right] \\
&= 2\,\mathbb{E}_Z\left[\mathrm{Var}_F\left(\mu(W_i)\right)\mathbb{E}\left[\mathbb{1}_{\{E^c\}}\,|\,Z_i\right]\right] \\
&\leq 2\,\mathbb{E}_{Z_i}\left[\mathbb{E}_F\left[\mu^2(W_i)\right]\mathbb{P}\left(h_{ii}\geq\frac{1}{2}\,|\,Z_i\right)\right] \\
&\leq 4\,\mathbb{E}_{Z_i}\left[\mathbb{E}_F\left[\mu^2(W_i)\right]\mathbb{E}\left[h_{ii}\,|\,Z_i\right]\right]
\end{aligned}
$$

where the second equality comes from the tower property of conditional expectation and $\mathrm{Var}_F\left(\mu(W_i)\right) \in \sigma(Z_i)$ and the last inequality holds due to Markov's inequality. Now we can compactly write down the following bound for $\mathrm{II}(\mu)$,

$$
\mathrm{II}(\mu) \leq \mathrm{II}_1 + \mathrm{II}_2 \leq 8\,\mathbb{E}_{Z_i}\left[\mathbb{E}_F\left[\mu^2(W_i)\right]\mathbb{E}\left[h_{ii}\,|\,Z_i\right]\right], \tag{A.120}
$$

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 (A.116), 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 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

$$
h_{ii} = u_i^\top\boldsymbol{A}^{-1}u_i - \frac{(u_i^\top\boldsymbol{A}^{-1}u_i)^2}{1 + u_i^\top\boldsymbol{A}^{-1}u_i} \leq u_i^\top\boldsymbol{A}^{-1}u_i. \tag{A.121}
$$

Since $\boldsymbol{A}$ also involves the unit vector $\mathbf{1}_{n-1}$, it is easier when we first project $\boldsymbol{Z}_{-i}$ on $\mathbf{1}_{n-1}$ then work with the orthogonal complement. Bearing this idea in mind, we denote $\boldsymbol{\Omega} = (\mathbf{1}_{n-1}, \boldsymbol{Z}_{-i})$ which is a $n-1$ by p matrix, then rewrite $\boldsymbol{A}$ as

$$\boldsymbol{A} = \boldsymbol{\Omega}^\top \boldsymbol{\Omega} = \begin{pmatrix} \mathbf{1}_{n-1}^\top \mathbf{1}_{n-1} & \mathbf{1}_{n-1}^\top \boldsymbol{Z}_{-i} \\ \boldsymbol{Z}_{-i}^\top \mathbf{1}_{n-1} & \boldsymbol{Z}_{-i}^\top \boldsymbol{Z}_{-i} \end{pmatrix}$$

where $\mathbf{I}_{n-1}$ is the $(n-1)$ dimensional identity matrix. Denote

$$\overline{\boldsymbol{Z}}_{-i} := \frac{1}{n-1} \sum_{m \neq i} Z_m = \frac{1}{n-1} \mathbf{1}_{n-1}^\top \boldsymbol{Z}_{-i} \quad \boldsymbol{\Gamma} := \begin{pmatrix} 1 & -\overline{\boldsymbol{Z}}_{-i} \\ \mathbf{0} & \mathbf{I}_{n-1} \end{pmatrix}, \tag{A.122}$$

we have

$$\begin{aligned} \boldsymbol{\Omega}\boldsymbol{\Gamma} = (\mathbf{1}_{n-1}, \boldsymbol{Z}_{-i})\boldsymbol{\Gamma} &= (\mathbf{1}_{n-1}, \boldsymbol{Z}_{-i} - \mathbf{1}_{n-1}\overline{\boldsymbol{Z}}_{-i}) \\ &= (\mathbf{1}_{n-1}, (\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}_{-i}). \end{aligned}$$

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

$$(\boldsymbol{\Omega}\boldsymbol{\Gamma})^\top \boldsymbol{\Omega}\boldsymbol{\Gamma} = \begin{pmatrix} n-1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z}_{-i}^\top (\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}_{-i} \end{pmatrix}$$

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

$$u_i^\top \boldsymbol{\Gamma} = (1, Z_i)\boldsymbol{\Gamma} = (1, Z_i - \overline{\boldsymbol{Z}}_{-i}). \tag{A.123}$$

Combining (A.122) with (A.123) yields the following

$$\begin{aligned} u_i^\top \boldsymbol{A}^{-1} u_i &= u_i^\top (\boldsymbol{\Omega}^\top \boldsymbol{\Omega})^{-1} u_i \\ &= u_i^\top \boldsymbol{\Gamma}((\boldsymbol{\Omega}\boldsymbol{\Gamma})^\top \boldsymbol{\Omega}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^\top w_i \\ &= \frac{1}{n-1} + (Z_i - \overline{\boldsymbol{Z}}_{-i})(\boldsymbol{Z}_{-i}^\top (\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}_{-i})^{-1}(Z_i - \overline{\boldsymbol{Z}}_{-i})^\top, \end{aligned}$$

which together with (A.121) implies $\mathbb{E}\left[h_{ii} \mid Z_i\right] \leq \mathbb{E}\left[u_i^\top \boldsymbol{A}^{-1} u_i \mid Z_i\right] = 1/(n-1) + \mathbb{E}\left[\boldsymbol{\Xi} \mid Z_i\right]$, where

$$\boldsymbol{\Xi} = (Z_i - \overline{\boldsymbol{Z}}_{-i})(\boldsymbol{Z}_{-i}^\top (\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}_{-i})^{-1}(Z_i - \overline{\boldsymbol{Z}}_{-i})^\top. \tag{A.124}$$

As the problem has been reduced to calculating $\mathbb{E}\left[\boldsymbol{\Xi} \mid Z_i\right]$, we arrive at the step (iv) now. Write $(\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}_{-i}) = (\boldsymbol{Z}_i - \boldsymbol{v}_0) - (\overline{\boldsymbol{Z}}_{-i} - \boldsymbol{v}_0)$, where $\boldsymbol{v}_0$ is the mean of Gaussian random variable Z, we can expand $\mathbb{E}\left[\boldsymbol{\Xi} \mid Z_i\right] = \mathrm{III}_1 + \mathrm{III}_2 + \mathrm{III}_3$, where

$$\mathrm{III}_1 = (\boldsymbol{Z}_i - \boldsymbol{v}_0)\mathbb{E}\left[(\boldsymbol{Z}_{-i}^\top (\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}_{-i})^{-1} \mid Z_i\right](\boldsymbol{Z}_i - \boldsymbol{v}_0)^\top \tag{A.125}$$

$$\mathrm{III}_2 = -2(\boldsymbol{Z}_i - \boldsymbol{v}_0)\mathbb{E}\left[(\boldsymbol{Z}_{-i}^\top (\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}_{-i})^{-1}(\overline{\boldsymbol{Z}}_{-i} - \boldsymbol{v}_0)^\top \mid Z_i\right] \tag{A.126}$$

$$\mathrm{III}_3 = \mathbb{E}\left[(\overline{\boldsymbol{Z}}_{-i} - \boldsymbol{v}_0)(\boldsymbol{Z}_{-i}^\top (\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}_{-i})^{-1}(\overline{\boldsymbol{Z}}_{-i} - \boldsymbol{v}_0)^\top \mid Z_i\right] \tag{A.127}$$

Below we are going to show $\mathrm{III}_2 = 0$ and derive $\mathrm{III}_1, \mathrm{III}_3$ carefully. Regarding the term III_1, we exactly write down its stochastic representation. Under the state Gaussian model, we have $\boldsymbol{Z}_{-i}^\top \sim \mathcal{N}\left(\boldsymbol{v}_0\mathbf{1}_{n-1}^\top, \mathbf{I}_{n-1} \otimes \boldsymbol{\Sigma}_0\right)$, then $(\boldsymbol{Z}_{-i}^\top (\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}_{-i})^{-1}$ follows an inverse Wishart distribution i.e.

$$(\boldsymbol{Z}_{-i}^\top (\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}_{-i})^{-1} \sim \mathcal{W}_{p-1}^{-1}(\boldsymbol{\Sigma}_0^{-1}, n-2)$$

and $\boldsymbol{Z}_{-i} \perp\!\!\!\perp \boldsymbol{Z}_i$, hence we can calculate

$$\mathbb{E}\left[(\boldsymbol{Z}_{-i}^\top (\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}_{-i})^{-1} \mid \boldsymbol{Z}_i\right] = \frac{\boldsymbol{\Sigma}_0^{-1}}{n-p-2}.$$

Plug in the above equation into (A.125), we have

$$\text{III}_1 = (\boldsymbol{Z}_i - \boldsymbol{v}_0)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{Z}_i - \boldsymbol{v}_0)^\top = \frac{\boldsymbol{\Phi}}{n-p-2}, \quad \text{where} \quad \boldsymbol{\Phi} \sim \chi_{p-1}^2, \; \boldsymbol{\Phi} \perp\!\!\!\perp \boldsymbol{Z}_{\text{-}i}. \tag{A.128}$$

Regarding the term III_2 in (A.126), we first denote $\boldsymbol{Z} = \boldsymbol{Z}_{\text{-}i} - \boldsymbol{1}_{n-1}\boldsymbol{v}_0$ and notice

$$\boldsymbol{Z} \sim \mathcal{N}\left(\boldsymbol{0}, \mathbf{I}_{n-1} \otimes \boldsymbol{\Sigma}_0\right), \quad \boldsymbol{1}_{n-1}^\top \boldsymbol{Z} = (n-1)(\overline{\boldsymbol{Z}}_{\text{-}i} - \boldsymbol{v}_0), \tag{A.129}$$

then rewrite III_2 as below

$$\text{III}_2 = -2(\boldsymbol{Z}_i - \boldsymbol{v}_0)\mathbb{E}\left[((\boldsymbol{Z} + \boldsymbol{1}_{n-1}\boldsymbol{v}_0)^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})(\boldsymbol{Z} + \boldsymbol{1}_{n-1}\boldsymbol{v}_0))^{-1}\frac{(\boldsymbol{1}_{n-1}^\top \boldsymbol{Z})^\top}{n-1}\right]$$

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 $(\boldsymbol{1}_{n-1}\boldsymbol{v}_0)^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1}) = \boldsymbol{0}$, we can simplify further

$$\text{III}_2 = -\frac{2}{n-1}(\boldsymbol{Z}_i - \boldsymbol{v}_0)\mathbb{E}\left[(\boldsymbol{Z}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z})^{-1}(\boldsymbol{1}_{n-1}^\top \boldsymbol{Z})^\top\right] \tag{A.130}$$

Notice in the above equation, $\boldsymbol{Z}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})$ is the orthogonal complement of $\boldsymbol{Z}^\top \boldsymbol{1}_{n-1}$, which implies independence under the Gaussian distribution assumption, which we will now use to prove the expectation in (A.130) equals zero. Formally, we first have $(\boldsymbol{Z}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1}), \boldsymbol{Z}^\top \boldsymbol{1}_{n-1})$ are multivariate Gaussian. Introducing the vectorization of matrix and the Kronecker product, we can express in the following way:

$$\text{vec}(\boldsymbol{Z}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})) = (\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1}) \otimes \mathbf{I}_{p-1}\text{vec}(\boldsymbol{Z}^\top), \quad \text{vec}(\boldsymbol{Z}^\top) = \boldsymbol{1}_{n-1} \otimes \mathbf{I}_{p-1}\text{vec}(\boldsymbol{Z}^\top).$$

Now we are ready to calculate the covariance

$$\text{Cov}\left(\text{vec}(\boldsymbol{Z}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})), \text{vec}(\boldsymbol{Z}^\top \boldsymbol{1}_{n-1})\right)$$
$$= \quad ((\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1}) \otimes \mathbf{I}_{p-1})(\mathbf{I}_{n-1} \otimes \boldsymbol{\Sigma}_0)(\boldsymbol{1}_{n-1} \otimes \mathbf{I}_{p-1})^\top$$
$$= \quad ((\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\mathbf{I}_{n-1}\boldsymbol{1}_{n-1}) \otimes (\mathbf{I}_{p-1}\boldsymbol{\Sigma}_0\mathbf{I}_{p-1}) = \boldsymbol{0}$$

where in above equalities we use the fact $\text{Var}\left(\text{vec}(\boldsymbol{Z}^\top)\right) = \mathbf{I}_{n-1} \otimes \boldsymbol{\Sigma}_0$ in (A.129) and the mixed-product property of the Kronecker product. Therefore

$$\boldsymbol{Z}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1}) \perp\!\!\!\perp \boldsymbol{Z}^\top \boldsymbol{1}_{n-1} \implies \text{III}_2 = 0 \tag{A.131}$$

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

$$\begin{aligned}
\text{III}_3 &= \mathbb{E}\left[(\overline{\boldsymbol{Z}}_{\text{-}i} - \boldsymbol{v}_0)(\boldsymbol{Z}_{\text{-}i}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z}_{\text{-}i})^{-1}(\overline{\boldsymbol{Z}}_{\text{-}i} - \boldsymbol{v}_0)^\top \mid \boldsymbol{Z}_i\right] \\
&= \mathbb{E}\left[\boldsymbol{1}_{n-1}^\top \boldsymbol{Z}(\boldsymbol{Z}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{1}_{n-1}\right]/(n-1)^2 \\
&= \mathbb{E}\left[\text{Tr}\left(\boldsymbol{1}_{n-1}^\top \boldsymbol{Z}(\boldsymbol{Z}^\top(\mathbf{I}_{n-1} - \boldsymbol{P}_{n-1})\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{1}_{n-1}\right)\right]/(n-1)^2 \\
&= \mathbb{E}\left[\text{Tr}(\boldsymbol{\Psi}_1\boldsymbol{\Psi}_2^{-1})\right]/(n-1) \\
&= \text{Tr}\,\mathbb{E}\left[\boldsymbol{\Psi}_1\boldsymbol{\Psi}_2^{-1}\right]/(n-1) \\
&= \text{Tr}(\mathbb{E}\left[\boldsymbol{\Psi}_1\right]\mathbb{E}\left[\boldsymbol{\Psi}_2^{-1}\right])/(n-1) \\
&= \text{Tr}(\boldsymbol{\Sigma}_0\frac{\boldsymbol{\Sigma}_0^{-1}}{n-p-2})/(n-1) \\
&= \frac{p}{(n-1)(n-p-2)}
\end{aligned} \tag{A.132}$$

where the first equality is from (A.127), the second equality is similarly obtained as (A.130), 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 (A.121), (A.128), (A.131) and (A.132), we finally obtain

$$
\begin{aligned}
\mathbb{E}\left[h_{ii} \mid Z_i\right] &\leq \mathbb{E}\left[u_i^\top \boldsymbol{A}^{-1} u_i \mid Z_i\right] \\
&\leq \frac{1}{n-1} + \mathbb{E}\left[\boldsymbol{\Xi} \mid Z_i\right] \\
&= \frac{1}{n-1} + \mathrm{III}_1 + \mathrm{III}_2 + \mathrm{III}_3 \\
&\leq \frac{1}{n-1} \cdot \frac{n-2}{n-p-2} + \frac{\boldsymbol{\Phi}}{n-p-2}
\end{aligned} \tag{A.133}
$$

Recall the bound for $\mathrm{II}(\mu)$ in (A.120), 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

$$
\begin{aligned}
\mathrm{II}(\mu) &\leq 8 \, \mathbb{E}_{Z_i}\left[\mathbb{E}_F\left[\mu^2(W_i)\right] \mathbb{E}\left[h_{ii} \mid Z_i\right]\right] \\
&\leq \frac{8(n-2)\mathbb{E}\left[\mu^2(W_i)\right]}{(n-1)(n-p-2)} + \frac{8\sqrt{\mathbb{E}\left[\boldsymbol{\Phi}^2\right]}}{n-p-2}\sqrt{\mathbb{E}_{\boldsymbol{Z}_i}\left[\mathbb{E}\left[\mu^4(W_i) \mid \boldsymbol{Z}_i\right]\right]} \\
&\leq \frac{8\sqrt{\mathbb{E}\left[\mu^4(X, Z)\right]}}{n-p-2}\left(1 + \sqrt{\mathbb{E}\left[\boldsymbol{\Phi}^2\right]}\right)
\end{aligned} \tag{A.134}
$$

where in the above equality, $\boldsymbol{\Phi} \sim \chi^2_{p-1}$ and is independent from $\boldsymbol{Z}_{-i}$. 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 $\mathrm{II}(\mu)$,

$$
\mathrm{II}(\mu) = O\left(\frac{p}{n-p-2}\right). \tag{A.135}
$$

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 $\mathrm{II}(\mu^\star)$. Shifting back to the n_2 notation, we finally establish (F.1), i.e.

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

$\square$

A.7.2 Proofs in Section F.3

Proof of Proposition F.2. 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 F.1. Following the derivations up to (A.115) in the proof of Proposition F.1, it suffices to deal with the following term:

$$
\mathrm{II}(\mu) := \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_F\left(\mu(W_i)\right) \mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\chi^2(F^{\boldsymbol{T}}\|F)\right]\right].
$$

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:

$$F : \mathbb{P}\left(X_i \,|\, \boldsymbol{Z}\right) = \prod_{k=1}^{K}\left(q(k, k_1, k_2)\right)^{\mathbb{1}\left\{X_i=k, W_{i,j-1}=k_1, W_{i,j+1}=k_1\right\}} \tag{A.136}$$

$$F_{\boldsymbol{T}} : \mathbb{P}\left(X_i \,|\, \boldsymbol{Z}, \boldsymbol{T}\right) = \prod_{k=1}^{K}\left(\hat{q}(k, k_1, k_2)\right)^{\mathbb{1}\left\{X_i=k, W_{i,j-1}=k_1, W_{i,j+1}=k_1\right\}} \tag{A.137}$$

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

$$\chi^2(F_{\boldsymbol{T}}||F) = \sum_{k=1}^{K} \frac{(\hat{q}(k, k_1, k_2) - q(k, k_1, k_2))^2}{q(k, k_1, k_2)}$$

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,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\chi^2(F_{\boldsymbol{T}}||F)\right] &= \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\frac{(\hat{q}(k, k_1, k_2) - q(k, k_1, k_2))^2}{q(k, k_1, k_2)}\right] \\
&= \sum_{k=1}^{K} \frac{q(k, k_1, k_2)(1 - q(k, k_1, k_2))}{N(:, k_1, k_2)q(k, k_1, k_2)} \\
&= \sum_{k=1}^{K} \frac{K - 1}{N(:, k_1, k_2)} \tag{A.138}
\end{aligned}$$

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.

$$\begin{aligned}
\Pi(\mu) &= \mathbb{E}_{\boldsymbol{Z}}\left[\mathrm{Var}_F\left(\mu(W_i)\right) \mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\chi^2(F_{\boldsymbol{T}}||F)\right]\right] \\
&= \mathbb{E}_{Z_i}\left[\mathrm{Var}_F\left(\mu(W_i)\right) \mathbb{E}\left[\mathbb{E}_{\boldsymbol{T}|\boldsymbol{Z}}\left[\chi^2(F_{\boldsymbol{T}}||F)\right]\,|\,Z_i\right]\right] \\
&= \mathbb{E}_{Z_i}\left[\mathrm{Var}_F\left(\mu(W_i)\right) \mathbb{E}\left[\frac{K - 1}{N(:, W_{i,j-1}, W_{i,j+1})}\,\Big|\,Z_i\right]\right] \\
&= \mathbb{E}_{Z_i}\left[\mathrm{Var}_F\left(\mu(W_i)\right) \mathbb{E}\left[\frac{K - 1}{1 + N_{n-1}(W_{i,j-1}, W_{i,j+1})}\,\Big|\,Z_i\right]\right] \tag{A.139}
\end{aligned}$$

where the second equality comes from the tower property of conditional expectation, the third equality holds due to (A.138) 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 $Z_i = W_{i,-j}$

$$\mathbb{1}_{\{W_{m,j-1}=W_{i,j-1},\, W_{m,j+1}=W_{i,j+1}\}} \overset{i.i.d.}{\sim} \mathrm{Bern}(q(W_{i,j-1}, W_{i,j+1})), \quad m \in [n], \; m \neq i.$$

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,

$$\mathbb{E}\left[\frac{1}{1 + B}\right] = \frac{1}{(n + 1)q} \cdot (1 - (1 - q)^{n+1}). \tag{A.140}$$

hence we can bound the term $\Pi(\mu)$ as below

$$\Pi(\mu) = \frac{K-1}{n}\mathbb{E}_{Z_i}\left[\text{Var}_F\left(\mu(W_i)\right)\frac{1-(1-q(W_{i,j-1},W_{i,j+1}))^n}{q(W_{i,j-1},W_{i,j+1})}\right] \tag{A.141}$$

$$\leq \frac{K-1}{n}\mathbb{E}_{Z_i}\left[\text{Var}_F\left(\mu(W_i)\right)\right]\frac{K^2}{K^2\min\{q(k_1,k_2)\}} \tag{A.142}$$

$$\leq \frac{K^3}{n}\frac{\mathbb{E}\left[\mu^2(X,Z)\right]}{q_0} \tag{A.143}$$

where the equality holds as a result of (A.139) and (A.140). 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 F.1, 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. $\qquad\square$

A.7.3   Ancillary Lemmas

Lemma A.7. *The χ^2-divergence between $P:\mathcal{N}\left(\boldsymbol{a}_1,\Sigma_1\right)$ and $Q:\mathcal{N}\left(\boldsymbol{a}_2,\Sigma_2\right)$ 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 A.7. According to the definition of the χ^2-divergence, we have

$$\chi^2(P\|Q) := \int \left(\frac{dP}{dQ}\right)^2 dQ - 1 = \int \frac{p^2(x)}{q(x)}dx - 1, \tag{A.144}$$

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

$$f(x) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(x-\boldsymbol{a})^\top\Sigma^{-1}(x-\boldsymbol{a})\right\}, \quad x \in \mathbb{R}^d. \tag{A.145}$$

Hence we can calculate the χ^2-divergence as below,

$$\chi^2(P\|Q) = \frac{|\Sigma_2|^{\frac{1}{2}}}{|\Sigma_1|}\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{\frac{d}{2}}}\exp\left\{-\frac{1}{2}(x-\boldsymbol{a}_1)^\top(2\Sigma_1^{-1})(x-\boldsymbol{a}_1)+\frac{1}{2}(x-\boldsymbol{a}_2)^\top\Sigma_2^{-1}(x-\boldsymbol{a}_2)\right\}dx - 1$$

$$:= \frac{|\Sigma_2|^{\frac{1}{2}}}{|\Sigma_1|}\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{\frac{d}{2}}}\exp\left\{\text{II}_1 + \text{II}_2 + \text{II}_3\right\}dx - 1, \tag{A.146}$$

where the first equality holds following the definition in (A.144) and the second equality comes from expanding the term in the exponent and combining, together with the following new notations:

$$\text{II}_1 := -\frac{1}{2}x^\top(2\Sigma_1^{-1} - \Sigma_2^{-1})x \tag{A.147}$$

$$\text{II}_2 := -\frac{1}{2}\cdot(-2x^\top)(2\Sigma_1^{-1}\boldsymbol{a}_1 - \Sigma_2^{-1}\boldsymbol{a}_2) \tag{A.148}$$

$$\text{II}_3 := -\frac{1}{2}(2\boldsymbol{a}_1^\top\Sigma_1^{-1}\boldsymbol{a}_1 - \boldsymbol{a}_2\Sigma_2^{-1}\boldsymbol{a}_2) \tag{A.149}$$

57

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

$$
\begin{aligned}
(\Sigma_1^{-1}\Sigma_\star\Sigma_2^{-1})^{-1} &= \Sigma_2\Sigma_\star^{-1}\Sigma_1 = 2\Sigma_2 - \Sigma_1 & \text{(A.150)}\\
2\Sigma_\star\Sigma_1^{-1} - \mathrm{I}_d &= \Sigma_\star(2\Sigma_1^{-1} - \Sigma_\star^{-1}) = \Sigma_\star\Sigma_2^{-1} & \text{(A.151)}\\
\frac{1}{2}\boldsymbol{a}_\star^\top\Sigma_\star^{-1}\boldsymbol{a}_\star &= \frac{1}{2}(2\Sigma_1^{-1}\boldsymbol{a}_1 - \Sigma_2^{-1}\boldsymbol{a}_2)^\top\Sigma_\star(2\Sigma_1^{-1}\boldsymbol{a}_1 - \Sigma_2^{-1}\boldsymbol{a}_2)\\
&= 2\boldsymbol{a}_1^\top\Sigma_1^{-1}\Sigma_\star\Sigma_1^{-1}\boldsymbol{a}_1 - 2\boldsymbol{a}_1^\top\Sigma_1^{-1}\Sigma_\star\Sigma_2^{-1}\boldsymbol{a}_2 + \frac{1}{2}\boldsymbol{a}_2^\top\Sigma_2^{-1}\Sigma_\star\Sigma_2^{-1}\boldsymbol{a}_2\\
&= 2\boldsymbol{a}_1^\top\Sigma_1^{-1}\Sigma_\star\Sigma_1^{-1}\boldsymbol{a}_1 - 2\boldsymbol{a}_1^\top(2\Sigma_2 - \Sigma_1)^{-1}\boldsymbol{a}_2 + \frac{1}{2}\boldsymbol{a}_2^\top\Sigma_2^{-1}\Sigma_\star\Sigma_2^{-1}\boldsymbol{a}_2 & \text{(A.152)}
\end{aligned}
$$

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 (A.150). The above equations will be used a lot for the incoming derivations. Now the term in the exponent can be written as

$$
\begin{aligned}
&\mathrm{II}_1 + \mathrm{II}_2 + \mathrm{II}_3\\
=\ & -\frac{1}{2}(x^\top\Sigma_\star^{-1}x - 2x^\top\Sigma_\star^{-1}\boldsymbol{a}_\star) + \mathrm{II}_3\\
=\ & -\frac{1}{2}(x - \boldsymbol{a}_\star)^\top\Sigma_\star^{-1}(x - \boldsymbol{a}_\star) + \frac{1}{2}\boldsymbol{a}_\star^\top\Sigma_\star^{-1}\boldsymbol{a}_\star - \frac{1}{2}(2\boldsymbol{a}_1^\top\Sigma_1^{-1}\boldsymbol{a}_1 - \boldsymbol{a}_2\Sigma_2^{-1}\boldsymbol{a}_2)\\
=\ & \lambda(x) + \boldsymbol{a}_1^\top\Sigma_1^{-1}(2\Sigma_\star\Sigma_1^{-1} - \mathrm{I}_d)\boldsymbol{a}_1 - 2\boldsymbol{a}_1^\top(2\Sigma_2 - \Sigma_1)^{-1}\boldsymbol{a}_2 + \frac{1}{2}\boldsymbol{a}_2^\top\Sigma_2^{-1}(\Sigma_\star\Sigma_2^{-1} + \mathrm{I}_d)\boldsymbol{a}_2\\
=\ & \lambda(x) + \boldsymbol{a}_1^\top\Sigma_1^{-1}\Sigma_\star\Sigma_2^{-1}\boldsymbol{a}_1 - 2\boldsymbol{a}_1^\top(2\Sigma_2 - \Sigma_1)^{-1}\boldsymbol{a}_2 + \boldsymbol{a}_2^\top\Sigma_2^{-1}\Sigma_\star\Sigma_1^{-1}\boldsymbol{a}_2\\
=\ & \lambda(x) + \boldsymbol{a}_1^\top(2\Sigma_2 - \Sigma_1)^{-1}\boldsymbol{a}_1 - 2\boldsymbol{a}_1^\top(2\Sigma_2 - \Sigma_1)^{-1}\boldsymbol{a}_2 + \boldsymbol{a}_2^\top(2\Sigma_2 - \Sigma_1)^{-1}\boldsymbol{a}_2\\
=\ & \lambda(x) + (\boldsymbol{a}_1 - \boldsymbol{a}_2)^\top(2\Sigma_2 - \Sigma_1)^{-1}(\boldsymbol{a}_1 - \boldsymbol{a}_2) := \lambda(x) + Q(\boldsymbol{a}_1, \boldsymbol{a}_2, \Sigma_1, \Sigma_2) & \text{(A.153)}
\end{aligned}
$$

where the first equality holds by the definition of $\Sigma_\star$, $\boldsymbol{a}_\star$ and (A.147), (A.148), and the second equality holds due to (A.149). 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 (A.152) 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 (A.151) 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

$$
\begin{aligned}
\chi^2(P\|Q) &= \frac{|\Sigma_2|^{\frac{1}{2}}}{|\Sigma_1|}\exp\{Q(\boldsymbol{a}_1, \boldsymbol{a}_2, \Sigma_1, \Sigma_2)\}\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{\frac{d}{2}}}\exp\{\lambda(x)\}\,dx - 1\\
&= \frac{|\Sigma_2|^{\frac{1}{2}}}{|\Sigma_1|}\exp\{Q(\boldsymbol{a}_1, \boldsymbol{a}_2, \Sigma_1, \Sigma_2)\}\int_{\mathbb{R}^d}\frac{|\Sigma_\star|^{\frac{1}{2}}}{(2\pi)^{\frac{d}{2}}|\Sigma_\star|^{\frac{1}{2}}}\exp\{\lambda(x)\}\,dx - 1\\
&= \frac{|\Sigma_2|^{\frac{1}{2}}}{|\Sigma_1|}|\Sigma_\star|^{\frac{1}{2}}\exp\{Q(\boldsymbol{a}_1, \boldsymbol{a}_2, \Sigma_1, \Sigma_2)\} - 1\\
&= \frac{|\Sigma_2|}{|\Sigma_1|^{\frac{1}{2}}}|\Sigma_1^{-1}\Sigma_\star\Sigma_2^{-1}|^{\frac{1}{2}}\exp\{Q(\boldsymbol{a}_1, \boldsymbol{a}_2, \Sigma_1, \Sigma_2)\} - 1\\
&= \frac{|\Sigma_2|}{|\Sigma_1|^{\frac{1}{2}}|2\Sigma_2 - \Sigma_1|^{\frac{1}{2}}}\exp\left\{(\boldsymbol{a}_1 - \boldsymbol{a}_2)^\top(2\Sigma_2 - \Sigma_1)^{-1}(\boldsymbol{a}_1 - \boldsymbol{a}_2)\right\} - 1
\end{aligned}
$$

where the first equality comes from (A.146) and (A.153), 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 (A.150). $\qquad\square$

B A troublesome 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 $(\mathbb{E}\left[W^\top W\right])^{-1} \mathbb{E}\left[WY\right] = (0, \frac{3}{4})^\top$, i.e., zero for the non-null covariate and non-zero for the null covariate.

C Hardness of upper confidence bounds

Let D_n be the i.i.d. samples $\{Y_i, X_i, Z_i\}_{i=1}^n$ and consider the mMSE gap $\mathcal{I}^2$, since the following theorem involves the mMSE gap under different joint laws over (Y, X, Z), we write $\mathcal{I}^2$ as a nonparametric functional explicitly, i.e.,

$$\mathcal{I}^2(F) = \mathbb{E}_F\left[\mathrm{Var}_F\left(\mathbb{E}_F\left[Y \mid X\right] \mid Z\right)\right], \tag{C.1}$$

where F denotes the joint law over (Y, X, Z). Also note the inferential target in following theorem is the squared version of the mMSE gap $\mathcal{I}^2$ instead of $\mathcal{I}$. We use $\mathcal{I}^2$ only for the convenience of calculation and this will not change the message.

Theorem C.1. *Consider any joint law over (Y, X, Z) such that $\mathrm{Var}(Y) < \infty$ and denote the class of these distributions by $\mathcal{F}$, given any confidence level $1 - \alpha$, there does not exist a non-trivial upper confidence bound for $\mathcal{I}^2(F)$ with asymptotic coverage, i.e., for any upper confidence bound procedure U that is pointwise asymptotically valid:*

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

we must have

$$\sup_{F \in \mathcal{F}} \limsup_{n \to \infty} \mathbb{P}_F\left(U(D_n) \leq \mathbb{E}_F\left[\mathrm{Var}_F\left(Y \mid Z\right)\right]\right) \leq \alpha \tag{C.2}$$

Proof. We prove by contradiction. Suppose there exists an upper confidence bound procedure ensuring asymptotic coverage such that (C.2) holds, that is, there exists a joint law over (Y, X, Z), denoted by $F_\infty \in \mathcal{F}$ such that

$$\limsup_{n \to \infty} \mathbb{P}_\infty\left(U(D_n) \leq \mathbb{E}_\infty\left[\mathrm{Var}_\infty\left(Y \mid Z\right)\right]\right) > \alpha. \tag{C.3}$$

where $\mathbb{P}_\infty$, $\mathbb{E}_\infty$, Var_∞ denote that the data generating distribution for i.i.d. sample D_n is F_∞. Further let $\lambda_1 = \mathbb{E}_\infty\left[\mathrm{Var}_\infty\left(Y \mid Z\right)\right]$, then we have $\lambda_1 > 0$, since otherwise $\mathbb{E}_\infty\left[\mathrm{Var}_\infty\left(Y \mid Z\right)\right] = \mathcal{I}^2(F_\infty) = 0$ and (C.3) does not hold. 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 (X, Z) follows the same distribution as that under F_∞ and so does the conditional distribution of $\epsilon \mid X, Z$, where $\epsilon = Y - \mathbb{E}\left[Y \mid X, Z\right]$, that is,

$$\mathbb{P}_k\left(X, Z\right) = \mathbb{P}_\infty\left(X, Z\right), \qquad \forall\, k \geq 1 \tag{C.4}$$
$$\mathbb{P}_k\left(\epsilon \mid X, Z\right) = \mathbb{P}_\infty\left(\epsilon \mid X, Z\right), \qquad \forall\, k \geq 1 \tag{C.5}$$

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) $\mathbb{E}_k\left[\mu_k^\star(X, Z) \mid Z\right] = \mathbb{E}_\infty\left[\mu_\infty^\star(X, Z) \mid Z\right]$ when $Z \in A_k$;

(d) $\mathrm{Var}_k\left(\mu_k^\star(X, Z) \mid Z\right) = \mathrm{Var}_\infty\left(\mu_\infty^\star(X, Z) \mid Z\right) + k\left(2\lambda_1 - \mathcal{I}^2(F_\infty)\right)$ 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 $\mu_k^\star(X, Z) :=$ $\mathbb{E}_k[Y \mid X, Z]$, $\mu_\infty^\star(X, Z) := \mathbb{E}_\infty[Y \mid X, Z]$. Note here $\mathbb{E}_k[\cdot \mid Z]$, $\text{Var}_k(\cdot \mid Z)$ are the same as $\mathbb{E}_\infty[\cdot \mid Z]$, $\text{Var}_\infty(\cdot \mid Z)$ due to (C.4). Hence we can calculate $\mathcal{I}(F_k)$ through the following

$$
\begin{aligned}
\mathcal{I}^2(F_k) - \mathcal{I}^2(F_\infty) &= \mathbb{E}_\infty\left[\mathbb{1}_{\{A_k\}}\left(\text{Var}_\infty\left(\mu_k^\star(X, Z) \mid Z\right) - \text{Var}_\infty\left(\mu_\infty^\star(X, Z) \mid Z\right)\right)\right] \\
&= \mathbb{E}_\infty\left[\mathbb{1}_{\{A_k\}} k\left(2\lambda_1 - \mathcal{I}^2(F_\infty)\right)\right] \\
&= 2\lambda_1 - \mathcal{I}^2(F_\infty) := \lambda_2
\end{aligned} \tag{C.6}
$$

where the first equality comes from (C.1), (C.4) 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

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

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 (C.4), the fourth equality holds due to (C.5), (b) and (c), the fifth equality comes from (C.6). 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

$$
\mathbb{P}_k\left(U(D_n) \geq 2\lambda_1\right) \geq 1 - \alpha + o_k(1) \tag{C.7}
$$

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

$$
\left|\mathbb{P}_\infty\left(U(D_n) \geq 2\lambda_1\right) - \mathbb{P}_k\left(U(D_n) \geq 2\lambda_1\right)\right| \leq d_{TV}(F_k, F_\infty) \leq \frac{1}{k}, \quad \forall\, k \geq 1 \tag{C.8}
$$

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 (C.7) and (C.8) yields the following

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

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

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

which contradicts

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

$\square$

D 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 (2.5):

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

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 D.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

$$\frac{Q(Y, X, Z) Q(\tilde{X} \mid Z)}{P(Y, X, Z) P(\tilde{X} \mid Z)} = \frac{Q(Y \mid X, Z)}{P(Y \mid X, Z)} \frac{Q(X, Z)}{P(X, Z)} \frac{Q(\tilde{X} \mid Z)}{P(\tilde{X} \mid Z)} = w_1(\tilde{X}, Z) w(X, Z), \tag{D.2}$$

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 (D.1), applying the central limit theorem, and combining them with the delta method.

E A general algorithm for inference on the MACM gap

Algorithm 2 involves computing the terms $\mathbb{E}[\mu(X_i, Z_i) \mid Z_i]$ 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 2.3, 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 regression function μ.

$$U := \mu(X, Z), \quad g(z) := \mathbb{E}[\mu(X, Z) \mid Z = z],$$
$$G_z(u) := \mathbb{P}\left(U < u \mid Z = z\right), \quad F_z(u) := \mathbb{P}\left(U \le u \mid Z = z\right).$$
$$\varsigma(z) := \sqrt{\mathrm{Var}\left(\mu(X, Z) \mid Z = z\right)},$$
$$C_{u,z,y} := \frac{\max\{|G_{z,y}(u) - G_{z,y}(g(z))|, |F_{z,y}(u) - F_{z,y}(g(z))|\}}{|u - g(z)|} \tag{E.1}$$

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 E.1.

Assumption E.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

$$C_{u,z,y} \le C_{z,y} \quad when \quad |u - g(z)| \le \varsigma(z) \delta_{z,y}.$$

61

(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) $\mathbb{E}\left[\varsigma^2(Z)\right] < \infty, \ \mathbb{E}\left[\frac{\mathbb{E}\left[|\mu(X,Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]|^3\,|\,Z\right]}{\varsigma^3(Z)}\right] < \infty.$

These assumptions are placed because we have to construct the Monte Carlo estimator of $\mathbb{E}\left[\mu(X,Z)\,|\,Z\right]$ then plug it into the discontinuous indicator functions in $f_{\ell_1}(\mu)$. Assumptions E.1(a) and E.1(b) are smoothness requirements on the the CDF of $\mu(X,Z)\,|\,Z,Y$ around $\mathbb{E}\left[\mu(X,Z)\,|\,Z\right]$. Assumption E.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 E.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 $g(Z_i)$ 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 E.2. *Under the same setting as in Theorem 3.3, if either (i) $\mathbb{E}\left[\mathrm{Var}(\mu(X,Z)\,|\,Z)\right] = 0$ or (ii) $\mathbb{E}\left[\mathrm{Var}\left(\mathbb{1}_{\{Y\cdot[\mu(X,Z)-\mathbb{E}[\mu(X,Z)\,|\,Z]]<0\}}\,|\,Z,Y\right)\right] > 0$ holds together with Assumption E.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 2 satisfies*

$$\mathbb{P}\left(L_{n,M,K}^\alpha(\mu) \le \mathcal{I}_{\ell_1}\right) \ge 1 - \alpha + o(1).$$

The proof can be found in Appendix A.5. 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).

E.1 Illustration of assumption E.1

We consider the joint distribution over W to be p-dimensional multivariate Gaussian with $X = W_j, Z = W_{-j}$ for some $1 \le j \le p$, and Y follows a generalized linear model with logistic link. That is,

$$W \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}\right), \ \ \mu^\star(W) = 2\mathbb{P}\left(Y = 1\,|\,W\right) - 1, \ \ \text{where } \mathbb{P}\left(Y = 1\,|\,W\right) = \frac{\exp\left(W\beta^\star\right)}{1 + \exp\left(W\beta^\star\right)}, \ \beta^\star \in \mathbb{R}^p.$$

Choosing logistic regression as the fitting algorithm, we have $U := \mu(X,Z)$ takes the following form

$$U := \mu(W) = \frac{2\exp\left(W\beta\right)}{1 + \exp\left(W\beta\right)} - 1$$

where $\beta \in \mathbb{R}^p$ is the fitted regression coefficient vector and $\beta_j \ne 0$ whenever $\mathbb{E}\left[\mathrm{Var}(\mu(X,Z)\,|\,Z)\right] > 0$. 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

$$h_{\text{logit}}(u) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(\text{logit}(u)-a)^2}{2\sigma^2}\right)\frac{1}{u(1-u)}, \ \ u \in (0,1) \tag{E.2}$$

62

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

$$h(x \mid Z = z, Y = 1) = \frac{h(x \mid Z = z)\mathbb{P}\left(Y = 1 \mid W = w\right)}{\int h(x \mid Z = z)\mathbb{P}\left(Y = 1 \mid W = w\right) dx} \tag{E.3}$$

where $w_j = x, w_{-j} = z$, $h(x \mid Z = z, Y = 1)$ 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 (E.3)

$$h_{z,1}(u) = \frac{h_{\text{logit}}(u)\mathbb{P}\left(Y = 1 \mid W = w\right)}{\int h_{\text{logit}}(u)\mathbb{P}\left(Y = 1 \mid W = w\right) du} \tag{E.4}$$

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.

F   Co-sufficient floodgate details

F.1   Monte Carlo analogue of co-sufficient floodgate

Similarly as in Section 2, when the conditional expectations in Algorithm 3 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 $\boldsymbol{X}_m \mid \boldsymbol{Z}_m, \boldsymbol{T}_m$, conditionally independently of $\boldsymbol{X}_m$ and $\boldsymbol{y}$ and thus replace R_m and V_m, respectively, by the sample estimators

$$(R_m^K, V_m^K) = \frac{1}{n_2}\left(\sum_{i=(m-1)n_2+1}^{mn_2} Y_i\left(\mu(X_i, Z_i) - \frac{1}{K}\sum_{k=1}^K \mu(\tilde{X}_i^{(k)}, Z_i)\right),\right.$$

$$\left.\sum_{i=(m-1)n_2+1}^{mn_2} \frac{1}{K-1}\sum_{k=1}^K\left(\mu(X_i^{(k)}, Z_i) - \frac{1}{K}\sum_{k=1}^K \mu(\tilde{X}_i^{(k)}, Z_i)\right)^2\right)$$

We defer to future work a proof of validity of the Monte Carlo analogue of co-sufficient floodgate following similar techniques as Theorem 2.4.

F.2   Low-dimensional multivariate Gaussian model

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

Proposition F.1. *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 $Z_i \sim \mathcal{N}\left(\boldsymbol{v}_0, \boldsymbol{\Sigma}_0\right)$. Assume σ^2 is known, the batch size n_2 satisfies $n_2 > p + 2$ and choose $\mathcal{T}$ to 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

$$f(\mu) - f_n^{\mathcal{T}}(\mu) = O\left(\frac{p}{n_2 - p - 2}\right) \tag{F.1}$$

The proof can be found in Appendix A.7.1. Note the condition $n_2 > p + 2$ is not surprising as we are aware that 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 (F.1) allows p to grow with n in general, but when p is fixed, it gives the rate of $O(n_2^{-1})$, as mentioned in Section 3.2.

F.3   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 $X = W_j$, $Z = W_{-j}$, then the model parameters include the initial probability vector $\pi^{(1)} \in \mathbb{R}^K$ with $\pi_k^{(1)} = \mathbb{P}(W_1 = k)$ and the transition probability matrix $\Pi^{(j)} \in \mathbb{R}^{K \times K}$ (between W_{j-1} and $X = W_j$) with $\Pi_{k,k'}^{(j)} = \mathbb{P}(W_j = k' \mid W_{j-1} = k)$. Further denoting $q(k, k_1, k_2) = \mathbb{P}(W_j = k | W_{j-1} = k_1, W_{j+1} = k_2)$, we have

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

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

$$\mathbb{P}(\boldsymbol{X}_m \mid \boldsymbol{Z}_m) = \prod_{k,k_1,k_2 \in [K]} (q(k, k_1, k_2))^{N(k,k_1,k_2)}, \tag{F.2}$$

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 conclude that $\{N(k, k_1, k_2)\}_{(k,k_1,k_2 \in [K])}$ is sufficient immediately, and we proceed with this sufficient statistic.

Proposition F.2. *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 assume $\mathbb{E}\left[(\mu^\star)^2(X, Z)\right], \mathbb{E}\left[\mu^2(X, Z)\right] < \infty$, we have

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

The proof can be found in Appendix A.7.2. Note that $\mathcal{T}$ here is not minimal sufficient and the above rate is cubic in K. The non-minimal sufficient statistics is adopted for the discrete Markov chain models 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 statistics. Again, K is allowed to grow with n at certain rate in general, but when it is fixed we get a rate of $O(n_2^{-1})$, as mentioned in Section 3.2.

G   Further simulation details

G.1   Nonlinear model setup

Consider W which follows a Gaussian copula distribution with $X = W_{j_0}, Z = W_{-j_0}$ for some j_0 ($1 \leq j_0 \leq p$), i.e.,

$$W^{\text{latent}} \sim AR(1), \quad W_j = 2\varphi(X_j^{\text{latent}}) - 1, \quad \forall\ 1 \leq j \leq p. \tag{G.1}$$

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,

$$\mu^\star(x, z) = \mu^\star(w) := \sum_{j \in S^1} g_j(w_j) + \sum_{(j,l) \in S^2} g_j(w_j)g_l(w_l) + \sum_{(j,l,m) \in S^3} g_j(w_j)g_l(w_l)g_m(w_m) \tag{G.2}$$

64

where each function $g_j(x)$ is randomly chosen from the following:

$$\sin(\pi x), \ \cos(\pi x), \ \sin(\pi x/2), \ \cos(\pi x)I(x > 0), \ x\sin(\pi x), \ x, \ |x|, \ x^2, \ x^3, \ e^x - 1. \tag{G.3}$$

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, (G.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 $S_{\mathrm{wl}} = S_\star$. 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 $2|S_{\mathrm{wl}}|/|S_\star|$ for variables in S_{wl}, $|S_\star \setminus S_{\mathrm{wl}}|/|S_\star|$ 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 $|S_{\mathrm{wl}}| \leq 5$. Regarding the construction of S^3, each time we randomly pick 3 variables from $S_\star$ with the unscaled weight being $1.5|S_{\mathrm{wl}}|/|S_\star|$ for variables in S_{wl}, $|S_\star \setminus S_{\mathrm{wl}}|/|S_\star|$ 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 $|S_{\mathrm{wl}}| = 0$.

G.2 Implementation details of fitting algorithms

Regarding how to obtain the regression function estimator, there will be four different fitting algorithms.

- *LASSO*: We fit a linear model by 10-fold cross-validated LASSO and output a regression function estimator. 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 2.3, 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.

G.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

$$\mathrm{LCB}_{\mathrm{OLS}} = \left\{ \begin{array}{ll} \mathrm{LCI}\sqrt{\mathbb{E}[\mathrm{Var}(X \mid Z)]} & \text{if } \beta > 0 \\ -\mathrm{UCI}\sqrt{\mathbb{E}[\mathrm{Var}(X \mid Z)]} & \text{if } \beta < 0 \end{array} \right. \tag{G.4}$$

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).

H   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. (2020).

Regarding the genotype modelling, we consider the hidden Markov models (HMM) (Scheet and Stephens, 2006), as used in Sesia et al. (2020, 2017), which provides a good description of the linkage disequilibrium (LD) structure. We obtain the fitted HMM parameters from Sesia et al. (2020) 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:

(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

$$A_{g_j} \mid A_{\text{-}g_j}, W_{\text{-}g_j} \stackrel{d}{=} A_{g_j} \mid A_{\text{-}g_j}, \tag{H.1}$$

$$W_{g_j} \mid A_{g_j}, A_{\text{-}g_j}, W_{\text{-}g_j} \stackrel{d}{=} W_{g_j} \mid A_{g_j}. \tag{H.2}$$

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. (2020) 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 (H.1) and (H.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 KnockoffZoom analysis (Sesia et al., 2020) 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., 2020) and included both genotypes (SNPs from chromosomes 1–22) and the 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. (2020). Although their selection used all of the data (including the data 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.