Variance Minimization in the Wasserstein Space for Invariant Causal Prediction

Guillaume Martinet* Princeton University

Alexander Strzalkowski* Princeton University

Barbara E. Engelhardt Princeton University, Gladstone Institutes

Abstract

Selecting powerful predictors for an outcome is a cornerstone task for machine learning. However, some types of questions can only be answered by identifying the predictors that causally affect the outcome. A recent approach to this causal inference problem leverages the invariance property of a causal mechanism across differing experimental environments (Peters et al., 2016; Heinze-Deml et al., 2018). This method, *invariant causal prediction* (ICP), has a substantial computational defect – the runtime scales exponentially with the number of possible causal variables. In this work, we show that the approach taken in ICP may be reformulated as a series of nonparametric tests that scales linearly in the number of predictors. Each of these tests relies on the minimization of a novel loss function – the Wasserstein variance – that is derived from tools in optimal transport theory and is used to quantify distributional variability across environments. We prove under mild assumptions that our method is able to recover the set of identifiable direct causes, and we demonstrate in our experiments that it is competitive with other benchmark causal discovery algorithms.

1 INTRODUCTION

Distinguishing between correlation and causation is a fundamental challenge that has been studied extensively over the years (Pearl, 2009). This distinction is necessary, for instance, to understand the behavior of regression under interventions. Although regression is well understood in statistics and machine learning, when the same regression model is applied in different experimental conditions, the results may differ dramatically. Identifying which predictors are causal for an outcome is central to solving this limitation, since causal mechanisms by definition remain invariant across different experimental settings (Peters et al., 2017).

Causal relationships are often represented by a directed *causal* graph, where each arrow signifies a direct cause-effect relationship between two variables. Usually, the approach to causal discovery has been to learn from observational or interventional data the entire causal graph of the variables, sometimes only up to Markov equivalence. Many methods have been developed that use a variety of assumptions. For example, methods such as Inductive Causation (IC, Pearl (2009)), Fast Causal Inference, and Peter and Clark's algorithm (FCI and PC, Spirtes et al. (2000)) identify the Markov equivalence class of the causal graph using conditional independence tests under the so-called *faithfulness* assumption, that all observable conditional independences stem only from the graph. Score-based methods such as Greedy Equivalence Search (GES, Chickering (2002)) and Greedy Interventional Equivalence Search (GIES, Hauser and Bühlmann (2012, 2015)) try to find the graph that maximizes some score function. On the other hand, methods like Linear Non-Gaussian Additive Models (LiNGAM, Shimizu et al. (2006, 2011)), Regression with Subsequent Independence Test (RESIT, Peters et al. (2014)), or Causal Additive Models (CAM, Bühlmann et al. (2014)) rely on model restrictions, such as additive nonlinear structural equations or non-Gaussian noises. Another example is the Greedy Sparsest Permutation (GSP) family of methods (Solus et al., 2017; Wang et al., 2017; Squires et al., 2020), which combine conditional independence tests with score-based ideas.

In practice, however, learning the whole causal graph is excessive. Often we are only interested in determining which variables are a *direct cause* of a specific target

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. * Equal contribution.

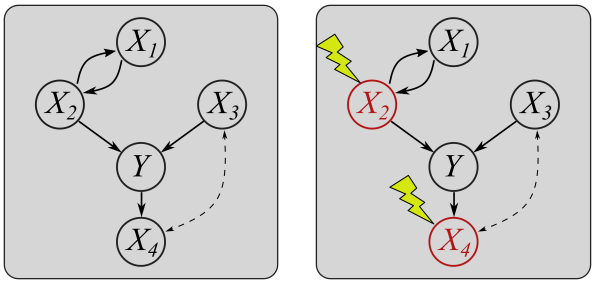
variable. Here, we define the *direct causes* as the parents of the target in the causal graph, which means that their causal effect on the target is not fully mediated by other observed variables.

A useful framework has been developed for inferring the direct causes of a target variable that – instead of using conditional independence tests, score maximization, or model assumptions – uses the stability of causal relationships across environments (Peters et al., 2016). This approach, known as invariant causal prediction (ICP), leverages a key property of causal mechanisms: the conditional distribution of the target given its direct causes will not change when we intervene on any of the observed variables excluding the target. This method has desirable advantages over previous approaches (e.g., in general, conditional independence is not a testable hypothesis (Shah et al., 2020)) and has been the source of inspiration for many recent algorithms (Rothenhäusler et al., 2015; Ghassami et al., 2017; Rothenhäusler et al., 2019; Arjovsky et al., 2019).

Unfortunately, the number of tests that ICP needs to perform scales exponentially in the number of predictors. Thus, ICP often cannot be used even when the number of predictors is moderate. In general, ICP is applied to only a *small* subset of the predictors, preselected by a sparse regression technique such as Lasso (Tibshirani, 1996) or boosting (Friedman, 2001; Hastie et al., 2009). This preselection step may severely reduce the power of ICP by rejecting variables that are direct causes, while including others that are not.

In this work, we show that the approach taken in ICP may be reformulated as a multiple-testing problem, where the number of tests scales linearly in the number of predictors. Given data from different experimental environments, we propose, for each predictor, to test for the existence of an invariant causal mechanism that does not involve the predictor in question. Each test involves a statistic based on a new loss function – the Wasserstein variance – that is used to quantify distributional variability across environments. More precisely, each of these statistics is obtained by solving a Wasserstein variance minimization (WVM) program over a restricted class of functions; when the resulting value surpasses some threshold, we declare the corresponding predictor as causal.

This paper is organized as follows: Section 2 introduces the setting, ICP, and our reformulation; Section 3 defines useful concepts from optimal transport and introduces the WVM algorithm; Section 4 derives theoretical guarantees about WVM; Section 5 describes implementation details of WVM; Section 6 compares WVM against other standard methods on experiments; Section 7 concludes.



(a) Environment $e = 1$. (b) Environment $e = 2$.

Figure 1: An SCM with $p = 4$, $S^* = \{2, 3\}$, a feedback loop between X_1 and X_2, and a hidden confounder between X_3 and X_4: (a) in an observational setting $e = 1$; (b) in an interventional setting $e = 2$ with interventions on X_2, X_4.

2 BACKGROUND

Suppose we are given data from E distinct experimental environments $e \in \mathcal{E} \doteq \{1, \dots, E\}$. Let $X^e \doteq (X_k^e)_{k=1,\dots,p} \in \mathbb{R}^p$ denote the p predictors and $Y^e \in \mathbb{R}$ denote the target variable in environment e. For each environment, we observe n_e i.i.d. samples. The main assumption of our paper is that the causal mechanism that relates the target variable to its direct causes is invariant across all environments. This is the invariance property that ICP exploits. Like ICP, we model the causal mechanism as a structural equation (SE) with additive noise (Peters et al., 2017).

Assumption 1 (Invariant SE)**.** *Denote $S^* \subseteq \{1, \dots, p\}$ as the set of direct causes. Let $\mathcal{F}$ represent a class of functions of the predictors, and $\mathcal{F}_{S^*} \subseteq \mathcal{F}$ a subclass of functions that depend only on the direct causes. For some fixed and unknown distribution D and function $f^* \in \mathcal{F}_{S^*}$, $\forall e \in \mathcal{E}$,*

$$Y^e = f^*(X^e) + \varepsilon^e, \quad \varepsilon^e \sim D, \quad \varepsilon^e \perp\!\!\!\perp X^e_{S^*} \doteq (X^e_k)_{k \in S^*}. \tag{1}$$

Typically, this invariance property arises in situations where the data are generated by interventions on variables other than the target. Suppose that in an observational setting $e = 1$ the variables X^1 and Y^1 are generated by a structural causal model (SCM; Figure 1a). We allow the SCM to admit feedback loops and hidden confounders as long as they do not affect the causal mechanism between the target and its direct causes. In another setting $e = 2$, if some potentially unknown variables other than the target are intervened on (Figure 1b), then the SE between Y^2 and $X^2_{S^*}$ remains unchanged, and Assumption 1 is satisfied between the two environments. Note that the interventions can either remove causal relationships or modify SEs by

changing their functions or the distribution of their noise. This type of scenario has been studied for instance by Meinshausen et al. (2016) in the context of a gene deletion experiment in yeast, where different environments are generated by knocking out one or more of the genes, and ICP is used to predict the causal effect of future interventions.

In addition, when only observational data are available, it is also possible to generate different environments satisfying Assumption 1 by splitting up the data according to the values of variables that are nondescendant of the target (e.g., an instrumental variable) in the original SCM (see Peters et al. (2016) for more details).

Invariant Causal Prediction (ICP). The goal of ICP is to recover the set S^* for the target variable Y from the data. We will denote by $\mathcal{F}_S$ the set of functions from $\mathcal{F}$ that depend only on predictors from S. The invariance property in Equation (1) offers a way to infer which predictors are direct causes by looking at subsets of predictors $S \subseteq \{1, \dots, p\}$ that satisfy the following null hypothesis:

$$H_{0,S}(\mathcal{E}) : \begin{cases} \text{for fixed } f \in \mathcal{F}_S \text{ and distr. } D, \forall e \in \mathcal{E}, \\ Y^e = f(X^e) + \varepsilon^e, \quad \varepsilon^e \sim D,\ \varepsilon^e \perp\!\!\!\perp X_S^e. \end{cases}$$

Assumption 1 implies that $H_{0,S^*}(\mathcal{E})$ is true. However, this is not sufficient to guarantee the full identifiability of S^*, since other subsets S of predictors may also satisfy the hypothesis. Instead, ICP seeks to recover the set of *identifiable* causal predictors that are defined to be predictors common to all S for which $H_{0,S}(\mathcal{E})$ is true (Peters et al., 2016).

Definition 1 (Identifiable causal predictors)**.** *Under Assumption 1, the set of identifiable causal predictors is:*

$$S(\mathcal{E}) \doteq \bigcap_{S:\ H_{0,S}(\mathcal{E})\ \textit{is true}} S \subseteq S^*. \tag{2}$$

For linear Gaussian SCMs, Peters et al. (2016) provides several sufficient conditions on the types of interventions applied in experimental environments that imply the identifiability of the direct causes, that is $S(\mathcal{E}) = S^*$. Regardless of identifiability, both the linear (Peters et al., 2016) and nonlinear (Heinze-Deml et al., 2018) versions of ICP derive an estimator of $S(\mathcal{E})$ from the data by testing $H_{0,S}(\mathcal{E})$ for all subsets S, and taking the intersection as in Equation (2). One way to test the hypothesis $H_{0,S}(\mathcal{E})$ is to regress the target variable on the set S of predictors and to test whether the resulting noise has an invariant distribution across environments, e.g., by using a Kolmogorov-Smirnov test. However, this formulation requires a combinatorial search over all subsets S. ICP's runtime hence scales exponentially with p, as it needs to perform 2^p tests in total.

Another Formulation of ICP. The exponential scaling of ICP prohibits its application to settings with even a moderate number of variables. As we show, it is in fact possible to estimate the set of identifiable causal predictors with potentially many fewer tests. For $k \in \{1, \dots, p\}$, consider the following null hypothesis:

$$H'_{0,k}(\mathcal{E}) : \begin{cases} \exists S \not\ni k, \exists f \in \mathcal{F}_S \text{ and distr. } D \text{ s.t. } \forall e \in \mathcal{E}, \\ \varepsilon^e \doteq Y^e - f(X^e) \sim D \text{ and } \varepsilon^e \perp\!\!\!\perp X_S^e. \end{cases}$$

In other words, the hypothesis $H'_{0,k}(\mathcal{E})$ means that it is possible to find a function in $\mathcal{F}$ that does not depend on the predictor k and yet satisfies Equation (1). It is easy to prove that $S(\mathcal{E})$ can in fact be expressed as the set of predictors k such that $H'_{0,k}(\mathcal{E})$ is false (proof in Appendix A):

Lemma 1 (Reformulation of $S(\mathcal{E})$)**.** *Under Assumption 1, the set of identifiable causal predictors can be expressed as:*

$$S(\mathcal{E}) = \left\{ k : H'_{0,k}(\mathcal{E})\ \textit{is false} \right\}. \tag{3}$$

Lemma 1 suggests that the approach taken in ICP may be treated as a multiple testing problem. We build an estimator of $S(\mathcal{E})$ by collecting all predictors k such that the null hypothesis $H'_{0,k}(\mathcal{E})$ may be rejected with high enough confidence. This formulation of the problem requires p tests, instead of 2^p tests as in the original ICP, scaling linearly with the number of predictors.

To test each hypothesis $H'_{0,k}(\mathcal{E})$, we rely on the minimization of a new loss function, the Wasserstein variance, over $\mathcal{F}_{-k}$ defined as the set of functions in $\mathcal{F}$ that do not depend on the predictor k. When the resulting minimum is above some threshold, we reject the hypothesis $H'_{0,k}(\mathcal{E})$. More precisely, the Wasserstein variance is used to quantify the distributional variability of the residuals $Y^e - f(X^e)$ across environments, in the sense that a high Wasserstein variance means that the residuals' distributions differ substantially across environments; conversely, a Wasserstein variance equal to zero means that these distributions are identical. Hence, a high value of the minimal Wasserstein variance over $\mathcal{F}_{-k}$ provides evidence of the nonexistence of a function f in this class such that the residuals have the same distribution across environments. Thus, a high value of the minimal Wasserstein variance means that $H'_{0,k}(\mathcal{E})$ should be rejected. We provide more details on our algorithm in the following sections.

Additional Notation. For the remainder, we introduce the following additional notation. We define $[p] \doteq \{1, \dots, p\}$ for any $p \in \mathbb{N}$, the total number of observations as $n \doteq \sum_{e=1}^{E} n_e$, and the minimum over the n_es as $n_0 \doteq \min_{e \in [E]} n_e$. Also, we call P_2 the set of probability measures on $\mathbb{R}$ with finite second moment,

and δ_x is the Dirac measure at x. Λ will refer to the set $\Lambda \doteq \{\boldsymbol{w} = (w_e)_{e=1}^E : \sum_{e=1}^E w_e = 1 \text{ and } w_e > 0,\ \forall e \in [E]\}$. We call $\mathbf{x}^e \doteq (x_i^e)_{e=1}^{n_e}$ the n_e observations of X^e and $\mathbf{x} \doteq (\mathbf{x}^e)_{e=1}^E$, and define $y_i^e, \mathbf{y}^e, \mathbf{y}$ similarly for the target variable Y^e. We also write $s \wedge t \doteq \min(s,t)$ for $s, t \in \mathbb{R}$.

3 WASSERSTEIN VARIANCE MINIMIZATION (WVM)

In order to test whether $H'_{0,k}(\mathcal{E})$ can be rejected, we need to measure the difference in distribution between the residuals $Y^e - f(X^e)$ across environments; we rely on the Wasserstein distance to quantify that difference. Compared to other metrics used in machine learning, such as the Kullback-Leibler (KL) divergence or maximum mean discrepancy (MMD, Gretton et al. (2012)), with Wasserstein distances it is possible to quantify the variability of multiple distributions in both an efficient and nonparametric way. For instance, the KL divergence requires parametric models of the distributions, and while MMD is nonparametric its complexity is $\mathcal{O}(n^2)$ compared to $\mathcal{O}(n \log n)$ for the Wasserstein distance in our setting. We rely on the Wasserstein variance, a quantity we derive from the notion of Wasserstein barycenter first introduced by Agueh and Carlier (2011). We introduce these concepts below before presenting our method, the Wasserstein variance minimization (WVM) algorithm.

Wasserstein Distance. The 2-Wasserstein distance (squared) W_2^2 is the optimal transportation cost between two probability distributions $\nu_1, \nu_2 \in P_2$ with a squared Euclidean cost function:

$$W_2^2(\nu_1, \nu_2) \doteq \inf_{\pi \in \Pi(\nu_1, \nu_2)} \int |x - y|^2 d\pi(x, y).$$

Here, $\Pi(\nu_1, \nu_2)$ is the set of all joint distributions, also called couplings, with marginals equal to ν_1 and ν_2. The Wasserstein distance defines a metric on P_2 (Theorem 7.3, Villani (2003)). Thus, $W_2(\nu_1, \nu_2) = 0$ iff $\nu_1 = \nu_2$. The resulting metric space (P_2, W_2) is also called the *Wasserstein space*.

Wasserstein Barycenter and Variance. In a Euclidean space, the barycenter x of E points $(x_e)_{e=1}^E$ with respective weights $(w_e)_{e=1}^E \in \Lambda$ minimizes $x \mapsto \sum_e w_e |x_e - x|^2$, and the resulting minimal value is their variance. By analogy, Agueh and Carlier (2011) defines the Wasserstein barycenter of E probability distributions $(\nu_e)_{e=1}^E$ as above by simply replacing the Euclidean distance by W_2. Similarly, we define the Wasserstein variance as the resulting minimal value:

Definition 2 (Wasserstein variance)**.** *Let $\boldsymbol{\nu} \doteq (\nu_e)_{e=1}^E$ be probability distributions from P_2. Their Wasserstein variance w.r.t. the weights $\boldsymbol{w} \doteq (w_e)_{e=1}^E \in \Lambda$ is defined as follows:*

$$WV_{\boldsymbol{w}}(\boldsymbol{\nu}) \doteq \inf_{\nu \in P_2} \sum_{e=1}^E w_e \cdot W_2^2(\nu_e, \nu).$$

A minimizer ν^* of the above infimum is called a Wasserstein barycenter*, and there always exists at least one Wasserstein barycenter (see Proposition 2.3, Agueh and Carlier (2011)).*

The Wasserstein variance is a practical tool to quantify the variability of different probability distributions; a low Wasserstein variance means that the distributions are more similar. In particular, the next result follows directly from the fact that W_2 is a metric:

Lemma 2 (A zero Wasserstein variance means no variability)**.** *Let $\boldsymbol{\nu}$, $\boldsymbol{w}$ be as in Definition 2. Then,*

$$WV_{\boldsymbol{w}}(\boldsymbol{\nu}) = 0 \iff \nu_1 = \nu_2 = \ldots = \nu_E.$$

WVM Algorithm. Call $\boldsymbol{\nu}(f) \doteq (\nu_e(f))_{e=1}^E$ the distribution of the residuals $Y^e - f(X^e)$ for $e \in [E]$. Fix weights $\boldsymbol{w} \doteq (w_e)_{e=1}^E \in \Lambda$. We propose to test $H'_{0,k}(\mathcal{E})$ for each $k \in [p]$ by checking whether a zero optimal value is obtained for the following population-wise Wasserstein variance minimization (WVM):

$$\Gamma_{\boldsymbol{w}}(\mathcal{F}_{-k}) \doteq \inf_{f \in \mathcal{F}_{-k}} \mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(f)). \tag{4}$$

As a consequence of Lemma 2 and the definition of $H'_{0,k}(\mathcal{E})$, we have that $H'_{0,k}(\mathcal{E})$ is false whenever $\Gamma_{\boldsymbol{w}}(\mathcal{F}_{-k}) > 0$. The WVM algorithm thus aims at testing whether the following null hypotheses may be rejected: for $k \in [p]$,

$$\tilde{H}_{0,k}(\mathcal{E}) :\ \Gamma_{\boldsymbol{w}}(\mathcal{F}_{-k}) = 0, \ \text{ vs } \ \tilde{H}_{1,k}(\mathcal{E}) :\ \Gamma_{\boldsymbol{w}}(\mathcal{F}_{-k}) > 0.$$

We form an estimator of $S(\mathcal{E})$ by collecting every predictor k such that $\tilde{H}_{0,k}(\mathcal{E})$ may be rejected with high enough confidence. Note that $\tilde{H}_{0,k}(\mathcal{E})$ is a weaker null hypothesis than $H'_{0,k}(\mathcal{E})$, as the latter implies the former but not the converse, since $H'_{0,k}(\mathcal{E})$ also implies that the residuals are independent of the causal predictors. Thus, in some situations our approach may yield a conservative estimate of $S(\mathcal{E})$, even in the limit of infinite data.

Definition 3 (WVM's identifiable causal predictors)**.** *Under Assumption 1, we define the set of identifiable causal predictors for the WVM algorithm to be:*

$$\tilde{S}(\mathcal{E}) \doteq \left\{ k : \tilde{H}_{0,k}(\mathcal{E}) \text{ is false} \right\} \subseteq S(\mathcal{E}) \subseteq S^*. \tag{5}$$

Several remarks are in order. First, for practical reasons, ICP also tests hypotheses that are effectively

Algorithm 1: WVM Algorithm

Input: $\mathbf{x}, \mathbf{y}, \alpha,$ get_threshold()
Output: $\hat{S}(\mathcal{E})$
Initialize $\hat{S}(\mathcal{E}) \leftarrow \emptyset$;
for $k \in [p]$ **do**
  Obtain $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_{-k})$ and $\hat{f}_k$;
  Set $t \leftarrow$ get_threshold$(\alpha, \boldsymbol{w}, \mathbf{x}, \mathbf{y}, \hat{f}_k)$;
  if $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_{-k}) > t$ **then** $\hat{S}(\mathcal{E}) \leftarrow \hat{S}(\mathcal{E}) \cup \{k\}$;
end

weaker than $H_{0,S}(\mathcal{E})$ (see Section 3.1 from Peters et al. (2016)). Moreover, most of the known identifiability conditions for ICP (i.e., Theorem 2 from Peters et al. (2016)) apply here since their proofs rely only on the invariant distribution of the residuals. Thus, those conditions are also sufficient to have $\tilde{S}(\mathcal{E}) = S^*$. Finally, a weaker null hypothesis means that the WVM algorithm would also work under less restrictive conditions than Assumption 1. In particular, the independence condition in (1) excludes any possibility of a hidden confounder between Y and X_{S^*}.

Peters et al. (2016) also considers a more general setting with instrumental variables that allows the presence of hidden confounders; they show that ICP may be adapted to this setting at the cost of having to perform an extensive grid search over all regressors in order to test each null hypothesis $H_{0,S}(\mathcal{E})$. Under this general setting, the WVM algorithm may be used to recover the same set of direct causes as ICP, and its advantage there is twofold since it avoids the combinatorial search of Equation (2) and also the extensive grid search; we include details in Appendix B.

The statistic that we use for testing $\tilde{H}_{0,k}(\mathcal{E})$ is the minimal value (4), where each distribution $\nu_e(f)$ is replaced by its empirical counterpart $\hat{\nu}_e(f) \doteq n_e^{-1} \sum_i \delta_{y_i^e - f(x_i^e)}$. More precisely, we compute $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_{-k}) \doteq \inf_{f \in \mathcal{F}_{-k}} \mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(f))$ for every $k \in [p]$, and we reject $\tilde{H}_{0,k}(\mathcal{E})$ whenever it is above some threshold; we also call $\hat{f}_k$ the resulting minimizer (see Algorithm 1). We discuss how these thresholds are chosen and how the optimization is performed below.

Connection with Likelihood Ratio Tests. The test we propose is similar to the classical *likelihood ratio test* (LRT). If we were interested in testing the statistical significance of a predictor X_k in a regression model parametrized by $\theta \in \Theta$, we might use a LRT to test whether $\inf_{\theta \in \Theta_k} -l(\theta) - \inf_{\theta \in \Theta} -l(\theta)$ is zero or strictly positive, where $l(\theta)$ is the log likelihood and $\Theta_k \subset \Theta$ is a restricted model that excludes X_k from the regression. Under Assumption 1, we can rewrite $\tilde{H}_{0,k}(\mathcal{E})$ as $\Gamma_{\boldsymbol{w}}(\mathcal{F}_{-k}) - \Gamma_{\boldsymbol{w}}(\mathcal{F}) = 0$. Thus, the WVM test essentially replaces the negative log likelihood from the LRT, which measures lack-of-fit, by the Wasserstein variance, which measures distributional variability instead.

Note that we can use a LRT to test a more restricted model, say Θ_S, that excludes a subset S of predictors such that $|S| \geq 1$; this extension to subset exclusion is another advantage of WVM over ICP. Then, WVM may be used to detect whether *at least one* of the predictors from S is causal, which can be useful in situations where these predictors are correlated and thus their effects are hard to distinguish statistically. ICP generally cannot be extended to subset exclusion. We discuss this extension in Appendix C.

4 THEORETICAL ANALYSIS

In this section, we first establish a new uniform bound for finite samples between the Wasserstein variance and its empirical counterpart in terms of the Rademacher complexity (Shalev-Shwartz and Ben-David, 2014). This guaranties that the Wasserstein variance is no more prone to over-fitting than any of the classical loss functions used in machine learning, and in particular that $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_{-k})$ will get close to $\Gamma_{\boldsymbol{w}}(\mathcal{F}_{-k})$ in finite samples for a suitable function class $\mathcal{F}_{-k}$. The proof is in Appendix D.

Theorem 1 (Uniform Bound)**.** *Let $\delta \in (0,1)$ and $\mathcal{G}$ be some class of functions of the predictors. If, for each $e \in [E]$, the variable $Z_e \doteq \sup_{g \in \mathcal{G}} |Y^e - g(X^e)|$ is sub-Gaussian, then with probability at least $1-\delta$ we have:*

$$\forall g \in \mathcal{G}, \quad |WV_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(g)) - WV_{\boldsymbol{w}}(\boldsymbol{\nu}(g))| \tag{6}$$
$$\leq \sum_{e=1}^{E} w_e \left(\frac{A_{\delta,n}}{\sqrt{n_e}} + B_{\delta,n}(1+\log(n_e))\mathcal{R}_{n_e}(\mathcal{G}) \right) + \frac{C_{\delta,n}}{n},$$

where $\mathcal{R}_{n_e}(\mathcal{G})$ is the Rademacher complexity of $\mathcal{G}$ under environment e (see Definition 5 in Appendix D), and $A_{\delta,n}, B_{\delta,n}, C_{\delta,n} = O(\log(n/\delta))$. Also, if the variables Z_e are bounded with probability one, then $A_{\delta,n}, B_{\delta,n}, C_{\delta,n}$ are just $O(\log(1/\delta))$. As a consequence, the bound from (6) *is also verified for $|\inf_{g \in \mathcal{G}} WV_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(g)) - \inf_{g \in \mathcal{G}} WV_{\boldsymbol{w}}(\boldsymbol{\nu}(g))|$ with probability at least $1-\delta$.*

We derived the bound in Theorem 1 for more general function classes than $\mathcal{F}$. Often in practice the minimizer of a loss over some (large) class $\mathcal{F}$ belongs to a more restricted class $\mathcal{G} \subset \mathcal{F}$, which is more useful for the convergence analysis (Bartlett et al., 2005). Theorem 1 establishes a guarantee on the ability of the WVM algorithm to recover $\tilde{S}(\mathcal{E})$ in finite samples, and is used for the asymptotic results of Theorem 2. This bound (Equation (6)) shows that, with high probability

and enough data, $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_{-k})$ is close to $\Gamma_{\boldsymbol{w}}(\mathcal{F}_{-k})$ for all k, and therefore it is possible to distinguish the identifiable causal predictors from the others. This means that there exists a choice of threshold t in Algorithm 1 such that the output $\hat{S}(\mathcal{E})$ is equal to $\tilde{S}(\mathcal{E})$ with high probability when n is large enough; below we discuss how to choose this threshold.

Theorem 1 also offers insight on how to choose the weights $\boldsymbol{w}$. Since the Rademacher complexity converges to zero for the usual classes of functions, in general at a $\tilde{O}(n_e^{-1/2})$ rate (Bartlett and Mendelson, 2002), bound (6) suggests that we use smaller weights in environments with less data. In the next section, we set $w_e = n_e/n$, which leads to a $\tilde{O}(n^{-1/2})$ bound in Equation (6).

Setting the Thresholds. We show how the thresholds can be set based on the asymptotic distribution of the Wasserstein variance under $\tilde{H}_{0,k}(\mathcal{E})$. For a probability distribution ν on the real line, and $\hat{\nu}_n$, its empirical estimate with n samples, the asymptotic distribution of $nW_2^2(\hat{\nu}_n, \nu)$ has already been established in the literature; see Del Barrio et al. (2005) for a complete treatment. More precisely, under some technical conditions we have:

$$nW_2^2(\hat{\nu}_n, \nu) \xrightarrow[n\to\infty]{d} \int_0^1 B^2(t)\, q^2(t) dt, \tag{7}$$

where $(B(t))_{t\in[0,1]}$ is the Brownian bridge between 0 and 1, i.e., a Gaussian process with covariance function $\eta(s,t) \doteq t \wedge s - st$, and q is the quantile density of ν, i.e., the derivative of its quantile function. The asymptotic result of Equation (7) holds when the CDF of ν is twice differentiable and $\int_0^1 \eta(t,t)q^2(t)dt$ is finite, along with other regularity conditions (Del Barrio et al., 2005); we provide the full list of these conditions in Appendix E.1. Similarly, and under the same set of conditions, we derive the following asymptotic result for the Wasserstein variance:

Proposition 1. *Assume that data from different environments are independent of each other, and set $w_e = n_e/n$. Let f be any function in $\mathcal{F}$ such that $WV_{\boldsymbol{w}}(\boldsymbol{\nu}(f)) = 0$, i.e., there exists $\nu(f)$ such that $\forall e \in [E]$, $\nu_e(f) = \nu(f)$. Assuming $\nu(f)$ respects the assumptions needed for Equation (7) to hold, we have:*

$$n\, WV_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(f)) \xrightarrow[n_0\to\infty]{d} \sum_{e=1}^{E-1} \int_0^1 B_e^2(t) q_f^2(t) dt, \tag{8}$$

where q_f is the quantile density of $\nu(f)$, and $(B_e(t))_{e=1}^{E-1}$ are $E-1$ independent Brownian bridges.

To test $\tilde{H}_{0,k}(\mathcal{E})$ at confidence level α, we set the threshold at the $(1-\alpha)$-quantile of the limit distribution from (8), for $f = \hat{f}_k$. Because the quantile density q_f is unknown, we propose to estimate it within each environment by a kernel quantile density estimator (Sheather and Marron, 1990; Jones, 1992).

Algorithm 2: get_threshold

Input: $\alpha, \boldsymbol{w} = n_e/n, \mathbf{x}, \mathbf{y}, f = \hat{f}_k$
Output: $\hat{t}_\alpha$
for $e \in [E]$ **do**
  Estimate $\hat{q}_e$ using kernel estimator (9) ;
end
Set $\hat{t}_\alpha$ as the $(1-\alpha)$-quantile of variable (10) ;

Definition 4 (Kernel quantile density estimator)**.** *Denote by $(\epsilon^e_{(i)}(f))_{i=1}^{n_e}$ the residuals for function f in environment e and sorted in increasing order; we use the following quantile density estimator with bandwidth $h_e \propto n_e^{-1/3}$ to estimate the quantile density of $\nu_e(f)$, for any $t \in [0,1]$:*

$$\hat{q}_f^e(t) \doteq \sum_{i=2}^{n_e} \left(\epsilon^e_{(i)}(f) - \epsilon^e_{(i-1)}(f)\right) K_{h_e}\left(t - \frac{i-1}{n_e}\right), \tag{9}$$

where $K_h(u) \doteq h^{-1}K(u/h)$, and K is a Lipschitz kernel supported on $[-1,1]$ such that $\int K(u)du = 1$.

In Theorem 2 we show that, by choosing the threshold as discussed above, with the quantile density replaced by its estimator from Definition 4, we get a consistent test of level α for $\tilde{H}_{0,k}(\mathcal{E})$. We need however to impose some additional assumptions; in particular, we assume that $\hat{f}_k$ is a bounded function in a Sobolev space, a class of functions used in nonparametric statistics (Tsybakov, 2008). A detailed list of these regularity conditions in Assumption 2 may be found in Appendix E.3.

Assumption 2 (Summary of the reg. conditions)**.** *For any $e \in [E]$, X^e is bounded with probability one, Y^e is sub-Gaussian, and $n_e \geq \lambda n$ for a constant $\lambda > 0$. Also, data from different environments are independent. For n large enough, and any $k \in [p]$ we have with high probability that $\hat{f}_k$ belongs to a fixed bounded set in a Sobolev space $W^{d,2}$ with $d > p/2$. Furthermore, uniformly over all functions f in this set and $e \in [E]$, $\nu_e(f)$ satisfies the conditions needed for the asymptotic result in Equation (7) to hold.*

Now we present Theorem 2; see Appendix E for proof.

Theorem 2 (Asymptotic guaranties)**.** *Assume Assumption 2 is true and let $k \in [p]$. For every $e \in [E]$, set $w_e = n_e/n$ and, for simplicity, call $\hat{q}_e$ the quantile density estimator from Equation (9) for $f = \hat{f}_k$, the minimizer of $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_{-k})$. Set $\hat{q} \doteq \sum_{e=1}^E w_e \hat{q}_e$ and let $\hat{t}_\alpha$ be the $(1-\alpha)$-quantile of the variable:*

$$\frac{1}{n}\sum_{e=1}^{E-1} \int_0^1 B_e^2(t)\, \hat{q}^2(t) dt, \tag{10}$$

where $(B_e(t))_{e=1}^{E-1}$ *are* $E-1$ *independent Brownian bridges between* 0 *and* 1. *Rejecting* $\tilde{H}_{0,k}(\mathcal{E})$ *whenever* $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_{-k}) > \hat{t}_\alpha$ *forms a consistent test of asymptotic level* α. *That is:*

$$\textit{Under } \tilde{H}_{0,k}(\mathcal{E}) : \limsup_{n\to\infty} \mathbb{P}(\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_{-k}) > \hat{t}_\alpha) \leq \alpha,$$
$$\textit{and under } \tilde{H}_{1,k}(\mathcal{E}) : \lim_{n\to\infty} \mathbb{P}(\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_{-k}) > \hat{t}_\alpha) = 1.$$

In the appendix we prove a slightly more general result than Theorem 2 to allow for the use of function classes that depend on the sample size (Theorem 4 in E.3).

Multiple Testing Correction. Theorem 2 says that, for each $k \in [p]$, testing for $\tilde{H}_{0,k}(\mathcal{E})$ has an asymptotic probability less than α to return a false positive. Since we need to perform p tests, if we want to control for the total number of false positives, one option is to correct for these multiple tests by choosing a lower α. One possibility, to control the family-wise error rate (FWER), is to use Bonferroni correction. We argue that, for WVM, such corrections may lead to conservative results. In general, Bonferroni correction is appropriate in situations where the tests are mostly independent of one another. In the case of WVM, however, the statistics $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_{-k})$ for $k \notin S^*$ may often be well correlated since they are all bounded by the same quantity $\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(f^*))$ that converges to 0. As we show in Theorem 3, correction is not needed when identifiability holds, since under identifiability the thresholds derived in Theorem 2 for $k \notin S^*$ all converge toward the $(1-\alpha)$-quantile of the limit distribution in (8) for $f = f^*$. In that case, the probability of *any* false positive across all of the tests is already bounded by α asymptotically. By identifiability, we mean that f^* is the only function in the closure of $\mathcal{F}$ such that $\mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(f^*)) = 0$. In the case of linear functions, the sufficient conditions of Theorem 2 from Peters et al. (2016) also imply identifiability in this sense.

Theorem 3 (When no correction is needed)**.** *Assume Assumptions 1 and 2 are true, and set* $w_e = n_e/n$. *Call* $\hat{S}(\mathcal{E})$ *the output of Algorithm 1 where the confidence level for the thresholds returned by Algorithm 2 is set at a fixed* $\alpha > 0$. *Under identifiability of Equation* (1)*, we have:*

$$\liminf_{n\to\infty} \mathbb{P}(\hat{S}(\mathcal{E}) = \tilde{S}(\mathcal{E})) = \liminf_{n\to\infty} \mathbb{P}(\hat{S}(\mathcal{E}) = S^*) \geq 1-\alpha.$$

5 IMPLEMENTATION DETAILS

We now explain how the optimization of $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_{-k})$ is performed, and how we approximate the distribution of Equation (10) to obtain the thresholds. Additional details can be found in Appendix F.

Optimization. Since the residuals are one dimensional, the Wasserstein variance here admits a closed form. More precisely, for any collection of distributions $\boldsymbol{\nu} = (\nu_e)_{e=1}^E$ defined on $\mathbb{R}$ let F_e^{-1} denote the quantile function of ν_e. By remarks 2.30 and 9.6 from Peyré et al. (2019), we have:

$$\mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}) = \int_0^1 \sum_{e=1}^E w_e \cdot \left(F_e^{-1}(t) - \sum_{e'=1}^E w_{e'} F_{e'}^{-1}(t) \right)^2 dt. \tag{11}$$

The closed form (11) may be efficiently computed when each ν_e is an empirical distribution such as $\hat{\nu}_e(f)$; this mainly requires sorting the residuals in each environment. Another useful property of Equation (11) is that when the functions are parametrized, that is, $f(\cdot) = f(\cdot\,;\theta)$ for some θ, the Wasserstein variance $\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(f(\cdot\,;\theta)))$ is almost everywhere differentiable w.r.t. θ. Thus, to minimize $\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(f(\cdot\,;\theta)))$, one can use any gradient-based optimization method to obtain $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_{-k})$; we use L-BFGS in our experiments.

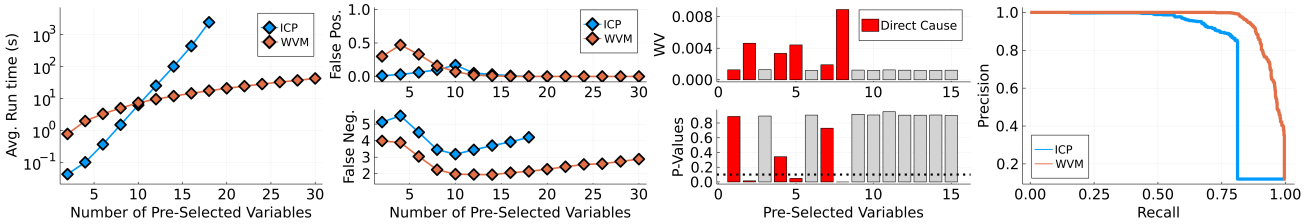
Approximation of the Asymptotic Distribution. From Del Barrio et al. (2005), the RHS of Equation (7) is a generalized χ-square distributed variable, and so can be expressed in distribution as the sum $\sum_i \lambda_i Z_i^2$, where the Z_is are i.i.d. standard normal variables and the λ_is are the eigenvalues of the integral operator with kernel $\eta'(s,t) = \eta(s,t)q(s)q(t)$. The generalized χ-square distribution may be accurately approximated by a Gamma distribution with the same mean and variance (Gretton et al., 2007; Johnson et al., 1995; Kankainen, 1995); in our case, the mean and variance may be expressed in terms of the trace and Hilbert-Schmidt norms of the above-mentioned integral operator:

Proposition 2. *Let* $\hat{\eta}(s,t) \doteq \eta(s,t)\hat{q}(s)\hat{q}(t)$ *and* $E' \doteq E-1$. *The mean* $\hat{m}$ *and variance* $\hat{\sigma}^2$ *of* (10) *are as follows:*

$$\hat{m} = \frac{E'}{n}\int_0^1 \hat{\eta}(t,t)dt, \quad \hat{\sigma}^2 = \frac{2E'}{n^2}\iint_0^1 \hat{\eta}^2(s,t)dsdt.$$

We can therefore approximate the distribution of Equation (10) as a Gamma distribution with shape parameter $\hat{\alpha} = \hat{m}^2/\hat{\sigma}^2$ and scale parameter $\hat{\theta} = \hat{\sigma}^2/\hat{m}$. In practice, one can estimate the above integrals using Monte Carlo integration.

Bootstrap Approximation. Since the statistics $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_{-k})$ result from a minimization, the thresholds based on the asymptotic distribution in Equation (10) may be too conservative in finite samples. Instead, we find empirically that using a bootstrap estimate of the expectation and variance of $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F})$ to set the Gamma distribution leads to a better and less conservative threshold. We use this heuristic in our experiments.



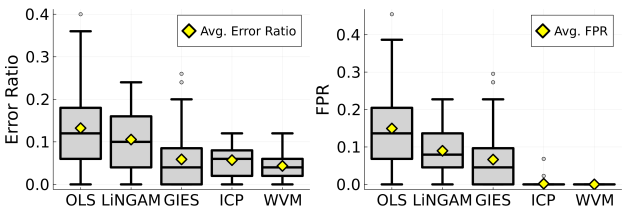
(a) (b) (c) (d)

Figure 2: (a): Run time in seconds for different numbers of pre-selected variables for ICP and WVM, averaged over the 100 simulations; we stopped at 18 variables for ICP as it took > 50 hours for this data-point. (b): Average number of false positives (top) and false negatives (bottom) for different numbers of pre-selected variables. (c): An example of the outputs of WVM with 15 pre-selected variables. (d): Precision-recall curves for ICP and WVM, averaged over the 100 simulations.

6 EXPERIMENTS

We now analyze the performance of the WVM algorithm compared to related algorithms in simulations. Additional experiments and details are in Appendix G. The code to reproduce all experiments is available at `https://github.com/astrzalk/WVM_reproducibility`.

Data Generating Process. We focus on the case where the causal model in Equation (1) is linear, i.e., $Y^e = \beta^{*T} X^e + \varepsilon^e$, where $\beta^*_k = 0$ for $k \notin S^*$. In other words, we consider $\mathcal{F} = \{f : f(x;\beta) = \beta^T x,\ \beta \in \mathbb{R}^p\}$ and $\mathcal{F}_{-k}$ consists of all functions $f(\,\cdot\,;\beta)$ from $\mathcal{F}$ such that $\beta_k = 0$. For our simulations, we use linear SCMs with independent Gaussian noise (i.e., no hidden confounders) for the observations. We sample 100 random graphs with 51 variables ($p = 50$) with average degree 12, and we fix for all graphs the number of direct causes to be $|S^*| = 6$. For each of these settings, the graph coefficients and noise variances are randomly sampled from uniform distributions. We generate four interventional environments by applying simple mechanism change interventions on random subsets of the variables (excluding the target variable). This leads to a total of $E = 5$ environments. For each environment, we generate $n_e = 500$ i.i.d. samples resulting in $n = 2500$ samples in total across all environments.

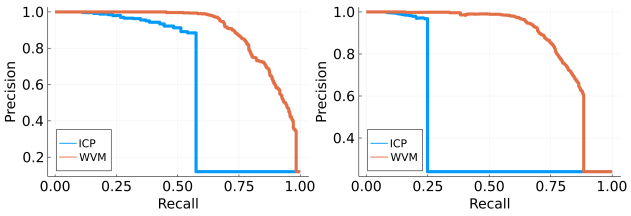


(a) (b)

Figure 3: (a): Error ratios of the algorithms. (b): False positive rates (FPR) of the algorithms.

Benchmarking. We compare WVM with naive OLS regression and three baseline causal discovery algorithms: LiNGAM (Shimizu et al., 2011), GIES (Hauser and Bühlmann, 2012), and ICP (Peters et al., 2016). ICP and WVM both use confidence level $\alpha = 0.1$; the inferred direct causes for OLS are the significant predictors at level α/p. As in ICP, we preselect variables using Lasso before applying WVM to improve its power. However we are not constrained by computation and can preselect as many predictors as desirable. We preselect 18 variables for this experiment while ICP fixes the total number of preselected variables at 8. Furthermore, LiNGAM is applied on the aggregated dataset across environments, and we specify that all non-target variables are intervened on for GIES. Across methods, let $\hat{S} \subset [p]$ denote the inferred direct causes for the target. Define the false positives as $FP \doteq \{k \in [p] : k \in \hat{S} \text{ and } k \notin S^*\}$, and similarly the false negatives as $FN \doteq \{k \in [p] : k \notin \hat{S} \text{ and } k \in S^*\}$. To evaluate the performance of causal discovery algorithms, we use the *Error Ratio* $\doteq (|FP| + |FN|)/p$ and the false positive rate $FPR \doteq |FP|/(p - |S^*|)$. We also consider the *Precision* $\doteq 1 - |FP|/|\hat{S}|$ and the *Recall* $\doteq 1 - |FP|/|S^*|$. WVM outperforms ICP and the other algorithms in terms of the error ratio (Figure 3a), and behaves similarly to ICP in terms of the false positive rate (Figure 3b).

Further Comparison between WVM and ICP. We now investigate the run time and power of WVM and ICP for different numbers of preselected variables. For moderate to large numbers of preselected variables, ICP's exponential scaling is much slower than WVM's runtime (Figure 2a). For instance, with 18 preselected variables, ICP takes 2443s on average while WVM takes only 17s, a 100 times speed-up. WVM's power is also less sensitive to the number of preselected variables (Figure 2b), and WVM identifies 1 to 2 (out of 6) more causes on average than ICP when applied on the same set of preselected variables. Even though Equation (5)



(a) (b)

Figure 4: Average precision-recall curves of ICP and WVM when (a): $n_e = 100$; and (b): $|S^*| = 12$.

suggests for infinite data WVM may be less powerful than ICP, for finite samples the converse is often true, since ICP's output is an intersection of exponentially many accepted sets of potential causes.

The statistics $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_{-k})$ returned by WVM are good indicators of the "strength" of a potential cause, and in many situations it is possible to identify causal predictors by looking at these values (Figure 2c). We further analyze the potential of these statistics to recover direct causes by looking at the precision-recall curve constructed for different choices of thresholds. In particular, WVM often recovers more causes than ICP with higher precision (Figure 2d).

Additional Settings. We consider two additional settings, one where the sample size is reduced to $n_e = 100$, and another where the number of direct causes is set to $|S^*| = 12$. The average precision-recall curves for ICP and WVM again show that WVM recovers more causes than ICP with higher precision (Figures 4a and 4b).

The advantages of WVM over ICP are more prominent in these situations. When $|S^*| = 12$, ICP's performance quickly deteriorates compared to WVM (Figure 4b). This scenario is of interest since ICP may often require fewer than $|S^*|$ preselected variables. This shows that ICP's computational complexity constrains its statistical power, and that WVM's practical improvement over ICP is more than run time.

7 DISCUSSION

In this paper we show that causal inference using ICP may be reformulated as a multiple hypothesis testing problem with only p tests to perform, compared to the 2^p tests that the original ICP requires. Each of those tests is similar to a likelihood ratio test, where the negative log likelihood is replaced by a new loss function that we call Wasserstein variance, which quantifies the distributional variability of the residuals across environments. WVM is nonparametric and can easily adapt to more general settings than ICP (see remarks after Definition 3 and Appendix B). We derived asymptotic guarantees on the ability of WVM to recover the direct causes with a limited number of false positives, and our simulations confirm our theoretical results.

There are possible improvements and extensions that we leave for future work. In practice, the thresholds based on our asymptotic results and bootstrap approximation may sometimes be conservative. Therefore, deriving a more accurate limit distribution for the statistics $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_{-k})$ under $\tilde{H}_{0,k}(\mathcal{E})$ for some specific classes of functions is of interest. We would like to stress however that under our rather weak assumptions on the class of functions $\mathcal{F}$, the asymptotic distribution in (10) is the best achievable limit distribution – when $\mathcal{F}_{-k}$ is finite and f^* is identifiable, the distributions of $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_{-k})$ and (10) coincide asymptotically under $\tilde{H}_{0,k}(\mathcal{E})$. Finally, the primary assumption that WVM relies on is the additive noise specification from Equation (1). Even though additive noise models are used by many causal discovery algorithms, ICP included, such an assumption may be too restrictive in some situations. Adapting WVM to more general functional relationships with nonadditive noise is another question left for future work.

References

Robert A Adams and John JF Fournier. *Sobolev spaces*. Elsevier, 2003.

Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

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

Peter Bühlmann, Jonas Peters, Jan Ernest, et al. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Erhan Cinlar. *Probability and stochastics*, volume 261. Springer, 2011.

Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2002.

Lokenath Debnath and Piotr Mikusinski. *Introduction to Hilbert spaces with applications*. Academic press, 2005.

Eustasio Del Barrio, Evarist Giné, Frederic Utzet, et al. Asymptotics for l2 functionals of the empirical quantile process, with applications to tests of fit based on weighted wasserstein distances. *Bernoulli*, 11(1): 131–189, 2005.

RM Dudley. Universal donsker classes and metric entropy. *The Annals of Probability*, pages 1306–1326, 1987.

Michael Falk. On the estimation of the quantile density function. *Statistics & Probability Letters*, 4(2):69–73, 1986.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Kun Zhang. Learning causal structures using regression invariance. In *Advances in Neural Information Processing Systems*, pages 3011–3021, 2017.

Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, Alexander J Smola, et al. A kernel statistical test of independence. In *Nips*, volume 20, pages 585–592. Citeseer, 2007.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13(1):2409–2464, 2012.

Alain Hauser and Peter Bühlmann. Jointly interventional and observational data: estimation of interventional markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 291–318, 2015.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018.

Norman L Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. *Continuous univariate distributions, volume 2*, volume 289. John wiley & sons, 1995.

M Chris Jones. Estimating densities, quantiles, quantile densities and density quantiles. *Annals of the Institute of Statistical Mathematics*, 44(4):721–727, 1992.

Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11): 1–26, 2012. URL `https://www.jstatsoft.org/article/view/v047i11`.

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

Annaliisa Kankainen. *Consistent testing of total independence based on the empirical characteristic function*. 1995.

Christian Kleiber and Achim Zeileis. *Applied Econometrics with R*. Springer-Verlag, New York, 2008. URL `https://CRAN.R-project.org/package=AER`. ISBN 978-0-387-77316-2.

Olivier Marchal and Julyan Arbel. On the subgaussianity of the beta and dirichlet distributions. *Electronic Communications in Probability*, 22:1–14, 2017.

Nicolai Meinshausen, Alain Hauser, Joris M Mooij, Jonas Peters, Philip Versteeg, and Peter Bühlmann. Methods for causal inference from gene perturbation experiments and validation. *Proceedings of the National Academy of Sciences*, 113(27):7361–7368, 2016.

Axel Munk and Claudia Czado. Nonparametric validation of similar distributions and assessment of goodness of fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):223–241, 1998.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 5(78):947–1012, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference*. The MIT Press, 2017.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11 (5-6):355–607, 2019.

Dominik Rothenhäusler, Christina Heinze, Jonas Peters, and Nicolai Meinshausen. Backshift: Learning causal cyclic graphs from unknown shift interventions. In *Advances in Neural Information Processing Systems*, pages 1513–1521, 2015.

Dominik Rothenhäusler, Peter Bühlmann, Nicolai Meinshausen, et al. Causal dantzig: fast inference in linear structural equation models with hidden variables under additive interventions. *The Annals of Statistics*, 47(3):1688–1722, 2019.

Cecilia Elena Rouse. Democratization or diversion? the effect of community colleges on educational attainment. *Journal of Business & Economic Statistics*, 13(2):217–224, 1995.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

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

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Simon J Sheather and James Stephen Marron. Kernel quantile estimators. *Journal of the American Statistical Association*, 85(410):410–416, 1990.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *The Journal of Machine Learning Research*, 12:1225–1248, 2011.

Liam Solus, Yuhao Wang, and Caroline Uhler. Consistency guarantees for greedy permutation-based causal inference algorithms. *arXiv preprint arXiv:1702.03530*, 2017.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

Chandler Squires, Yuhao Wang, and Caroline Uhler. Permutation-based causal structure learning with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence*, pages 1039–1048. PMLR, 2020.

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

Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Yuhao Wang, Liam Solus, Karren Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. In *Advances in Neural Information Processing Systems*, pages 5822–5831, 2017.

Supplementary Material: Variance Minimization in the Wasserstein Space for Invariant Causal Prediction

A PROOF OF LEMMA 1

Recall that Assumption 1 implies that $H_{0,S^*}(\mathcal{E})$ is true. Equation (3) then results directly from the following chain of logical equivalences:

$$\begin{aligned} k \in S(\mathcal{E}) &\iff \forall S \subseteq [p] \text{ s.t. } H_{0,S}(\mathcal{E}) \text{ is true, } k \in S \\ &\iff \nexists S \subseteq [p]/\{k\} \text{ s.t. } H_{0,S}(\mathcal{E}) \text{ is true} \\ &\iff H'_{0,k}(\mathcal{E}) \text{ is false.} \end{aligned}$$

B A MORE GENERAL SETTING

Peters et al. (2016) propose some extensions of ICP to settings where Assumption 1 is violated (e.g., we refer to Section 5 in their paper). In particular, they consider the case of a general form of SCM that allows for the presence of hidden confounders and feedback loops between the target and the causal predictors, and only imposes that the environment variable $I \in [E]$ acts as an instrumental variable on the predictors X (see Figure 5 below). We show in this section that WVM is directly applicable in this setting, whereas ICP's extension is computationally intractable.

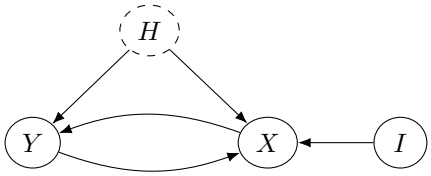


Figure 5: An SCM with an unobserved confounder, node H in the graph, with a feedback cycle between X and Y, and where the instrumental variable $I \in [E]$ indicates from which environment an observation is drawn.

The SCM illustrated in Figure 5 contains a hidden confounder H and a feedback loop between variables X and Y. Concretely, the underlying system of structural equations for the above SCM is:

$$\begin{cases} X = h(I, H, Y, \eta), \\ Y = f^*(X) + g(H, \varepsilon), \end{cases} \tag{12}$$

where $f^* \in \mathcal{F}_{S^*}$, and H, I, ϵ, η are mutually independent.

Note that Assumption 1 no longer holds for the structural equations (12) as the residual $g(H, \varepsilon)$ is no longer independent of the covariates. To deal with this more general setting Peters et al. (2016) introduced a relaxed null hypothesis that removes the assumption of the independence of the noises from the covariates:

$$H_{0,S,hidden}(\mathcal{E}) : \left\{ \exists f \in \mathcal{F}_S \text{ such that the distribution of } Y^e - f(X^e) \text{ is identical for all } e \in \mathcal{E}. \right.$$

$H_{0,S^*,hidden}(\mathcal{E})$ is true under model (12), since the environment variable I is independent of H and ϵ. Given this new weaker null hypothesis, they propose to recover the set $S_H(\mathcal{E})$ of identifiable causal predictors under model (12) defined as:

$$S_H(\mathcal{E}) \doteq \bigcap_{S:\, H_{0,S,hidden}(\mathcal{E}) \text{ is true}} S \subseteq S^*. \tag{13}$$

It turns out however that it is computationally challenging to test for each hypothesis $H_{0,S,hidden}(\mathcal{E})$. The main reason is that we can no longer use regression techniques to recover the residuals since they are dependent on the covariates. The only solution Peters et al. (2016) propose for testing $H_{0,S,hidden}(\mathcal{E})$ is to go through all functions in $\mathcal{F}_S$ (or over some approximating grid of it) and to check if for at least one of them the resulting residuals have an invariant distribution across environments. Such an approach is of course quite intractable in practice.

On the other hand, WVM can recover without any modification, and thus in a tractable way, the set $S_H(\mathcal{E})$ from the data. To see this, first define $H'_{0,k,hidden}(\mathcal{E})$ as follows:

$$H'_{0,k,hidden}(\mathcal{E}) : \begin{cases} \exists S \not\ni k, \exists f \in \mathcal{F}_S \text{ and a fixed distribution } D \text{ s.t.} \\ \text{for all } e \in \mathcal{E}, Y^e - f(X^e) \sim D. \end{cases} \tag{14}$$

Note by the same reasoning as the proof for Lemma 1, we have that $S_H(\mathcal{E}) = \{k : H'_{0,k,hidden}(\mathcal{E}) \; is \; false\}$. The important observation is that, by Lemma 2, $H'_{0,k,hidden}(\mathcal{E})$ is false if and only if $\tilde{H}_{0,k}(\mathcal{E})$ is false.[1] Therefore, the sets of identifiable causes for the WVM algorithm and for the above extension of ICP are exactly the same, that is $\tilde{S}(\mathcal{E}) = S_H(\mathcal{E})$; compared with this extension of ICP however, WVM is much more computationally efficient.

C AN EXTENSION OF THE WVM TEST TO BLOCKS OF VARIABLES

It is possible to extend WVM to detect whether there is a direct cause among a set S of several predictors, instead of testing for each of the predictors separately using $\Gamma_{\boldsymbol{w}}(\mathcal{F}_{-k})$. Let $\mathcal{F}_{-S}$ denote the set of functions in $\mathcal{F}$ that don't depend on any of the predictors from S and consider the following hypotheses:

$$\tilde{H}_{0,S}(\mathcal{E}) : \;\; \Gamma_{\boldsymbol{w}}(\mathcal{F}_{-S}) = 0, \quad \text{against} \quad \tilde{H}_{1,S}(\mathcal{E}) : \;\; \Gamma_{\boldsymbol{w}}(\mathcal{F}_{-S}) > 0.$$

From Assumption 1 (or even under the more general setting considered in Section B), we have that $\Gamma_{\boldsymbol{w}}(\mathcal{F}_{-S}) = 0$ whenever the set S does not include any of the direct causes. Therefore, if we observe with enough confidence that $\Gamma_{\boldsymbol{w}}(\mathcal{F}_{-S}) > 0$ then we can conclude that S contains at least one direct cause. Such a test is apparently not possible within the framework of ICP since the independence property of the noise from Assumption 1 can easily be violated when we group variables together; for instance, when S includes a variable dependent on the residual from (2) (e.g. a descendant of Y).

We can push this extension of WVM further by considering a partition $\mathcal{P} \doteq \{S_1, \dots, S_m\}$ of the p predictors and in the same spirit as Equation (5), we can seek to recover the collection of identifiable blocks of variables containing at least one cause:

$$\tilde{S}_{\mathcal{P}}(\mathcal{E}) \doteq \left\{ S_i : \tilde{H}_{0,S_i}(\mathcal{E}) \text{ is false}, i \in [m] \right\}. \tag{15}$$

Grouping variables and testing with WVM in such a way can be beneficial in situations where some of the variables are highly correlated. To see this, consider the situation where a predictor k_1 is a direct cause and another predictor k_2 is highly correlated with k_1. In this case, $\Gamma_{\boldsymbol{w}}(\mathcal{F}_{-k_1})$ can be equal to 0, or close to it, since k_1 in the regression can easily be substituted by k_2. On the other hand, by grouping them in a set S we might have $\Gamma_{\boldsymbol{w}}(\mathcal{F}_{-S})$ large enough to detect that at least one of them is causal. Therefore, using such an extension may potentially recover more information about the causal structure of the data when some of the predictors are highly correlated.

This issue also was discussed for ICP in Heinze-Deml et al. (2018). Indeed, ICP can return an empty set in the presence of highly correlated variables for the same reason discussed above. The authors propose to change the output of ICP so that it includes defining sets (see Section 2.2 in Heinze-Deml et al. (2018)), in which at least one variable is a direct cause with high probability. However, the concept of a defining set is hard to translate to the WVM algorithm. Instead, in the situation where some variables are highly correlated, we propose to first group the predictors into a collection $\mathcal{P}$ of clusters of highly correlated variables (where such clusters can potentially contain only one predictor) and then use WVM as mentioned above to recover the set $\tilde{S}_{\mathcal{P}}(\mathcal{E})$ from Equation (15).

[1] As a point of rigor, this equivalence might not be true for some classes of functions $\mathcal{F}$. Indeed, it is technically possible to have $\Gamma_{\boldsymbol{w}}(\mathcal{F}_{-k}) = 0$ while $\mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(f)) > 0$ for any $f \in \mathcal{F}_{-k}$ for some k. Note that when this happens, the set of identifiable causal predictors for WVM will be smaller than $S_H(\mathcal{E})$, similarly to what we saw in Definition 3. However, we believe that for the usual classes of functions one encounters in practice this equivalence holds. This is true for instance when the predictors are bounded and $\mathcal{F}$ is a class of linear functions with bounded coefficients; then in that case $\inf_{f \in \mathcal{F}_{-k}} \mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(f))$ is in fact a minimum – this is a direct consequence of the "continuity" of the Wasserstein variance as expressed in Lemma 8.

D PROOF OF THEOREM 1

In order to prove Theorem 1, we use the dual formulation of the Wasserstein barycenter optimization problem from Agueh and Carlier (2011) as an alternative expression for the Wasserstein variance. Recall from Definition 2 that the optimal value of this optimization problem is simply what we call the Wasserstein variance. As we shall see below, this formulation will be useful in our derivation of Theorem 1. The following result is an adaptation of Proposition 2.2 form Agueh and Carlier (2011):

Proposition 3 (Proposition 2.2 from Agueh and Carlier (2011))**.** *Define* $\mathcal{C}_{b,2}(\mathbb{R}) \doteq \left\{ f \in C(\mathbb{R}) : \frac{f}{1+|\,.\,|^2} \text{ is bounded} \right\}$, *where* $C(\mathbb{R})$ *is the set of continuous functions defined on* $\mathbb{R}$. *Let* $\boldsymbol{\nu} = (\nu_i)_{i=1}^E$ *be probability distributions from* P_2 *and* $\boldsymbol{w} = (w_i)_{i=1}^E \in \Lambda$ *some weights. Then the Wasserstein variance (Definition 2) admits the following dual formulation:*

$$WV_{\boldsymbol{w}}(\boldsymbol{\nu}) = \sup \left\{ \sum_{e=1}^{E} \int S_{w_e} f_e d\nu_e : \sum_{e=1}^{E} f_e = 0,\ f_e \in \mathcal{C}_{b,2}(\mathbb{R}) \right\}, \tag{16}$$

where

$$S_{w_e} f(x) \doteq \inf_{y \in \mathbb{R}} \left\{ w_e |x-y|^2 - f(y) \right\},\ \forall x \in \mathbb{R}, f \in \mathcal{C}_{b,2}(\mathbb{R}), w_e > 0. \tag{17}$$

One of the main advantages of using the dual formulation of Equation (16) is that it expresses the Wasserstein variance almost as sum of expectations; the only difference is of course the supremum over a subset of $\mathcal{C}_{b,2}$ in front of it. In general, available tools to derive uniform bounds in the spirit of Theorem 1 are essentially meant for loss functions that can be expressed as an expectation of a penalty term, therefore the main difficulty here is the presence of the supremum. We show however that, by a chaining argument and an adaptation of Massart Lemma, this supremum will add only a $O(\log(n_e))$ factor in front of the Rademacher complexity compared to classical uniform bounds (Shalev-Shwartz and Ben-David, 2014).

It is also possible to derive a uniform bound by using the explicit formulation of the Wasserstein variance given in Equation (11) instead of the dual formulation in Equation (16). However, by using this approach in a first attempt we obtained a bound that was slightly worse with a higher power for the log-factor; and the proof was essentially using similar steps and wasn't necessarily shorter. More importantly, the proof we provide based on (16) can be easily adapted to situations where the target is multi-dimensional, while (11) can be used only when Y is one-dimension.

Finally, we prove Theorem 1 by assuming only that the data are independent (more precisely, i.i.d.) *within* each environment but not necessarily *across* environments; that is, for every $e \in [E]$ we assume that the data $(\mathbf{x}_e, \mathbf{y}_e)$ are i.i.d. but not necessarily that $(\mathbf{x}_e, \mathbf{y}_e)$ is independent of $(\mathbf{x}_{e'}, \mathbf{y}_{e'})$ for another environment e'. This means that our bound will also hold in situations where each environment is created by splitting an original observational data set, and where the same observations can appear in different environments; doing so might be useful for instance to increase the number of observations by environment, and thus obtain better bounds.

Before starting our proof, recall the definition of the Rademacher complexity (e.g., see Shalev-Shwartz and Ben-David (2014)):

Definition 5 (Rademacher complexity)**.** *Let* $e \in [E]$ *and* $\boldsymbol{\xi} = (\xi_i)_{i=1}^{n_e}$ *be independent Rademacher variables. For a fixed data* $\boldsymbol{x}_e$ *we define the empirical Rademacher complexity of a class of functions* $\mathcal{G}$ *to be:*

$$R_{\boldsymbol{x}_e}(\mathcal{G}) \doteq \frac{1}{n_e} \mathbb{E}_{\boldsymbol{\xi}} \left[\sup_{g \in \mathcal{G}} \sum_{i=1}^{n_e} \xi_i g(x_i^e) \right].$$

The Rademacher complexity for $\mathcal{G}$ *and environment* e *is then defined as:*

$$\mathcal{R}_{n_e}(\mathcal{G}) \doteq \mathbb{E}_{\boldsymbol{x}_e} \left[R_{\boldsymbol{x}_e}(\mathcal{G}) \right].$$

D.1 Short Discussion of the Assumptions

We briefly discuss the assumptions that the variables $Z_e \doteq \sup_{g \in \mathcal{G}} |g(X^e) - Y^e|$ are either sub-Gaussian or bounded. We argue that such assumptions are not particularly restrictive. For instance, in practice it is reasonable

to assume that there is a large enough constant $M > 0$ (potentially very large) such that all variables X^e and Y^e are bounded (in absolute value) by M; also that for a reasonable choice for $\mathcal{G}$, $g(X^e)$ is uniformly bounded with probability one. Under that scenario the Z_es are therefore bounded with probability one. The main reason we consider the weaker sub-Gaussian assumption is to include the possibility of data generated by a linear Gaussian SCM, a model that is often used in causal inference; in that case if $\mathcal{G}$ is a class composed of linear functions with bounded norm, then the sub-Gaussian assumption holds for the Z_e's – this fact, in addition to the sub-Gaussianity of the next example, can be shown by using point (II) of Theorem 2.1 from Wainwright (2019). Note that nonlinear models are also possible in that case: For instance, if one takes $\mathcal{G}$ to be a bounded subset of an RKHS, and that the related kernel is also bounded with probability one w.r.t. X^e, then if Y^e is sub-Gaussian we have Z_e sub-Gaussian too.

D.2 First Steps

To prove the bound from equation (6) we derive upper-bounds for both $\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(g)) - \mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g))$ and $\mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g)) - \mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(g))$ separately as they need (slightly) different steps. First, we start by bounding the former term uniformly, we then focus on the latter one. Furthermore, in order to improve the exposition of Theorem 1's proof, we often use directly some technical results as lemmas and postpone their proofs to Section D.6.

Let $m = m(\mathbf{x}, \mathbf{y}) \doteq \max_{e\in[E]} \max_{i\in[n_e]} \sup_{g\in\mathcal{G}} |g(x_i^e) - y_i^e|$, and denote $\mathcal{C}_m \doteq \{f \in C(\mathbb{R}) : \forall x \in B(0,m)^c, f(x) = f(mx/|x|)\}$, where $B(0,m)$ refers to the ball (or interval, as we are in $\mathbb{R}$) of center 0 and radius m. In other words, $\mathcal{C}_m$ is the set of continuous functions defined on $\mathbb{R}$ that are constant on $(-\infty, -m]$ and on $[m, +\infty)$. We prove in Lemma 4 that the dual formulation of $\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(g))$ can be written as a supremum over $\mathcal{C}_m$ instead of $\mathcal{C}_{b,2}(\mathbb{R})$; hence by Lemma 4 we have:

$$\begin{aligned}
\textcircled{1} \doteq \sup_{g\in\mathcal{G}} \left(\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(g)) - \mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g))\right) &= \sup_{g\in\mathcal{G}} \left(\sup_{\substack{f_e\in\mathcal{C}_m,\\ \sum_e f_e = 0}} \sum_{e=1}^{E} \int S_{w_e} f_e d\hat{\nu}_e(g) - \sup_{\substack{f'_e\in\mathcal{C}_{b,2},\\ \sum_e f'_e = 0}} \sum_{e=1}^{E} \int S_{w_e} f'_e d\nu_e(g) \right) \\
&\leq \sup_{g\in\mathcal{G}} \sup_{\substack{f_e\in\mathcal{C}_m,\\ \sum_e f_e = 0}} \sum_{e=1}^{E} \int S_{w_e} f_e d(\hat{\nu}_e(g) - \nu_e(g)).
\end{aligned}$$

We can also write:

$$\sum_{e=1}^{E} \int S_{w_e} f_e d\nu_e(g) = \mathbb{E}_{(\mathbf{x}',\mathbf{y}')} \left[\sum_{e=1}^{E} \int S_{w_e} f_e d\hat{\nu}'_e(g) \right],$$

where $(\mathbf{x}', \mathbf{y}')$ is virtual data drawn from the exact same distribution as $(\mathbf{x}, \mathbf{y})$, and $\hat{\nu}'_e(g)$ is defined as $\hat{\nu}_e(g)$ but with the new data instead. Furthermore, we will also define $m' = m'(\mathbf{x}', \mathbf{y}')$ as we defined m, but again using data $(\mathbf{x}', \mathbf{y}')$ instead of $(\mathbf{x}, \mathbf{y})$. We get:

$$\begin{aligned}
\textcircled{1} &\leq \sup_{g\in\mathcal{G}} \sup_{\substack{f_e\in\mathcal{C}_m,\\ \sum_e f_e = 0}} \mathbb{E}_{(\mathbf{x}',\mathbf{y}')} \left[\sum_{e=1}^{E} \int S_{w_e} f_e d(\hat{\nu}_e(g) - \hat{\nu}'_e(g)) \right] \\
&\leq \mathbb{E}_{(\mathbf{x}',\mathbf{y}')} \left[\sup_{g\in\mathcal{G}} \sup_{\substack{f_e\in\mathcal{C}_m,\\ \sum_e f_e = 0}} \sum_{e=1}^{E} \int S_{w_e} f_e d(\hat{\nu}_e(g) - \hat{\nu}'_e(g)) \right] \\
&\leq \mathbb{E}_{(\mathbf{x}',\mathbf{y}')} \left[\sup_{g\in\mathcal{G}} \sup_{f_e\in\mathcal{C}_{m\vee m'}} \sum_{e=1}^{E} \int S_{w_e} f_e d(\hat{\nu}_e(g) - \hat{\nu}'_e(g)) \right]. \qquad (18)
\end{aligned}$$

Since for any $g \in \mathcal{G}$ both $\hat{\nu}_e(g)$ and $\hat{\nu}'_e(g)$ are supported on $B(0, m \vee m')$, by Lemma 3, $S_{w_e} f_e$ is $4w_e(m \vee m')$-Lipschitz. Furthermore, notice that for any constant c, $\int c d(\hat{\nu}_e(g)) - \hat{\nu}'_e(g)) = 0$; so we can always modify the f_e functions up by an additive constant to get $S_{w_e} f_e(0) = 0$ without changing the value of the expression inside the supremum. Finally, we can obviously switch the two 'sup'. Based on all these remarks, we obtain the following

new bound:

$$\begin{aligned}
① &\leq 4\mathbb{E}_{(\mathbf{x}',\mathbf{y}')}\left[\sup_{\phi_e\in\mathcal{L}^0_{m\vee m'}}\sup_{g\in\mathcal{G}}\sum_{e=1}^{E} w_e\int\phi_e d(\hat{\nu}_e(g)-\hat{\nu}'_e(g))\right]\\
&\leq 4\mathbb{E}_{(\mathbf{x}',\mathbf{y}')}\left[\sum_{e=1}^{E} w_e\sup_{\phi_e\in\mathcal{L}^0_{m\vee m'}}\sup_{g\in\mathcal{G}}\int\phi_e d(\hat{\nu}_e(g)-\hat{\nu}'_e(g))\right],
\end{aligned}$$

where we define $\mathcal{L}^0_{m\vee m'}$ as the set of all $m\vee m'$-Lipschitz functions defined on $B(0,m\vee m')$ that are equal to zero at the origin, and that we extend outside of $B(0,m\vee m')$ the same way we did for functions in $\mathcal{C}_{m\vee m'}$:

$$\mathcal{L}^0_{m\vee m'} \doteq \{\phi : \phi(0)=0, \phi \text{ is } (m\vee m')\text{-Lipschitz and } \forall x\notin B(0,m\vee m'), \phi(x)=\phi((m\vee m')x/|x|)\}.$$

Now we simplify the upper bound by expending it into a sum of expectations that depend only on the observations coming from one of the environments. Beside improving clarity, since we have to treat each of these expectations separately, another important reason for this step is to allow us derive the bound (6) under the (weaker) assumption that the observations are only independent within each environment, and not necessarily across environments (see discussion at the beginning of Section D).

In what follows, let $M>0$ be a constant to be chosen later on. We have:

$$\begin{aligned}
① &\leq 4\mathbb{E}_{(\mathbf{x}',\mathbf{y}')}\left[\mathbb{1}_{m\vee m'\leq M}\sum_{e=1}^{E} w_e\sup_{\phi_e\in\mathcal{L}^0_M}\sup_{g\in\mathcal{G}}\int\phi_e d(\hat{\nu}_e(g)-\hat{\nu}'_e(g))\right]+8\mathbb{E}_{(\mathbf{x}',\mathbf{y}')}\left[(m\vee m')^2\mathbb{1}_{m\vee m'>M}\right]\\
&\leq 4\sum_{e=1}^{E} w_e\mathbb{1}_{m_e\leq M}\mathbb{E}_{(\mathbf{x}'_e,\mathbf{y}'_e)}\underbrace{\left[\sup_{\phi_e\in\mathcal{L}^0_M}\sup_{g\in\mathcal{G}}\int\phi_e d(\hat{\nu}_e(g)-\hat{\nu}'_e(g))\right]}_{\doteq G_M((\mathbf{x}_e,\mathbf{y}_e),(\mathbf{x}'_e,\mathbf{y}'_e))}+8\mathbb{E}_{m'}\underbrace{\left[(m\vee m')^2\mathbb{1}_{m\vee m'>M}\right]}_{\doteq H_M(m,m')},
\end{aligned}\tag{19}$$

where we used the fact that for any $\phi\in\mathcal{L}^0_{m\vee m'}$ we have that $|\phi|\leq M^2$ and for all $e\in[E]$ let $m_e\doteq\max_{i\in[n_e]}\sup_{g\in\mathcal{G}}|g(x_i^e)-y_i^e|$. Now we turn to finding an upper bound for:

$$② \doteq \sup_{g\in\mathcal{G}}\left(\mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g))-\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(g))\right).$$

We define $(\mathbf{x}',\mathbf{y}')$, $\hat{\nu}'_e(g)$ and m' the same way as we did above. We then get, for any fixed $g\in\mathcal{G}$:

$$\mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g))=\sup_{\substack{f_e\in\mathcal{C}_{b,2},\\ \sum_e f_e=0}}\mathbb{E}_{(\mathbf{x}',\mathbf{y}')}\left[\sum_{e=1}^{E}\int S_{w_e}f_e d\hat{\nu}'_e(g))\right]\leq\mathbb{E}_{(\mathbf{x}',\mathbf{y}')}\left[\sup_{\substack{f_e\in\mathcal{C}_{b,2},\\ \sum_e f_e=0}}\sum_{e=1}^{E}\int S_{w_e}f_e d\hat{\nu}'_e(g))\right],$$

which in turn means that:

$$\begin{aligned}
② &\leq \mathbb{E}_{(\mathbf{x}',\mathbf{y}')}\left[\sup_{g\in\mathcal{G}}\left(\sup_{\substack{f_e\in\mathcal{C}_{b,2},\\ \sum_e f_e=0}}\sum_{e=1}^{E}\int S_{w_e}f_e d\hat{\nu}'_e(g))-\sup_{\substack{f'_e\in\mathcal{C}_{b,2},\\ \sum_e f'_e=0}}\sum_{e=1}^{E}\int S_{w_e}f'_e d\hat{\nu}_e(g))\right)\right]\\
&\leq \mathbb{E}_{(\mathbf{x}',\mathbf{y}')}\left[\sup_{g\in\mathcal{G}}\sup_{f_e\in\mathcal{C}_{m\vee m'}}\sum_{e=1}^{E}\int S_{w_e}f_e d(\hat{\nu}'_e(g)-\hat{\nu}_e(g))\right],
\end{aligned}$$

where we used Lemma 4 in the last inequality. Therefore we fall back to the bound in Equation (18), where $\hat{\nu}_e(g)$ and $\hat{\nu}'_e(g)$ are switched. Following the same steps as before, we arrive at the following bound:

$$② \leq 4\sum_{e=1}^{E} w_e\mathbb{1}_{m_e\leq M}\mathbb{E}_{(\mathbf{x}'_e,\mathbf{y}'_e)}\left[G_M\left((\mathbf{x}'_e,\mathbf{y}'_e),(\mathbf{x}_e,\mathbf{y}_e)\right)\right]+8\mathbb{E}_{m'}\left[H_M\left(m',m\right)\right].\tag{20}$$

As it is usually the case in the derivation of high probability bounds, we first bound the expectations of ① and ② (in Section D.3) and then prove some concentration bounds of ① and ② around their respective averages (in Section D.4).

D.3 Bounding the Expectations

Notice that since $(\mathbf{x}, \mathbf{y})$ and $(\mathbf{x}', \mathbf{y}')$ have the same distribution (and are independent of each other), the expectations of the bounds (19) and (20) are identical, that is, we have:

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})}\left[\textcircled{1}\right] \vee \mathbb{E}_{(\mathbf{x},\mathbf{y})}\left[\textcircled{2}\right] \leq 4\sum_{e=1}^{E} w_e \mathbb{E}_{(\mathbf{x}_e,\mathbf{y}_e),(\mathbf{x}'_e,\mathbf{y}'_e)}\left[G_M\left((\mathbf{x}_e,\mathbf{y}_e),(\mathbf{x}'_e,\mathbf{y}'_e)\right)\right] + 8\mathbb{E}_{m,m'}\left[H_M\left(m,m'\right)\right] \doteq \textcircled{3}.$$

So there are two quantities to bound for any $e \in [E]$: $\mathbb{E}_{(\mathbf{x}_e,\mathbf{y}_e),(\mathbf{x}'_e,\mathbf{y}'_e)}\left[G_M\left((\mathbf{x}_e,\mathbf{y}_e),(\mathbf{x}'_e,\mathbf{y}'_e)\right)\right]$ and $\mathbb{E}_{m,m'}\left[H_M\left(m,m'\right)\right]$. Let's start with the first one. Set $e \in [E]$, we have:

$$\mathbb{E}_{(\mathbf{x}_e,\mathbf{y}_e),(\mathbf{x}'_e,\mathbf{y}'_e)}\left[G_M\left((\mathbf{x}_e,\mathbf{y}_e),(\mathbf{x}'_e,\mathbf{y}'_e)\right)\right] = \mathbb{E}_{(\mathbf{x}_e,\mathbf{y}_e),(\mathbf{x}'_e,\mathbf{y}'_e)}\left[\sup_{\phi\in\mathcal{L}_M^0, g\in\mathcal{G}} \frac{1}{n_e}\sum_{i=1}^{n_e}\left(\phi(y_i^e - g(x_i^e)) - \phi(y_i'^e - g(x_i'^e))\right)\right] \doteq \textcircled{4}$$

As all the observations are i.i.d., the above expectation would remain unchanged if we switched any observation (x_i^e, y_i^e) with its counterpart $(x_i'^e, y_i'^e)$. Let $\boldsymbol{\xi} = (\xi_i)_{i=1}^{n_e}$ be i.i.d. Rademacher variables. By symmetry we thus have:

$$\begin{aligned}
\textcircled{4} &= \mathbb{E}_{(\mathbf{x}_e,\mathbf{y}_e),(\mathbf{x}'_e,\mathbf{y}'_e)}\left[\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\phi\in\mathcal{L}_M^0, g\in\mathcal{G}} \frac{1}{n_e}\sum_{i=1}^{n_e}\xi_i\left(\phi(y_i^e - g(x_i^e)) - \phi(y_i'^e - g(x_i'^e))\right)\right]\right] \\
&\leq 2\mathbb{E}_{(\mathbf{x}_e,\mathbf{y}_e)}\left[\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\phi\in\mathcal{L}_M^0}\sup_{g\in\mathcal{G}} \frac{1}{n_e}\sum_{i=1}^{n_e}\xi_i\phi(y_i^e - g(x_i^e))\right]\right]. \qquad (21)
\end{aligned}$$

The above quantity is close to the Rademacher complexity of the function class $\mathcal{G}$; if ϕ were a fixed M-Lipschitz function, then by the contraction lemma (e.g., Lemma 26.9 from Shalev-Shwartz and Ben-David (2014)) we would be able to bound it directly by $2M\mathcal{R}_{n_e}(\mathcal{G})$. However, the supremum over $\mathcal{L}_M^0$ inside the expectation makes it more difficult to derive such a bound involving the Rademacher complexity. We show below that this additional supremum will only add a $\log(n_e)$ factor and some $O(n_e^{-1/2})$ terms to the Rademacher complexity. We prove this using a chaining argument (Dudley, 1987); our next steps are inspired by the proof of Lemma 27.4 from Shalev-Shwartz and Ben-David (2014). Any $\phi \in \mathcal{L}_M^0$ can be decomposed as follows:

$$\phi = (\phi - \phi_K) + (\phi_K - \phi_{K-1}) + \ldots + (\phi_1 - \phi_0) + \phi_0, \qquad (22)$$

where for any $k \in \{0\} \cup [K]$, ϕ_k belongs to a $2^{-k}M^2$-cover (for the norm $\|\cdot\|_\infty$) of $\mathcal{L}_M^0$, and chosen such that: $\|\phi_k - \phi_{k+1}\|_\infty \leq 2^{-k}M^2$, where $\phi_{M+1} = \phi$. As any function in $\mathcal{L}_M^0$ is bounded by M^2, we can take $\phi_0 = 0$. Call B_k the above cover sets of $\mathcal{L}_M^0$ such that $\forall k, \phi_k \in B_k$, and for any $k \geq 1$ let $\hat{B}_k \doteq \{\phi_k - \phi_{k-1} : \phi_k \in B_k, \phi_{k-1} \in B_{k-1} \text{ and } \|\phi_k - \phi_{k-1}\|_\infty \leq 2^{-k+1}M^2\}$. Then from Equation (22) we have that for any fixed $\phi \in \mathcal{L}_M^0$,

$$\exists \psi_k \in \hat{B}_k \text{ for } k \in [K], \text{ and } \exists \psi \in \mathcal{L}_{2M}^0, \text{ s.t. } \|\psi\|_\infty \leq M^2 2^{-K} \text{ and } \phi = \psi + \sum_{k=1}^{K}\psi_k.$$

Therefore, for any fixed sample $(\mathbf{x}_e, \mathbf{y}_e)$ we get:

$$\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\phi\in\mathcal{L}_M^0}\sup_{g\in\mathcal{G}} \frac{1}{n_e}\sum_{i=1}^{n_e}\xi_i\phi(y_i^e - g(x_i^e))\right] \leq M^2 2^{-K} + \sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\psi\in\hat{B}_k}\sup_{g\in\mathcal{G}} \frac{1}{n_e}\sum_{i=1}^{n_e}\xi_i\psi(y_i^e - g(x_i^e))\right]. \qquad (23)$$

The advantage of this decomposition is that each set $\hat{B}_k$ is finite. Indeed, it is known that the metric entropy (the logarithm of the covering number) for an ϵ-cover of the class of L-Lipschitz functions defined on a ball of diameter D in $\mathbb{R}^d$ is of the order of $(LD/\epsilon)^d$. As we focus on Lipschitz functions on $\mathbb{R}$ here, we have that $|\hat{B}_k|$ is therefore of the order 2^k. More precisely, in Lemma 5 we (briefly) expose a possible construction for B_k, for which we have $\log|B_k| = 2\log(3)\cdot 2^k$. Furthermore, from the construction given in Lemma 5, the ϕ_k's from (22) can actually be chosen so that $\psi_k \doteq \phi_k - \phi_{k-1}$ is M-Lipschitz (instead of $2M$-Lipschitz) and $\|\psi_k\|_\infty \leq M^2 2^{-k}$ (instead of $M^2 2^{-k+1}$). In the following we will therefore consider that the ϕ_ks and the ψ_ks satisfy these conditions; note however that this change will just improve our bound up to some constant factor, so it can be ignored.

We can derive more explicit upper bounds for the terms in R.H.S. of Equation (23), using an adaptation of Massart Lemma's proof (see Lemma 6). More precisely, we show in Lemma 6 how to bound

$\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\psi\in B}\sup_{g\in\mathcal{G}}\frac{1}{n_e}\sum_{i=1}^{n_e}\xi_i\psi(y_i^e-g(x_i^e))\right]$ when B is a finite set of bounded and Lipschitz functions. Therefore, using Lemma 6 with $\epsilon = M^2 2^{-k}$ and $\log|\hat{B}_k| \leq \log|B_k|^2 = 4\log(3)\cdot 2^k$, we get:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\phi\in\mathcal{L}_M^0}\sup_{g\in\mathcal{G}}\frac{1}{n_e}\sum_{i=1}^{n_e}\xi_i\phi(y_i^e-g(x_i^e))\right] &\leq M^2 2^{-K} + M^2\sum_{k=1}^{K}2\sqrt{2}2^{-k}\sqrt{\frac{4\log(3)2^k}{n_e}} + KMR_{\mathbf{x}_e}(\mathcal{G}) \\ &\leq M^2 2^{-K} + KMR_{\mathbf{x}_e}(\mathcal{G}) + M^2\frac{4\sqrt{2\log(3)}}{\sqrt{n_e}\left(\sqrt{2}-1\right)},\end{aligned}$$

where we used in the last inequality that $\sum_{k=1}^{\infty}2^{-k/2} = \frac{1}{\sqrt{2}-1}$.

Finally, by setting $K = \lceil\log(n_e)\rceil$, and since $\log 2 \geq 0.5$, we get that:

$$\mathbb{E}_{(\mathbf{x}_e,\mathbf{y}_e),(\mathbf{x}'_e,\mathbf{y}'_e)}\left[G_M\left((\mathbf{x}_e,\mathbf{y}_e),(\mathbf{x}'_e,\mathbf{y}'_e)\right)\right] \leq \frac{M^2}{\sqrt{n_e}}\cdot\left(2+\frac{8\sqrt{2\log(3)}}{\sqrt{2}-1}\right) + 2\left(1+\log(n_e)\right)M\mathcal{R}_{n_e}(\mathcal{G}). \tag{24}$$

Now let's turn to bounding $\mathbb{E}_{m,m'}[H_M(m,m')] = \mathbb{E}_{m,m'}[(m\vee m')^2\mathbb{1}_{m\vee m'>M}]$. Using the notation from Theorem 1, we let $Z_e \doteq \sup_{g\in\mathcal{G}}|g(X^e)-Y^e|$; recall that from Theorem 1 we assume these variables are either bounded or sub-Gaussian (e.g., see Wainwright (2019), Chapter 2 for the definition). In particular, if the Z_es are all bounded by M, then $\mathbb{E}_{m,m'}[H_M(m,m')]$ is equal to zero.

Assume now that the Z_es are sub-Gaussian $\mathcal{G}(\mu_e,\sigma_e)$, that is with mean μ_e and sub-Gaussian parameter σ_e. For simplicity call $Z \doteq m\vee m$; if we set $M = \sqrt{2}(M'+\mu)$ for some $M'>0$ and where $\mu \doteq \max_e\mu_e$, we prove in Lemma 7 that:

$$\mathbb{E}_Z[Z^2\mathbb{1}_{Z>M}] \leq 2\sum_{e=1}^{E}n_e\left(2(M'+\mu)^2e^{-M'^2/\sigma_e^2} + 4\sigma_e^2e^{-M'^2/2\sigma_e^2}\right).$$

Call $\sigma^2 = \max_{e\in[E]}\sigma_e^2$, and let $\delta\in(0,1)$ be the probability from Theorem 1 – we introduce it now but it will be useful in Section D.4. We set $M' = 2\sqrt{\log(n/\delta)}\sigma$ (recall that $n = \sum_e n_e$). Hence, for $M = \sqrt{2}\left(2\sqrt{\log(n/\delta)}\sigma+\mu\right)$:

$$\begin{aligned}\mathbb{E}_{m,m'}[H_M(m,m')] &\leq 2\sum_{e=1}^{E}\frac{\delta^2n_e}{n^2}\cdot\left(2\left(2\sqrt{\log(n/\delta)}+\mu\right)^2\delta^2/n^2 + 4\sigma^2\right) \\ &\leq \frac{4\delta^2}{n}\left(\left(2\sqrt{\log(n/\delta)}+\mu\right)^2\delta^2/n^2 + 2\sigma^2\right).\end{aligned} \tag{25}$$

To conclude this section, let's summarize the bounds we derived based on Equations (24) and (25). When the Z_es are sub-Gaussian $\mathcal{G}(\mu_e,\sigma_e)$, we have:

$$③ \leq \sum_{e=1}^{E}w_e\left(\frac{A^0_{\delta,n}}{\sqrt{n_e}} + B^0_{\delta,n}(1+\log(n_e))\mathcal{R}_{n_e}(\mathcal{G})\right) + \frac{C^0_{\delta,n}}{n}, \tag{26}$$

where $A^0_{\delta,n} = 8\left(2\sqrt{\log(n/\delta)}\sigma+\mu\right)^2\cdot\left(2+\frac{8\sqrt{2\log 3}}{\sqrt{2}-1}\right)$, $B^0_{\delta,n} = 8\sqrt{2}\left(2\sqrt{\log(n/\delta)}\sigma+\mu\right)$ and $C^0_{\delta,n} = 32\delta^2\left((2\sqrt{\log(n/\delta)}\sigma+\mu)^2\delta^2/n^2+2\sigma^2\right)$.

When the Z_e's are bounded with probability one by some constant $M>0$, we have:

$$③ \leq \sum_{e=1}^{E}w_e\left(\frac{A^0}{\sqrt{n_e}} + B^0(1+\log(n_e))\mathcal{R}_{n_e}(\mathcal{G})\right), \tag{27}$$

where $A^0 = 4\left(2+\frac{8\sqrt{2\log(3)}}{\sqrt{2}-1}\right)M^2$ and $B^0 = 8(1+\log(n_e))M$.

D.4 Concentration Bounds

Take any $e \in [E]$ and $M > 0$. Notice that $\mathbb{1}_{m_e \leq M}\mathbb{E}_{(\mathbf{x}'_e,\mathbf{y}'_e)}\left[G_M\left((\mathbf{x}_e,\mathbf{y}_e),(\mathbf{x}'_e,\mathbf{y}'_e)\right)\right]$, as a function of the data $(\mathbf{x}_e,\mathbf{y}_e)$, satisfies the bounded difference condition for McDiarmid's inequality (e.g., see Lemma 26.4 from Shalev-Shwartz and Ben-David (2014)) with constant $2M^2/n_e$. Hence with probability at least $1-\delta/E$, we have:

$$\mathbb{1}_{m_e \leq M}\mathbb{E}_{(\mathbf{x}'_e,\mathbf{y}'_e)}\left[G_M\left((\mathbf{x}_e,\mathbf{y}_e),(\mathbf{x}'_e,\mathbf{y}'_e)\right)\right] \leq M^2\sqrt{\frac{2\log(E/\delta)}{n_e}} + \mathbb{E}_{(\mathbf{x}_e,\mathbf{y}_e),(\mathbf{x}'_e,\mathbf{y}'_e)}\left[G_M\left((\mathbf{x}_e,\mathbf{y}_e),(\mathbf{x}'_e,\mathbf{y}'_e)\right)\right].$$

Similarly, we have also with probability at least $1-\delta/E$:

$$\mathbb{1}_{m_e \leq M}\mathbb{E}_{(\mathbf{x}'_e,\mathbf{y}'_e)}\left[G_M\left((\mathbf{x}'_e,\mathbf{y}'_e),(\mathbf{x}_e,\mathbf{y}_e)\right)\right] \leq M^2\sqrt{\frac{2\log(E/\delta)}{n_e}} + \mathbb{E}_{(\mathbf{x}_e,\mathbf{y}_e),(\mathbf{x}'_e,\mathbf{y}'_e)}\left[G_M\left((\mathbf{x}'_e,\mathbf{y}'_e),(\mathbf{x}_e,\mathbf{y}_e)\right)\right].$$

Finally, note that we have also

$$\begin{aligned}\mathbb{E}_{m'}\left[H_M\left(m,m'\right)\right] = \mathbb{E}_{m'}\left[H_M\left(m',m\right)\right] &\leq \mathbb{E}_{m'}\left[m'^2\mathbb{1}_{m'>M}\right] + m^2\mathbb{1}_{m>M} \\ &\leq \mathbb{E}_{m,m'}\left[H_M\left(m,m'\right)\right] + m^2\mathbb{1}_{m>M}.\end{aligned}$$

When the Z_e's are all bounded by M with probability one, then $m^2\mathbb{1}_{m>M} = 0$ (with probability one). However, when the Z_e's are only sub-Gaussian $\mathcal{G}(\mu_e,\sigma_e)$, the probability that $m^2\mathbb{1}_{m>M}$ is non-zero is

$$\mathbb{P}\left(m^2\mathbb{1}_{m>M} = 0 > 0\right) = \mathbb{P}\left(m > M\right) \leq \sum_{e=1}^{E} n_e\mathbb{P}\left(Z_e \geq M\right) \leq ne^{-M'^2/\sigma^2} \leq \delta,$$

when we choose $M' = 2\sqrt{\log(n/\delta)}\sigma$ and $M = \sqrt{2}(M'+\mu)$. Hence, combining the above inequalities and equations (19) and (20) we obtain that with probability at least $1-3\delta$ simultaneously:

$$\textcircled{1} \leq \textcircled{3} + 2\left(2\sqrt{\log(n/\delta)}\sigma + \mu\right)^2\sqrt{\frac{2\log(E/\delta)}{n_e}} \text{ and } \textcircled{2} \leq \textcircled{3} + 2\left(2\sqrt{\log(n/\delta)}\sigma + \mu\right)^2\sqrt{\frac{2\log(E/\delta)}{n_e}}, \quad (28)$$

when the Z_e's are sub-Gaussian. When they are just bounded by some $M > 0$, we have with probability at least $1-2\delta$ that simultaneously:

$$\textcircled{1} \leq \textcircled{3} + M^2\sqrt{\frac{2\log(E/\delta)}{n_e}} \quad \text{and} \quad \textcircled{2} \leq \textcircled{3} + M^2\sqrt{\frac{2\log(E/\delta)}{n_e}}. \quad (29)$$

D.5 Conclusion

We now combine the bounds that were obtained in the previous sections, in particular equations (19), (20), (26) Whenever the Z_e's are sub-Gaussian $\mathcal{G}(\mu_e,\sigma_e)$, we have that with probability at least $1-3\delta$:

$$\sup_{g\in\mathcal{G}}\left|\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(g)) - \mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g))\right| = \textcircled{1}\vee\textcircled{2} \leq \sum_{e=1}^{E} w_e\left(\frac{A_{\delta,n}}{\sqrt{n_e}} + B_{\delta,n}(1+\log(n_e))\mathcal{R}_{n_e}(\mathcal{G})\right) + \frac{C_{\delta,n}}{n},$$

where $A_{\delta,n} = A^0_{\delta,n} + 2\left(2\sqrt{\log(n/\delta)}\sigma+\mu\right)^2\cdot\sqrt{2\log(E/\delta)}$, $B_{\delta,n} = B^0_{\delta,n}$ and $C_{\delta,n} = C^0_{\delta,n}$.

Whenever the Z_e's are bounded with probability one by some constant $M > 0$, we have with probability at least $1-2\delta$:

$$\sup_{g\in\mathcal{G}}\left|\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(g)) - \mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g))\right| = \textcircled{1}\vee\textcircled{2} \leq \sum_{e=1}^{E} w_e\left(\frac{A_{\delta}}{\sqrt{n_e}} + B_{\delta}(1+\log(n_e))\mathcal{R}_{n_e}(\mathcal{G})\right),$$

where $A_\delta = A^0_\delta + M^2\cdot\sqrt{2\log(E/\delta)}$ and $B_\delta = B^0_\delta$.

Therefore the high probability uniform bound from Theorem 1 is obtained by replacing δ respectively by $\delta/3$ and $\delta/2$. What is left to show is the bound for the difference between the minimal Wasserstein variance and its empirical counterpart; this can be shown since:

$$\inf_{g\in\mathcal{G}}\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(g)) - \inf_{g\in\mathcal{G}}\mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g)) \leq \textcircled{1} \quad \text{and} \quad \inf_{g\in\mathcal{G}}\mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g)) - \inf_{g\in\mathcal{G}}\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(g)) \leq \textcircled{2}.$$

D.6 Supporting Lemmas

Recall that in Section D.2 we defined $\mathcal{C}_m$ as the set of continuous functions defined on $\mathbb{R}$ that are constant on $(-\infty, -m]$ and on $[m, +\infty)$, that is $\mathcal{C}_m \doteq \{f \in C(\mathbb{R}) : \forall x \in B(0,m)^c, f(x) = f(mx/|x|)\}$.

Lemma 3. *Let $w > 0$, $m > 0$ and $f \in \mathcal{C}_m$. Then for any $x \in B(0,m)$ we have:*

$$S_w f(x) = \inf_{y \in B(0,m)} \left\{ w|x-y|^2 - f(y) \right\},$$

and therefore $S_w f(x)$ is $4wm$-Lipschitz on $B(0,m)$.

Proof. Let $x \in B(0,m)$. Recall that we have:

$$S_w f(x) \doteq \inf_{y \in \mathbb{R}} \left\{ w|x-y|^2 - f(y) \right\}.$$

Take any $y \in B(0,m)^c$. As $f \in \mathcal{C}_m$ then $f(y) = f(my/|y|)$. Note that $my/|y|$ is the orthogonal projection of y on $B(0,m)$, hence by the contraction property of orthogonal projections we also have that $|x - my/|y||^2 \leq |x-y|^2$. This concludes the first point.

The second point is implied by the fact that, for all $y \in B(0,m)$, the functions $x \mapsto w|x-y|^2 - f(y)$ are $4wm$-Lipschitz on $B(0,m)$ (see for instance Box 1.8 from Santambrogio (2015)). □

Lemma 4. *For any probability distributions $\nu_1, \nu_2, \ldots, \nu_E$ that are all supported on the ball $B(0,m)$ for some $m > 0$, we have :*

$$WV_{\boldsymbol{w}}(\boldsymbol{\nu}) = \sup \left\{ \sum_{e=1}^{E} \int S_{w_e} f_e d\nu_e : f_e \in \mathcal{C}_m, \sum_{e=1}^{E} f_e = 0 \right\}. \tag{30}$$

Proof. As all function functions $f_e \in \mathcal{C}_m$ are continuous and bounded, they belong to $\mathcal{C}_{b,2}$, and it is clear that:

$$\mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}) \geq \sup \left\{ \sum_{e=1}^{E} \int S_{w_e} f_e d\nu_e : f_e \in \mathcal{C}_m, \sum_{e=1}^{E} f_e = 0 \right\}.$$

The inequality in the opposite direction is proved by modifying a dual solution for Equation (16) into a solution for Equation (30). Note that Proposition 2.3 from Agueh and Carlier (2011) proves that indeed the dual problem in Equation (16) admits a solution $(f_e^*)_{e=1}^E$. If we define a new operator $S_{w_e}^m$ on $\mathcal{C}_{b,2}$ as follows:

$$S_{w_e}^m f(x) \doteq \inf_{y \in B(0,m)} \left\{ w_e|x-y|^2 - f(y) \right\} \geq S_{w_e} f(x).$$

We then have that:

$$\mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}) = \sum_{e=1}^{E} \int S_{w_e} f_e^* d\nu_e \leq \sum_{e=1}^{E} \int S_{w_e}^m f_e^* d\nu_e.$$

For any $e \in [E]$, consider now another function $\tilde{f}_e \in \mathcal{C}_m$ defined as follows:

$$\tilde{f}_e(x) = \begin{cases} f_e^*(x) \text{ if } x \in B(0,m), \\ f_e^*(mx/|x|) \text{ otherwise.} \end{cases}$$

We may check that $\sum_{e=1}^{E} \tilde{f}_e = 0$, and that for any $x \in B(0,m)$ we have $S_{w_e}^m f_e^*(x) = S_{w_e}^m \tilde{f}_e(x) = S_{w_e} \tilde{f}_e(x)$, where the last equality comes from Lemma 3. As the distributions ν_e are supported on $B(0,m)$:

$$\mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}) \leq \sum_{e=1}^{E} \int S_{w_e} \tilde{f}_e d\nu_e,$$

which proves equality (30). □

Let $M > 0$. Recall the definition of $\mathcal{L}_M^0$ from Section D.2:

$$\mathcal{L}_M^0 \doteq \{\phi : \phi(0) = 0, \phi \text{ is } M\text{-Lipschitz and } \forall x \notin B(0,M), \phi(x) = \phi((M)x/|x|)\}.$$

Lemma 5. *For $M > 0$ and $k \geq 1$, it is possible to construct an $M^2 2^{-k}$-cover, called B_k, of $\mathcal{L}_M^0$ such that $\log |B_k| = 2\log(3) \cdot 2^k$.*

Proof. A straightforward construction of B_k can be done as follows: subdivide the interval $[-M, M]$ into a grid, each segment of length $M2^{-k}$ (there are 2^k of them on each side of the origin); set B_k to be composed of all (continuous) piece-wise linear functions equal to 0 at the origin and either increase or decrease by $M^2 2^{-k}$ or stay constant to the next point in the grid. These functions are also set to stay constant outside of $[-M, M]$. It is easy from there to check that this construction is indeed a $M^2 2^{-k}$-cover of $\mathcal{L}_M^0$, and that $\log |B_k| = 2\log(3) \cdot 2^k$. □

Lemma 6. *Let $M > 0$ and B a finite set of M-Lipschitz functions, that are bounded by some $\epsilon > 0$ for the infinite norm. For a fixed sample $(\boldsymbol{x}_e, \boldsymbol{y}_e)$, we have:*

$$\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\psi \in B}\sup_{g \in \mathcal{G}} \frac{1}{n_e}\sum_{i=1}^{n_e} \xi_i \psi(y_i^e - g(x_i^e))\right] \leq M R_{\boldsymbol{x}_e}(\mathcal{G}) + 2\sqrt{2}\epsilon\sqrt{\frac{\log |B|}{n_e}}.$$

Proof. Our proof follows similar steps than in the proof of the Massart Lemma; see Lemma 26.8 from Shalev-Shwartz and Ben-David (2014). Let $\lambda > 0$, we have:

$$\begin{aligned}
\lambda \mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\psi \in B}\sup_{g \in \mathcal{G}} \frac{1}{n_e}\sum_{i=1}^{n_e} \xi_i \psi(y_i^e - g(x_i^e))\right] &= \mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(\sup_{\psi \in B}\exp\left(\sup_{g \in \mathcal{G}} \frac{\lambda}{n_e}\sum_{i=1}^{n_e} \xi_i \psi(y_i^e - g(x_i^e))\right)\right)\right] \\
&\leq \mathbb{E}_{\boldsymbol{\xi}}\left[\log\left(\sum_{\psi \in B}\exp\left(\sup_{g \in \mathcal{G}} \frac{\lambda}{n_e}\sum_{i=1}^{n_e} \xi_i \psi(y_i^e - g(x_i^e))\right)\right)\right] \\
&\leq \log \sum_{\psi \in B} \mathbb{E}_{\boldsymbol{\xi}}\left[\exp\left(\lambda \sup_{g \in \mathcal{G}} \frac{1}{n_e}\sum_{i=1}^{n_e} \xi_i \psi(y_i^e - g(x_i^e))\right)\right] \doteq ④.
\end{aligned}$$

Where the last inequality comes from Jensen's inequality. By Azuma-Hoeffding's Theorem (see for instance Chapter 2 from Wainwright (2019)) the variable $\sup_{g \in \mathcal{G}} n_e^{-1} \sum_{i=1}^{n_e} \xi_i \psi(y_i^e - g(x_i^e))$ is sub-Gaussian with mean $R'(\psi) \doteq \mathbb{E}_{\boldsymbol{\xi}}[\sup_{g \in \mathcal{G}} n_e^{-1} \sum_{i=1}^{n_e} \xi_i \psi(y_i^e - g(x_i^e))]$ and parameter $\sigma^2 \doteq n_e^{-1} \sum_{i=1}^{n_e} 4\epsilon^2 = 4\epsilon^2/n_e$. Notice that by the Contraction Lemma (i.e. Lemma 26.9 from Shalev-Shwartz and Ben-David (2014)) we have $R'(\psi) \leq R_{\mathbf{x}_e}(\mathcal{G})M$. Hence we get this bound:

$$\begin{aligned}
④ &\leq \log \sum_{\psi \in B} \mathbb{E}_{\boldsymbol{\xi}}\left[\exp\left(\lambda\left(\sup_{g \in \mathcal{G}} \frac{1}{n_e}\sum_{i=1}^{n_e} \xi_i \psi(y_i^e - g(x_i^e)) - R'(\psi)\right) + \lambda R'(\psi)\right)\right] \\
&\leq \log\left(|B| e^{\lambda M R_{\mathbf{x}_e}(\mathcal{G})} \cdot e^{2\lambda^2\epsilon^2/n_e}\right) = \log|B| + \lambda M R_{\mathbf{x}_e}(\mathcal{G}) + 2\lambda^2\epsilon^2/n_e.
\end{aligned}$$

Hence,

$$\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\psi \in B}\sup_{g \in \mathcal{G}} \frac{1}{n_e}\sum_{i=1}^{n_e} \xi_i \psi(y_i^e - g(x_i^e))\right] \leq \log|B|/\lambda + M R_{\mathbf{x}_e}(\mathcal{G}) + 2\lambda\epsilon^2/n_e.$$

Taking $\lambda = \sqrt{\frac{n_e \log|B|}{2\epsilon^2}}$ we get the result. □

As in Section D.2 we let $m = m(\mathbf{x}, \mathbf{y}) \doteq \max_{e \in [E]} \max_{i \in [n_e]} \sup_{g \in \mathcal{G}} |g(x_i^e) - y_i^e|$. Also as in Section D.2, we let $m' = m'(\mathbf{x}', \mathbf{y}')$ to be similarly defined but for another (independent) data set $(\mathbf{x}', \mathbf{y}')$ drawn from the same distribution as $(\mathbf{x}, \mathbf{y})$.

Lemma 7. *Let $M > 0$ be a constant and call $Z \doteq m \vee m'$. Assume $Z_e \doteq \sup_{g \in \mathcal{G}} |g(X^e) - Y^e|$ is sub-Gaussian $\mathcal{G}(\mu_e, \sigma_e)$, for any $e \in [E]$. Finally, call $\mu \doteq \max_e \mu_e$. We have:*

$$\mathbb{E}_Z[Z^2 \mathbb{1}_{Z > M}] \leq 2\sum_{e=1}^{E} n_e \left(2(M' + \mu)^2 e^{-M'^2/\sigma_e^2} + 4\sigma_e^2 e^{-M'^2/2\sigma_e^2}\right),$$

whenever $M' \doteq M/\sqrt{2} - \mu > 0$.

Proof. We have:

$$\begin{aligned}\mathbb{E}_Z[Z^2\mathbb{1}_{Z>M}] &= \int_0^\infty \mathbb{P}\left(Z^2\mathbb{1}_{Z>M} > t\right) dt \\ &= \int_0^{M^2} \mathbb{P}\left(Z > M\right) dt + \int_{M^2}^\infty \mathbb{P}\left(Z^2 > t\right) dt \\ &= M^2\mathbb{P}\left(Z > M\right) + \int_{M^2}^\infty \mathbb{P}\left(Z > \sqrt{t}\right) dt.\end{aligned}$$

Notice that by a simple union bound we have $\mathbb{P}(Z > t) \leq 2\sum_{e=1}^E n_e \mathbb{P}(Z_e > t)$. Now recall that the Z_e's are sub-Gaussian with means μ_e and sub-Gaussian parameters σ_e, and that we defined $M' > 0$ such that $M = \sqrt{2}(M' + \mu)$. Finally, note that $\sqrt{2}\sqrt{t+\mu^2} \geq \sqrt{t} + \mu$, hence for any $e \in [E]$:

$$\mathbb{P}\left(Z_e \geq \sqrt{2}\sqrt{t+\mu^2}\right) \leq \mathbb{P}\left(Z_e \geq \sqrt{t} + \mu\right) \leq e^{-t/2\sigma_e^2}.$$

Using the change of variable $t \to 2(t + \mu^2)$ we get, for all $e \in [E]$:

$$\begin{aligned}\int_{M^2}^\infty \mathbb{P}\left(Z_e \geq \sqrt{t}\right) dt &\leq 2\int_{M^2/2-\mu^2}^\infty e^{-t/2\sigma_e^2} dt \\ &\leq 2\int_{M'^2}^\infty e^{-t/2\sigma_e^2} = 4\sigma_e^2 e^{-M'^2/2\sigma_e^2}.\end{aligned}$$

Also: $\mathbb{P}\left(Z_e \geq M\right) \leq \mathbb{P}\left(Z_e - \mu \geq \sqrt{2}M'\right) \leq e^{-M'^2/\sigma_e^2}$. Combining all of the above, we get the result. □

E PROOFS OF PROPOSITION 1, Theorem 2 and Theorem 3

We provide in this section the proofs of Proposition 1, Theorem 2 and Theorem 3, as well as the details of the regularity conditions that are needed for them.

E.1 Sufficient Conditions for the Asymptotic Limit in Equation (7)

Let $(z_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \nu$, where ν is a probability distribution on $\mathbb{R}$. Define $\hat{\nu} = n^{-1}\sum_{i=1}^n \delta_{z_i}$ the corresponding empirical distribution. In the following we call F the CDF of ν, $f \doteq F'$ its PDF, F^{-1} its quantile function and $q \doteq (F^{-1})'$ the quantile density. Then, from Del Barrio et al. (2005) Theorem 4.6 we have:

$$nW_2^2(\hat{\nu}, \nu) \xrightarrow[n\to\infty]{d} \int_0^1 B^2(t)q^2(t)dt,$$

where $B(t)$ is the Brownian bridge between 0 and 1, if the following conditions are satisfied:

Assumption 3 (Case (i) from Theorem 4.6 of Del Barrio et al. (2005))**.** *Using the above notations, the distribution ν satisfies the following properties:*

(i) *ν is supported on an interval (a_F, b_F), F is twice differentiable and $f > 0$ on (a_F, b_F); note this means F^{-1} is also twice differentiable on $(0, 1)$,*

(ii) $\sup_{0<t<1} t(1-t)|q'(t)|/q(t) < \infty$,

(iii) $\int_0^1 t(1-t)q^2(t)dt < \infty$,

(iv) *either $a_F > -\infty$ or $\liminf_{t\to 0^+} t|q'(t)|/q(t) > 0$ and either $b_F > -\infty$ or $\liminf_{t\to 0^+} t|q'(1-t)|/q(1-t) > 0$,*

(v) $\lim_{t\to 0^+} tq(t) = 0$ *and* $\lim_{t\to 0^+} tq(1-t) = 0$.

Del Barrio et al. (2005) also provides other distributional limit results for the square Wasserstein variance under assumptions different from Assumption 3. For simplicity, we limit ourselves to the above case as under this setting the asymptotic distribution is relatively simple, but it might happen that some realistic distributions does not satisfy Assumption 3. It turns out however that a small modification of the Wasserstein distance can alleviate this potential issue (see Remark 1 below).

Remark 1 (When Assumption 3 holds)**.** *It is easy to see that Assumption 3 would hold, for instance, for any distribution ν that is compactly supported with a continuously differentiable density that does not converge to zero too fast at the borders of its support, or is simply bounded away from zero. Furthermore, the assumption can hold also for distributions that are not compactly supported, as long as their tails of distribution are not too heavy. For instance, in Examples 4.1 and 4.2 Del Barrio et al. (2005) shows that for the Weibull distribution, or distributions with tails of the form $\exp(-|x|^\alpha)$, Assumption 3 holds only when $\alpha > 2$.*

This means in particular that the normal distribution unfortunately does not respect Assumption 3 and in fact falls into another category of distributions (see case (ii) from Theorem 4.6 of Del Barrio et al. (2005)) for which one needs to correct $nW_2^2(\hat{\nu}, \nu)$ by a drift that goes to infinity in order to obtain convergence in distribution. This drift for the normal distribution will actually diverge relatively slowly at a logarithmic rate though, so actually even when the distribution ν has tails relatively similar to a Gaussian, and therefore does not respect Assumption 3, setting the thresholds based on the asymptotic distribution in the R.H.S. of (7) (or the R.H.S. of (8)) might not be such a detriment in practice. For instance, we used Gaussian noises in our simulations' observational environments and didn't encounter any major issue.

However, if it is believed that the residuals don't respect Assumption 3 in a way that might affect the validity of the thresholds, there is in fact a very simple way to solve the issue. Del Barrio et al. (2005) actually derived their asymptotic results for a more general version of the Wasserstein distance called weighted Wasserstein distance; it is simply defined as the weighted L2 norm of the difference of the quantile functions. We could easily replace in our paper the Wasserstein distance by its weighted counterpart and most of our results would still hold for 'well-behaved' weight functions (Theorem 1's proof would probably be the most difficult to modify – for simplicity and clarity we focus only on the classical Wasserstein distance in our paper and leave this potential extension for future work); the only minor difference would be that the integrals in the R.H.S. of (7), (8) will include the weight function. In that case, since in general Assumption 3 fails because the quantile density diverge too fast at 0 and 1, choosing a weight function that goes to zero fast enough at 0 and 1 will allow our asymptotic analysis to hold for a potentially much wider range of distributions.

A more radical choice a weight function can be one that is equal to zero outside of an interval $[\alpha, 1-\alpha]$ for $\alpha \in (0, 1/2)$ – note that in that case we would loose the metric property needed for Lemma 2 to hold. This kind of weighted Wasserstein distance is sometimes called trimmed Mallows distance in the literature (Munk and Czado, 1998), and its asymptotic properties hold under quite weaker assumptions than Assumption 3.

E.2 Proof of Proposition 1

Using the notations from Section 5, let $\hat{F}_e^{-1}$ be the empirical quantile function for $\hat{\nu}_e(f)$, and similarly call F^{-1} the quantile function of the distribution $\nu(f)$ defined in Proposition 1. For any $e \in [E]$ and $t \in [0,1]$, we also define the empirical quantile process $v_{n_e}(t) \doteq \sqrt{n_e}\left(\hat{F}_e^{-1}(t) - F^{-1}(t)\right)$. Finally, let $\boldsymbol{n} \doteq (n_e)_{e=1}^E$ and $\mathbf{1}$ is a vector of size E composed only of ones. Based on equation (11), we have:

$$
\begin{aligned}
n\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(f)) &= \int_0^1 \sum_{e=1}^E n w_e \left(\hat{F}_e^{-1}(t) - \sum_{e'=1}^E w_{e'} \hat{F}_{e'}^{-1}(t)\right)^2 dt \\
&= \int_0^1 \sum_{e=1}^E w_e \left(w_e^{-1/2} v_{n_e}(t) - \sum_{e'=1}^E \sqrt{w_{e'}} v_{n_{e'}}(t)\right)^2 dt \\
&= \int_0^1 V_{\boldsymbol{n}}^T(t) \left(D_{\boldsymbol{w}}^{-1/2} - \mathbf{1}\mathbf{1}^T D_{\boldsymbol{w}}^{1/2}\right)^T D_{\boldsymbol{w}} \left(D_{\boldsymbol{w}}^{-1/2} - \mathbf{1}\mathbf{1}^T D_{\boldsymbol{w}}^{1/2}\right) V_{\boldsymbol{n}}(t) dt \\
&= \int_0^1 V_{\boldsymbol{n}}^T(t) \left(I_E - D_{\boldsymbol{w}}^{1/2}\mathbf{1}\mathbf{1}^T D_{\boldsymbol{w}}^{1/2}\right)^2 V_{\boldsymbol{n}}(t) dt,
\end{aligned}
$$

where I_E refers to the identity matrix of size E, $D_{\boldsymbol{w}} \doteq \mathrm{diag}(\boldsymbol{w})$,$V_{\boldsymbol{n}}(t) \doteq \left((v_{n_e}(t))_{e=1}^E\right)^T$ and $V_{\boldsymbol{n}}^T(t)$ is its transpose. In the following, we also call $A_{\boldsymbol{w}} \doteq I_E - D_{\boldsymbol{w}}^{1/2}\mathbf{1}\mathbf{1}^T D_{\boldsymbol{w}}^{1/2}$.

Under Assumption 3, note that we have for any $e \in [E]$,

$$
v_{n_e}(t) \xrightarrow[n_e \to \infty]{d} q_f(t) B_e(t),
$$

in $L_2(0,1)$ for a Brownian bridge $B_e(t)$ on $[0,1]$ – this is actually a consequence of Theorem 4.1 and Lemma 2.3 from Del Barrio et al. (2005) which show that a truncated version of $v_{n_e}(t)$ converges in distribution toward $q_f(t)B_e(t)$ in $L_2(0,1)$, and of their Lemma 2.4 which shows that the difference between this truncated version and the full process $v_{n_e}(t)$ goes to 0 in probability. Furthermore, recall that data from different environments are independent of each other. This means in particular that:

$$V_{\boldsymbol{n}}(t) \xrightarrow[n_0\to\infty]{d} q_f(t)\cdot \boldsymbol{B}(t),$$

in $L_2(0,1)^E$ with $\boldsymbol{B}(t) \doteq \left((B_e(t))_{e=1}^E\right)^T$ being a vector of E independent Brownian bridges on $[0,1]$.

To show Proposition 1, we need to establish that for any sequence $(\boldsymbol{n}(k))_{k\geq 0}$ of the numbers of observations per environment such that $n_0(k) \doteq \min_{e\in[E]} n_e(k) \to \infty$ as k goes to infinity, we have that:

$$\int_0^1 V_{\boldsymbol{n}(k)}^T(t)\cdot A_{\boldsymbol{w}(k)}^2 \cdot V_{\boldsymbol{n}(k)}(t)dt \xrightarrow[k\to\infty]{d} \sum_{e=1}^{E-1}\int_0^1 q_f^2(t)B_e^2(t)dt. \tag{31}$$

We are going to prove this by the selection principle (see for instance Proposition 1.6 in Chapter 3 of Cinlar (2011)), that is, we are going to show that every sub-sequence of the series in the LHS of equation (31) admits a further sub-sequence that converges in distribution to the RHS of the equation.

Consider a sub-sequence $(\boldsymbol{n}(\phi(k)))_{k\geq 0}$ (with ϕ strictly increasing) of $(\boldsymbol{n}(k))_{k\geq 0}$; since the weights $w_e(\phi(k))$ are in $[0,1]$ and therefore bounded, there is a further sub-sequence $(\boldsymbol{n}(\psi\circ\phi(k)))_{k\geq 0}$ such that:

$$\boldsymbol{w}(\psi\circ\phi(k)) \xrightarrow[k\to\infty]{} \bar{\boldsymbol{w}} \in \bar{\Lambda}.$$

As a consequence:

$$V_{\boldsymbol{n}(\psi\circ\phi(k))}^T(t)\cdot A_{\boldsymbol{w}(\psi\circ\phi(k))}^2 \cdot V_{\boldsymbol{n}(\psi\circ\phi(k))}(t) \xrightarrow[k\to\infty]{d} q_f^2(t)\cdot \boldsymbol{B}(t)^T A_{\bar{\boldsymbol{w}}}^2 \boldsymbol{B}(t).$$

in $L_2(0,1)$ by the continuous mapping theorem.

Since $A_{\bar{\boldsymbol{w}}}$ is symmetric, it can be diagonalized $A_{\bar{\boldsymbol{w}}} = R^T\cdot D\cdot R$ where R is an orthonormal matrix of size $E\times E$ and D is a diagonal matrix composed of the E eigen-values of $A_{\bar{\boldsymbol{w}}}$. Notice that $(\sqrt{\bar{w}_1},\ldots,\sqrt{\bar{w}_E})$ is an eigen vector of $A_{\bar{\boldsymbol{w}}}$ with eigen value 0. Also, notice that the matrix $D_{\bar{\boldsymbol{w}}}^{1/2}\mathbf{1}\mathbf{1}^T D_{\bar{\boldsymbol{w}}}^{1/2}$ in the RHS of the definition of $A_{\bar{\boldsymbol{w}}}$ is of rank one, and hence its null space is of dimension $E-1$; it is easy to see that each of these null vectors is an eigen-vector of $A_{\bar{\boldsymbol{w}}}$ with eigen-value 1. Hence, $D = \mathrm{diag}(\underbrace{1,1,\ldots,1}_{E-1 \text{ times}},0)$.

Call $\bar{\boldsymbol{B}}(t) \doteq R\boldsymbol{B}(t)$; since R is orthonormal, it is easy to check that $\bar{\boldsymbol{B}}(t)$ is a vector of E independent Brownian bridges. We then have

$$\boldsymbol{B}(t)^T A_{\bar{\boldsymbol{w}}}^2 \boldsymbol{B}(t) = \bar{\boldsymbol{B}}^T(t) D \bar{\boldsymbol{B}}(t) = \sum_{e=1}^{E-1}\bar{B}_e^2(t).$$

Therefore, we've just showed that, for any strictly increasing ϕ there exists ψ (also strictly increasing) such that:

$$n(\psi\circ\phi(k))\cdot \mathrm{WV}_{\boldsymbol{w}(\psi\circ\phi(k))}(\hat{\boldsymbol{\nu}}_{\psi\circ\phi(k)}(f)) \xrightarrow[k\to\infty]{d} \sum_{e=1}^{E-1}\int_0^1 B_e^2(t)q_f^2(t)dt.$$

As this limit is exactly the same in distribution regardless of the selected sequence $(\boldsymbol{n}(k))_{k\geq 0}$ (and further sub-sequence $(\boldsymbol{n}(\phi(k)))_{k\geq 0}$), this proves proposition 1.

E.3 Regularity Conditions for Theorem 2 and Theorem 4

We prove Theorem 2 for a generic class of functions $\mathcal{F}'$ – of course, we are mainly interested in the case $\mathcal{F}' \in \{\mathcal{F}_{-k}, k\in[p]\}$. The hypotheses of interest here are therefore the following:

$$\tilde{H}_0(\mathcal{E}) : \Gamma_{\boldsymbol{w}}(\mathcal{F}') = 0, \quad \text{against} \quad \tilde{H}_1(\mathcal{E}) : \Gamma_{\boldsymbol{w}}(\mathcal{F}') > 0.$$

In fact, in Theorem 4 below we prove a slightly more general result than Theorem 2. Indeed, we consider the case where the function class used to compute the test statistic depends on n:

$$\text{Test statistic:} \quad \hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}'_n) \doteq \inf_{f \in \mathcal{F}'_n} \mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(f)) \quad \text{where} \quad \forall n,\ \mathcal{F}'_n \subseteq \mathcal{F}'_{n+1} \text{ and } \mathcal{F}' = \overline{\bigcup_n \mathcal{F}'_n},$$

where the closure in the RHS is w.r.t. the $\|\cdot\|_\infty$ norm. We also call $\hat{f}_n$ the obtained minimizer for $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}'_n)$. For simplicity, we assume that the class of functions $\mathcal{F}'_n$ depends only on the total number of observations n, but one can easily extend our analysis to the case where this class depends on the full array of numbers of observations per environment $\boldsymbol{n} \doteq (n_e)_{e=1}^E$.

A possible choice for $\mathcal{F}'_n$ is the class of the regressors that are linear combinations of some basis functions, such as regression splines, Fourier features or wavelet bases for instance, where the number of bases increases with n; another option is to directly restrict the complexity of $\mathcal{F}'_n$ by adding a regularization term to the initial Wasserstein variance minimization program $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}')$, with its hyperparameter implicitly depending on n. In general, using a restricted class of function $\mathcal{F}'_n$ instead of $\mathcal{F}'$ in finite samples can potentially decrease the number of false negatives by reducing overfitting – at the risk of increasing the number of false positives, though. At least under some conditions, we prove that asymptotically such an approach constitutes a consistent test for $\tilde{H}_0(\mathcal{E})$.

In terms of notations, for a compact set $\mathcal{X} \subset \mathbb{R}^p$ we will call $C_0(\mathcal{X}, ||\cdot||_\infty)$ the set of real-valued continuous functions defined on $\mathcal{X}$, and $W^{d,2}(\mathcal{X})$ the Sobolev space on $\mathcal{X}$ with a degree $d \in \mathbb{N}$ of differentiability (see Definition 6.31 from Debnath and Mikusinski (2005), see also Adams and Fournier (2003)). For any subsets $\mathcal{G}_1, \mathcal{G}_2$ of $C_0(\mathcal{X}, ||\cdot||_\infty)$ we denote their Hausdorff distance by $d_H(\mathcal{G}_1, \mathcal{G}_2) \doteq \max(\sup_{g_1 \in \mathcal{G}_1} d(g_1, \mathcal{G}_2), \sup_{g_2 \in \mathcal{G}_2} d(g_2, \mathcal{G}_1))$ where $d(g_i, \mathcal{G}_j) \doteq \inf_{g_j \in \mathcal{G}_j} ||g_i - g_j||_\infty$ for $i, j \in \{1, 2\}$. Finally, we set $w_e = n_e/n$. We are going to use several useful properties of $W^{d,2}(\mathcal{X})$, for $d > p/2$, that are summarized in the remark below.

Remark 2 (Useful properties of the Sobolev space)**.** *When $d > p/2$ and $\mathcal{X}$ has a smooth boundary (e.g. $\mathcal{X}$ is a ball in $\mathbb{R}^p$), the Sobolev embedding theorem states that $W^{d,2}(\mathcal{X}) \subset C_0(\mathcal{X}, ||\cdot||_\infty)$, see Remark 3 of Cucker and Smale (2002); if we call B_R the ball centered at the origin and of radius $R > 0$ in the Sobolev Space $W^{d,2}(\mathcal{X})$, we also have that $\bar{B}_R$ is a compact subset of $C_0(\mathcal{X}, ||\cdot||_\infty)$, where the closure is w.r.t. the infinite norm's topology – in the rest of our paper the closure will always be meant in that sense. Furthermore, we have $\log(\mathcal{N}(\bar{B}_R, \epsilon)) = O((R/\epsilon)^{p/d})$, where $\mathcal{N}(\bar{B}_R, \epsilon)$ is the covering number of $\bar{B}_R$ by balls of radius $\epsilon > 0$. Finally, the space $W^{d,2}(\mathcal{X})$ is norm equivalent to the RKHS generated by the Matérn kernel $k_{d-p/2,h}$ of degree $d - p/2$ for any scale $h > 0$, see Example 2.6 from Kanagawa et al. (2018) and references therein.*

We can now fully detail our list of regularity conditions for Theorem 4 below (for Theorem 2 these conditions were first summarized in Assumption 2 where $\mathcal{F}' \in \{\mathcal{F}_{-k}, k \in [p]\}$ and $\mathcal{F}'_n = \mathcal{F}', \forall n$):

Assumption 4 (Full detail of the regularity conditions)**.** *The following properties are true:*

(1) The X^e's are bounded, that is there exists a compact set $\mathcal{X} \subseteq \mathbb{R}^p$ with smooth boundary (e.g. a ball) such that $\forall e \in [E]$, $\mathbb{P}(X^e \in \mathcal{X}) = 1$; and the Y^e's are sub-Gaussian,

(2) Data from different environments are independent of each other, that is $(\boldsymbol{x}^1, \boldsymbol{y}^1), \ldots, (\boldsymbol{x}^E, \boldsymbol{y}^E)$ are mutually independent,

(3) There is a constant $\lambda > 0$ independent of the n_e's such that $n_e \geq \lambda n, \forall e \in [E]$,

(4) For some integer $d > p/2$ and Sobolev space $W^{d,2}(\mathcal{X})$, we have: For any $\delta \in (0, 1)$ there exists $R_\delta > 0$ such that for n large enough, with probability at least $1 - \delta$, $\hat{f}_n \in B_{R_\delta}$, where B_{R_δ} is the ball of radius $R_\delta > 0$ centered at the origin in $W^{d,2}(\mathcal{X})$.

In what follows, we will call $\mathcal{G}_n^\delta \doteq \overline{\mathcal{F}'_n \bigcap B_{R_\delta}}$, $\mathcal{G}^\delta \doteq \overline{\mathcal{F}' \bigcap B_{R_\delta}}$ and $\mathcal{G}_0^\delta \doteq \{g \in \mathcal{G}^\delta : WV_{\boldsymbol{w}}(\boldsymbol{\nu}(g)) = 0\}$.

(5) For any $\delta \in (0, 1)$, we have $d_H(\mathcal{G}_n^\delta, \mathcal{G}^\delta) = o\left(n^{-1/2}\right)$,

(6) $\forall f \in \mathcal{G}^\delta, \forall e \in [E], \nu_e(f)$ satisfies Assumption 3 from Section E.1. Furthermore, conditions (ii) and (iii) from Assumption 3 are satisfied uniformly in the following sense:

(6a) $\forall s \in (0, 1/2), \exists M_{\delta,s} > 0$ s.t. $\forall f \in \mathcal{G}^\delta, \forall e \in [E]$, $\sup_{s \leq t \leq 1-s} t(1-t)|q_f^{e\prime}(t)|/q_f^e(t) < M_{\delta,s}$,

(6b) $\forall s \in (0,1/2), \exists M'_{\delta,s} > 0$ *s.t.* $\forall f \in \mathcal{G}^\delta, \forall e \in [E]$, $\int_s^{1-s} t(1-t)q_f^{e\,2}(t)dt < M'_{\delta,s}$,

(6c) And, $\forall e \in [E], \sup_{f\in\mathcal{G}_0^\delta}\left(\int_0^s t(1-t)q_f^{e\,2}(t)dt\right) \vee \left(\int_{1-s}^1 t(1-t)q_f^{e\,2}(t)dt\right) \xrightarrow[s\to 0]{} 0$,

where q_f^e *refers to the quantile density of* $\nu_e(f)$*, for any* $e \in [E]$ *and function* $f \in \mathcal{G}^\delta$.

Remark 3 (On condition (4) of Assumption 4)**.** *Condition (4) typically arises in situations where the class* $\mathcal{F}'_n$ *is composed of smooth functions, and is not too rich so that it doesn't tend to overfit the data by returning near-zero residuals in each environment. For instance, when* $\mathcal{F}'_n = \mathcal{F}', \forall n$*, where* $\mathcal{F}'$ *is the class of linear regressors, we often observe in practice that the minimizer* $\hat{f}_n$ *has coefficients that are not too extreme, which means in particular they are bounded in probability – condition (4) is valid in that case. As mentioned earlier, putting restrictions on the class* $\mathcal{F}'_n$ *by adding a regularization term to the optimization or by considering a number of basis functions that grows slowly in* n *are other ways to insure that condition (4) is valid.*

Remark 4 (On conditions (6a)–(6c) of Assumption 4)**.** *As long as Assumption 3 is true for all* $\nu_e(f)$ *with* $f \in \mathcal{G}^\delta$*, conditions (6a) and (6b) (respectively (6c)) are automatically verified when* $\mathcal{G}^\delta$ *(respectively* $\mathcal{G}_0^\delta$*) is finite. Furthermore, since the quantile density of the normal distribution depends only the variance parameter, if we focus only conditions (6a) and (6b), notice that these conditions are true whenever the observed variables* X^e *and* Y^e *are jointly Gaussian and* $\mathcal{G}^\delta$ *is a bounded class of linear regressors – even though Assumption 3 is itself not verified for the normal distribution. For that reason, if we use a weighted Wasserstein distance as suggested in Remark 1 with appropriate weight function, the conditions of Assumption 4 hold easily when data are generated by a linear Gaussian structural model, with the exception of condition (1) of course – we believe however that this condition can be weaken to* X^e *being sub-Gaussian, for simplicity we keep it as it is.*

Theorem 4 (More general asymptotic guaranties)**.** *Under Assumption 4, for any* $e \in [E]$ *set* $w_e = n_e/n$ *and let* $\hat{q}_n \doteq \sum_{e=1}^E w_e \hat{q}_f^e$ *for* $f = \hat{f}_n$*, where* $\hat{q}_f^e$ *is defined as in Definition 4. Call* $\hat{t}_\alpha$ *the* $(1-\alpha)$*-quantile of the following variable:*

$$\frac{1}{n}\sum_{e=1}^{E-1}\int_0^1 B_e^2(t)\,\hat{q}_n^2(t)dt, \tag{32}$$

where $(B_e(t))_{e=1}^{E-1}$ *are* $E-1$ *independent Brownian bridges between* 0 *and* 1*. Rejecting* $\tilde{H}_0(\mathcal{E})$ *whenever we have* $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}'_n) > \hat{t}_\alpha$ *forms a consistent test of (asymptotic) level* α *when* $n \to \infty$*. That is:*

$$\textit{Under } \tilde{H}_0(\mathcal{E}) : \limsup_{n\to\infty} \mathbb{P}(\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}'_n) > \hat{t}_\alpha) \le \alpha, \textit{ and under } \tilde{H}_1(\mathcal{E}) : \lim_{n\to\infty} \mathbb{P}(\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}'_n) > \hat{t}_\alpha) = 1.$$

Theorem 2 is a direct consequence of Theorem 4 by taking $\mathcal{F}' \in \{\mathcal{F}_{-k}, k \in [p]\}$, $\mathcal{F}'_n = \mathcal{F}', \forall n$ and assuming that Assumption 4 is true for any of these $\mathcal{F}'$'s (as summarized in Assumption 2). Note that condition (5) from Assumption 4 is automatically verified in that case.

E.4 Proofs of Theorems 2 and 4 under $\tilde{H}_0(\mathcal{E})$

As we've just mentioned, Theorem 2 is a direct consequence of Theorem 4, therefore we only focus on proving the latter. To remain concise, we are going to use directly several technical results that are presented as supporting lemmas and proved in Section E.8.

First Steps. Let's fix some arbitrary $\delta \in (0,1/2)$ and $\epsilon > 0$. Note that by condition (4) of Assumption 4, when $n \ge c_1(\delta)$ (for some constant $c_1(\delta)$ depending only on δ) with probability at least $1-\delta$ we have $\hat{f}_n \in \mathcal{G}_n^\delta$. In Lemma 8 we prove that for any functions $g, g' \in C_0(\mathcal{X}, ||\cdot||_\infty)$ we have:

$$\left|\sqrt{\mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g))} - \sqrt{\mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g'))}\right| \le \|g - g'\|_\infty.$$

As a consequence, with probability at least $1-\delta$:

$$\left|\sqrt{\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}'_n)} - \sqrt{\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{G}^\delta)}\right| = \left|\sqrt{\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{G}_n^\delta)} - \sqrt{\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{G}^\delta)}\right| \le d_H(\mathcal{G}_n^\delta, \mathcal{G}^\delta).$$

Therefore:

$$\mathbb{P}(\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}'_n) > \hat{t}_\alpha) \le \mathbb{P}\left(\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{G}^\delta) > \hat{t}_\alpha - 2\sqrt{\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{G}^\delta)}d_H(\mathcal{G}_n^\delta, \mathcal{G}^\delta) - d_H^2(\mathcal{G}_n^\delta, \mathcal{G}^\delta)\right) + \delta. \tag{33}$$

For simplicity we use the notation $q_{1-\alpha}(T)$ to refer to the $(1-\alpha)$-quantile of a real variable T. We are also going to define the following variables for any quantile density function q (potentially empirical) that satisfies condition (iii) from Assumption 3 and $s \in [0, 1/2)$:

$$T_s(q) \doteq \sum_{e=1}^{E-1} \int_s^{1-s} B_e^2(t) q^2(t) dt,$$

where $(B_e(t))_{e=1}^{E-1}$ are $E-1$ independent Brownian bridges. Note that $n\hat{t}_\alpha = q_{1-\alpha}(T_0(\hat{q}_n))$.

First, as a consequence of Theorem 1 and Lemma 8 we can show that $\mathcal{G}_0^\delta$ is an non-empty compact subset of $C_0(\mathcal{X}, ||\cdot||_\infty)$ (see Lemma 10 for a proof). For any $s \in [0, 1/2)$ and $\alpha \in (0, 1)$, we will define also:

$$t_{s,\alpha}^\delta \doteq \inf_{f \in \mathcal{G}_0^\delta} \left(q_{1-\alpha}(T_s(q_f)) \right), \tag{34}$$

where q_f is the quantile density of $\nu_e(f)$ – note that since $f \in \mathcal{G}_0^\delta$, $\nu_e(f)$ is identical for all $e \in [E]$.

Notice that the infimum in (34) is always attained by some function in $\mathcal{G}_0^\delta$, we prove this fact in Lemma 15. In particular we let $f_0^\delta \in \mathcal{G}_0^\delta$ to be a function such that:

$$t_{0,\alpha}^\delta = q_{1-\alpha}(T_0(q_{f_0^\delta})). \tag{35}$$

Since $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{G}^\delta) \leq \mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(f_0^\delta))$, from inequality (33) we get:

$$\mathbb{P}(\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_n') > \hat{t}_\alpha) \leq \mathbb{P}\left(n\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(f_0^\delta)) > n\hat{t}_\alpha - 2\sqrt{n\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(f_0^\delta))} n^{1/2} d_H(\mathcal{G}_n^\delta, \mathcal{G}^\delta) - n d_H^2(\mathcal{G}_n^\delta, \mathcal{G}^\delta) \right) + \delta.$$

Furthermore, because $n^{1/2} d_H(\mathcal{G}_n^\delta, \mathcal{G}^\delta) = o_n(1)$ and, from Proposition 1, $n\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(f_0^\delta))$ converges in distribution, we get that the term in the RHS of $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_n')$ above converges in probability to 0, that is for $n \geq c_2(\delta, \epsilon)$ we have with probability at least $1-\delta$:

$$2\sqrt{n\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(f_0^\delta))} n^{1/2} d_H(\mathcal{G}_n^\delta, \mathcal{G}^\delta) + n d_H^2(\mathcal{G}_n^\delta, \mathcal{G}^\delta) \leq \epsilon,$$

Therefore:

$$\limsup_{n\to\infty} \mathbb{P}(\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_n') > \hat{t}_\alpha) \leq \mathbb{P}\left(n\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(f_0^\delta)) > n\hat{t}_\alpha - \epsilon \right) + 2\delta.$$

Finally, in Lemma 15 we show that there exist $(s, \alpha') \in (0, 1/2) \times (0, 1)$ such that $\alpha' > \alpha$ and $t_{s,\alpha'}^\delta + \epsilon \geq t_{0,\alpha}^\delta$. From now on we fix s and α' to be as such. Note that $n\hat{t}_\alpha = q_{1-\alpha}(T_0(\hat{q}_n)) \geq q_{1-\alpha}(T_s(\hat{q}_n))$. We have:

$$\limsup_{n\to\infty} \mathbb{P}(\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_n') > \hat{t}_\alpha) \leq \limsup_{n\to\infty} \mathbb{P}\left(n\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(f_0^\delta)) > t_{0,\alpha}^\delta + \left(q_{1-\alpha}(T_s(\hat{q}_n)) - t_{0,\alpha}^\delta \right) - \epsilon \right) + 2\delta. \tag{36}$$

What's left is to study the asymptotic behavior of $q_{1-\alpha}(T_s(\hat{q}_n)) - t_{0,\alpha}^\delta$.

Asymptotic Behavior of $q_{1-\alpha}(T_s(\hat{q}_n)) - t_{0,\alpha}^\delta$. In the following, for any $f \in \mathcal{G}^\delta$ we will call q_f^e the quantile density of $\nu_e(f)$ and $q_f \doteq \sum_{e=1}^E w_e q_f^e$. Similarly, we call $\hat{q}_f \doteq \sum_{e=1}^E w_e \hat{q}_f^e$, where $\hat{q}_f^e$ is the kernel estimator from Definition 4 with bandwidth $h_e = \beta n_e^{-1/3}$ for some fixed $\beta > 0$.

First, we would like to measure the convergence of the quantile density estimator $\hat{q}_n$ toward $q_{\hat{f}_n}$. However, since $\hat{f}_n$ is random and not fixed, we cannot directly use convergence results like the one we proved in Theorem 5 from Section E.7. Instead, we will consider a finite cover of $\mathcal{G}^\delta$, fine enough such that there exists a function in this cover not too far from $\hat{f}_n$, so that their corresponding quantile density estimators are close to each other; and coarse enough so we can make sure that with high probability the quantile density estimators at each of the functions in this cover uniformly converge. Let $\epsilon_n \doteq n^{-1/3-r}$ with $r \doteq \frac{1}{3}\frac{2-p/d}{2+p/d} > 0$ (recall that $d > p/2$). By Remark 2 we know that there is a constant C_δ such that:

$$\log\left(\mathcal{N}\left(\mathcal{G}^\delta, \epsilon_n \right) \right) \leq C_\delta \epsilon_n^{-p/d}.$$

Call $\mathcal{C}_n^\delta$ the corresponding ϵ_n-cover of $\mathcal{G}^\delta$, i.e. $\log\left|\mathcal{C}_n^\delta\right| \leq C_\delta \epsilon_n^{-p/d}$. Next, fix $\epsilon' \in (0, 1/2)$ and set $\delta_n \doteq \delta/\left|\mathcal{C}_n^\delta\right|$. Recall that the bandwidths for the kernel estimators are all set to be $h_e = \beta n_e^{-1/3}$ for $e \in [E]$, and that $n \geq n_e \geq \lambda n$ for some constant λ under Assumption 4.

In Lemma 11 we show that (for any $s \in (0, 1/2)$) there exists a constant $C_{\delta,s}$ such that for any $g \in \mathcal{G}^\delta$ and $e \in [E]$ we have $\|q_g^e\|_{s,\infty} \leq C_{\delta,s}$ and $\|q_g^{e\prime}\|_{s,\infty} \leq C_{\delta,s}$, where we define $\|q\|_{s,\infty} \doteq \sup_{t\in[s,1-s]} |q(t)|$. Because of that, we can apply Theorem 5 from Section E.7; we then get that there exist two constants $c_3(s,\delta)$ and $c_4(s,\delta)$ such that for any $n \geq c_3(s,\delta)$ and for any $g \in \mathcal{C}_n^\delta$, with probability at least $1 - \delta_n$:

$$\|\hat{q}_g - q_g\|_{s,\infty} \leq c_4(s,\delta)\left(n^{-1/3} + \frac{\log(n/\delta_n)}{n^{2/3}} + \sqrt{\frac{\log(n/\delta_n)}{n^{2/3}}}\right).$$

Furthermore, one can easily check that $\log(n/\delta_n)n^{-2/3} \leq \log(n/\delta)n^{-2/3} + C_\delta\epsilon_n^{-p/d}n^{-2/3} = \log(n/\delta)n^{-2/3} + C_\delta n^{-2r}$. By a union bound, we therefore get that there exists a constant $c_5(s,\delta,\epsilon')$ such that if $n \geq c_5(s,\delta,\epsilon')$, with probability at least $1-\delta$ we have:

$$\forall g \in \mathcal{C}_n^\delta, \quad \|\hat{q}_g - q_g\|_{s,\infty} \leq \epsilon'. \tag{37}$$

Besides, from Lemma 16 we get that when $n \geq c_6(\delta,\epsilon')$ (for some constant $c_6(\delta,\epsilon')$), with probability at least $1-2\delta$:

$$\hat{f}_n \in \mathcal{G}^\delta, \quad \text{and} \quad \exists \bar{f}_n \in \mathcal{G}_0^\delta \quad \text{s.t.} \quad \|\bar{f}_n - \hat{f}_n\|_\infty < \epsilon'. \tag{38}$$

Under the above event, there exists a function $\bar{g}_n \in \mathcal{C}_n^\delta$ such that $\|\hat{f}_n - \bar{g}_n\|_\infty \leq \epsilon_n$. By Lemma 17 we can choose a constant $c_7(s,\epsilon')$ such that whenever $n \geq c_7(s,\epsilon')$, we have both $4\beta^{-1}Ln^{-r} \leq \epsilon'$ and:

$$\|\hat{q}_n - \hat{q}_{\bar{g}_n}\|_{s,\infty} \leq 4\beta^{-1}Ln^{1/3}\epsilon_n = 4\beta^{-1}Ln^{-r} \leq \epsilon'.$$

For simplicity, let's also assume that $c_7(s,\epsilon')$ was chosen so that $\epsilon_n \leq \epsilon'$ for $n \geq c_7(s,\epsilon')$. Under the event of equation (38), notice that for $n \geq \max(c_6(\delta,\epsilon'), c_7(s,\epsilon'))$, we have $\|\bar{g}_n - \bar{f}_n\| \leq 2\epsilon' < 1$, and Lemma 12 implies:

$$\left(\int_s^{1-s}\left(q_{\bar{g}_n}(t) - q_{\bar{f}_n}(t)\right)^2 dt\right)^{1/2} \leq A_{\delta,s}^{1/2}(2\epsilon')^{1/3},$$

for some constant $A_{\delta,s}$.

Consider the decomposition $\hat{q}_n - q_{\bar{f}_n} = \hat{q}_n - \hat{q}_{\bar{g}_n} + \hat{q}_{\bar{g}_n} - q_{\bar{g}_n} + q_{\bar{g}_n} - q_{\bar{f}_n}$. Combining the events of (37) and (38), by a union bound, we get that if $n \geq c_8(s,\delta,\epsilon') \doteq \max(c_5(s,\delta,\epsilon'), c_6(\delta,\epsilon'), c_7(s,\epsilon'))$, with probability at least $1-3\delta$:

$$\left(\int_s^{1-s}\left(\hat{q}_n(t) - q_{\bar{f}_n}(t)\right)^2 dt\right)^{1/2} \leq \epsilon' + \epsilon' + A_{\delta,s}^{1/2}(2\epsilon')^{1/3}.$$

Which, by Lemma 14 implies that:

$$\sqrt{q_{1-\alpha}(T_s(\hat{q}_n))} \geq \sqrt{q_{1-\alpha'}(T_s(q_{\bar{f}_n}))} - \frac{(E-1)^{1/2}}{\alpha'-\alpha}\left(2\epsilon' + A_{\delta,s}^{1/2}(2\epsilon')^{1/3}\right).$$

Because of our choice of α', s and that $\bar{f} \in \mathcal{G}_0^\delta$, by definition $q_{1-\alpha'}(T_s(q_{\bar{f}_n})) \geq t_{s,\alpha'}^\delta \geq t_{0,\alpha}^\delta - \epsilon$. Also, since ϵ' was arbitrarily chosen, we can take ϵ' small enough to obtain the following result: For $n \geq c_8(s,\delta,\epsilon')$, we have with probability at least $1-3\delta$:

$$q_{1-\alpha}(T_s(\hat{q}_n)) - t_{0,\alpha}^\delta \geq -2\epsilon. \tag{39}$$

Conclusion. Combining (36) with (39) yields the following result:

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}(\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_n') > \hat{t}_\alpha) &\leq \limsup_{n\to\infty} \mathbb{P}\left(n\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(f_0^\delta)) > t_{0,\alpha}^\delta - 3\epsilon\right) + 5\delta \\ &\leq \mathbb{P}\left(T_0(q_{f_0^\delta}) > t_{0,\alpha}^\delta - 3\epsilon\right) + 5\delta,\end{aligned}$$

where the last inequality (which is actually an equality) comes from Proposition 1. Since ϵ was arbitrarily chosen, we can send it to 0 and get that $\mathbb{P}\left(T_0(q_{f_0^\delta}) > t_{0,\alpha}^\delta - 3\epsilon\right)$ goes to α by equation (35). Finally, δ was also arbitrary, sending it to 0 then yields our result.

E.5 Proofs of Theorems 2 and 4 under $\tilde{H}_1(\mathcal{E})$

Again, since Theorem 2 is a direct consequence of Theorem 4 we only focus on proving the latter. The proof of the consistency of the test proposed in Theorem 4 is achieved in two steps: We show that, under $\tilde{H}_1(\mathcal{E})$, the statistic $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}'_n)$ is asymptotically lower bounded by a positive constant independent of n, while the threshold $\hat{t}_\alpha$ converges in probability toward 0. Furthermore, we set $\delta \in (0,1)$.

Lower Bound on $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}'_n)$. In the first steps of the proof of Theorem 4 under $\tilde{H}_0(\mathcal{E})$ (see Section E.4) we observed that with probability at least $1-\delta$:

$$\left|\sqrt{\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}'_n)} - \sqrt{\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{G}^\delta)}\right| \leq d_H(\mathcal{G}_n^\delta, \mathcal{G}^\delta).$$

Therefore, we can derive this first lower bound on $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}'_n)$:

$$\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}'_n) \geq \left(\sqrt{\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{G}^\delta)} - d_H(\mathcal{G}_n^\delta, \mathcal{G}^\delta)\right)^2 \mathbb{1}\left\{\sqrt{\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{G}^\delta)} - d_H(\mathcal{G}_n^\delta, \mathcal{G}^\delta) \geq 0\right\}. \tag{40}$$

Now, we are going to derive a lower bound on $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{G}^\delta)$. In the proof of Lemma 10 we showed that $\mathcal{R}_{n_e}(\mathcal{G}^\delta) = O(n_e^{-1/2})$ by using the fact that the Sobolev space $W^{d,2}(\mathcal{X})$ is norm-equivalent to the RKHS generated by the Matérn kernel (see Remark 2). By Theorem 1, such a property is indeed useful for deriving bounds on $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{G}^\delta)$. In particular, from equation (55) of the proof of Lemma 10 we have that with probability at least $1-\delta$:

$$\forall g \in \mathcal{G}^\delta, \quad |\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(g)) - \mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g))| \leq c_\delta \frac{\log^2(n)}{\sqrt{n}}, \tag{41}$$

for some constant c_δ that depends only on δ.

Furthermore, by Lemmas 8 and 9, since $\mathcal{G}^\delta$ is a compact subset of $C_0(\mathcal{X}, ||\cdot||_\infty)$ (see Remark 2), under condition (3) of Assumption 4 and $\tilde{H}_1(\mathcal{E})$, we can find $\gamma_0 > 0$ independent of $\boldsymbol{w}$ such that

$$\forall g \in \mathcal{G}^\delta, \quad \mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g)) \geq \gamma_0 > 0. \tag{42}$$

Combining (41) and (42) it is easy to see that for n large enough, we have with probability at least $1-\delta$ that $0 < \gamma_0/2 \leq \hat{\Gamma}_{\boldsymbol{w}}(\mathcal{G}^\delta)$. Recall also that by condition (5) of Assumption 4 we have $d_H(\mathcal{G}_n^\delta, \mathcal{G}^\delta) = o(n^{-1/2})$. Therefore, from (40) we get that there exist constants $\gamma > 0$ and $c(\delta, \gamma) > 0$ such that for any $n \geq c(\delta, \gamma)$, with probability at least $1-\delta$:

$$\forall \boldsymbol{w} \in \Lambda \text{ s.t. } \min_{e\in[E]} w_e \geq \lambda, \quad \hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}'_n) \geq \gamma > 0, \tag{43}$$

where $\lambda > 0$ is from condition (3) of Assumption 4.

$\hat{t}_\alpha$ Converges to 0. Call $Z_e^\delta \doteq \sup_{g\in\mathcal{G}^\delta} |Y^e - g(X^e)|$, for $e \in [E]$. Since $\mathcal{G}^\delta$ is a bounded subset in $C_0(\mathcal{X}, ||\cdot||_\infty)$ and Y^e is sub-Gaussian from condition (1) of Assumption 4, we have that Z_e^δ is sub-Gaussian $\mathcal{G}(\mu_e^\delta, \sigma_e^\delta)$ for some mean μ_e^δ and sub-Gaussian parameter σ_e^δ (see Remark D.1). Let $\mu_\delta \doteq \max_{e\in[E]} \mu_e^\delta$ and $\sigma_\delta \doteq \max_{e\in[E]} \sigma_e^\delta$. Call also for any $e \in [E]$ and $i \in [n_e]$, $z_{i,e}^\delta \doteq \sup_{g\in\mathcal{G}^\delta} |y_i^e - g(x_i^e)|$ and set $m_\delta \doteq \max_{e\in[E]} \max_{i\in[n_e]} z_{i,e}^\delta$. By a union bound and because the Z_e^δ's are sub-Gaussian, we have (see also Chapter 2 from Wainwright (2019)):

$$\mathbb{P}(m_\delta > M) \leq \sum_{e=1}^{E} n_e \mathbb{P}(Z_e > M) \leq n \exp\left(-\frac{(M-\mu_\delta)^2}{2\sigma_\delta^2}\right) \leq \delta, \tag{44}$$

when we set $M = \mu_\delta + \sigma_\delta\sqrt{2\log(n/\delta)}$. Therefore with probability at least $1-\delta$ we have $m_\delta \leq \mu_\delta + \sigma_\delta\sqrt{2\log(n/\delta)}$.

Recall that we set the bandwidth h_e for the kernel estimator in Definition 4 at $h_e = \beta n_e^{-1/3}$, for some constant $\beta > 0$. Notice that from Definition 4, for any $e \in [E]$, $f \in \mathcal{G}^\delta$ and $t \in [0,1]$:

$$|\hat{q}_f^e(t)| \leq \sum_{i=2}^{n_e} (\epsilon_{(i)}^e(f) - \epsilon_{(i-1)}^e(f)) \frac{\|K\|_\infty}{h_e} \leq (\epsilon_{(n_e)}^e(f) - \epsilon_{(1)}^e(f)) n^{1/3}\beta^{-1}\|K\|_\infty \leq 2m_\delta n^{1/3}\beta^{-1}\|K\|_\infty. \tag{45}$$

Along with condition (4) of Assumption 4, using (44) and (45) yields that there is a constant $c'(\delta)$ such that for any $n \geq c'(\delta)$ we have with probability at least $1 - 2\delta$:

$$\hat{f}_n \in \mathcal{G}^\delta, \quad \text{and} \quad \|\hat{q}_n\|_\infty \leq 2n^{1/3}\beta^{-1}\|K\|_\infty(\mu_\delta + \sigma_\delta\sqrt{2\log(n/\delta)}).$$

Furthermore, if we use the notations from Section E.4:

$$\hat{t}_\alpha = \frac{1}{n}q_{1-\alpha}(T_0(\hat{q}_n)) \leq \frac{1}{n\alpha}\mathbb{E}\left[\sum_{e=1}^{E-1}\int_0^1 B_e^2(t)\hat{q}_n^2(t)dt\right] = \frac{(E-1)}{n\alpha}\int_0^1 t(1-t)\hat{q}_n^2(t)dt.$$

Therefore for any $n \geq c'(\delta)$, with probability at least $1 - 2\delta$ we have:

$$\hat{t}_\alpha \leq 4\frac{(E-1)\|K\|_\infty^2}{n^{1/3}\alpha\beta^2}\left(\mu_\delta + \sigma_\delta\sqrt{2\log(n/\delta)}\right)^2 \xrightarrow[n\to\infty]{} 0. \tag{46}$$

Conclusion. As a consequence of both (43) and (46), we get:

$$\liminf_{n\to\infty} \mathbb{P}(\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}'_n) > \hat{t}_\alpha) = 1 - 3\delta.$$

Since δ was chosen arbitrarily, we can send it to 0; and it concludes our proof.

E.6 Proof of Theorem 3

First, notice that by identifiability it is direct that $\tilde{S}(\mathcal{E}) = S^*$ (we prove this fact in Lemma 18). For each $k \in [p]$ we will call $\hat{t}_\alpha^k$ the threshold used in Theorem 2 for the statistic $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_{-k})$. By a union bound:

$$\mathbb{P}(\hat{S}(\mathcal{E}) = \tilde{S}(\mathcal{E})) \geq 1 - \mathbb{P}\left(\exists k \notin \tilde{S}(\mathcal{E}) \text{ s.t. } \hat{\Gamma}(\mathcal{F}_{-k}) > \hat{t}_\alpha^k\right) - \sum_{k\in\tilde{S}(\mathcal{E})}\mathbb{P}\left(\hat{\Gamma}(\mathcal{F}_{-k}) \leq \hat{t}_\alpha^k\right).$$

First notice that, by Theorem 2, for $k \in \tilde{S}(\mathcal{E})$, $\mathbb{P}(\hat{\Gamma}(\mathcal{F}_{-k}) \leq \hat{t}_\alpha^k) \to_{n\to\infty} 0$. Then, because that $\hat{\Gamma}(\mathcal{F}_{-k}) \leq \mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(f^*))$ for $k \notin \tilde{S}(\mathcal{E}) = S^*$ we have:

$$\mathbb{P}\left(\exists k \notin \tilde{S}(\mathcal{E}) \text{ s.t. } \hat{\Gamma}(\mathcal{F}_{-k}) > \hat{t}_\alpha^k\right) \leq \mathbb{P}\left(\exists k \notin \tilde{S}(\mathcal{E}) \text{ s.t. } n\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(f^*)) > n\hat{t}_\alpha^k\right).$$

Set any $\delta \in (0, 1/2)$ and $\epsilon > 0$. Using the notations from Section E.4, we can see that by identifiability, for any $k \in [p]$, we have the corresponding $t_{0,\alpha}^\delta$ (when we set $\mathcal{F}' = \mathcal{F}_{-k}$ in the proof of Theorems 2 and 4 in Section E.4) is actually equal to $q_{1-\alpha}(T_0(q_{f^*}))$. Hence the result (39) from Section E.4 implies that for any $k \in [p]$, we have a constant c_k such that whenever $n \geq c_k$, with probability at least $1 - 3\delta$:

$$n\hat{t}_\alpha^k - q_{1-\alpha}(T_0(q_{f^*})) \geq -2\epsilon.$$

Therefore by a union bound and using Proposition 1 we get:

$$\begin{aligned}\limsup_{n\to\infty}\mathbb{P}\left(\exists k \notin \tilde{S}(\mathcal{E}) \text{ s.t. } n\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(f^*)) > n\hat{t}_\alpha^k\right) &\leq \limsup_{n\to\infty}\mathbb{P}\left(n\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(f^*)) > q_{1-\alpha}(T_0(q_{f^*})) - 2\epsilon\right) + 3p\delta \\ &\leq \mathbb{P}\left(T_0(q_{f^*}) > q_{1-\alpha}(T_0(q_{f^*})) - 2\epsilon\right) + 3p\delta.\end{aligned}$$

Since δ and ϵ can be chosen arbitrarily small and $\mathbb{P}\left(T_0(q_{f^*}) > q_{1-\alpha}(T_0(q_{f^*}))\right) = \alpha$, the above steps implies that:

$$\liminf_{n\to\infty}\mathbb{P}(\hat{S}(\mathcal{E}) = \tilde{S}(\mathcal{E})) \geq 1 - \alpha.$$

E.7 High Probability Error Bounds for the Quantile Density Estimator

We derive here high probability bounds on the supremum of the absolute difference between a quantile density q and its estimator $\hat{q}$, as defined in Definition 4, over intervals of the form $[s, 1-s]$, where $s \in (0, 1/2)$. Such high probability bounds are needed for our proof of Theorem 2 and, to the best of our knowledge, cannot be found in the literature. The known theoretical guarantees for kernel quantile or quantiles density estimators (Falk, 1986; Sheather and Marron, 1990; Jones, 1992) only focus on the mean square error at a fixed point in $[0, 1]$ and require that q is twice differentiable, an assumption we do not make. Therefore the following result can also be of independent interest.

Theorem 5 (Quantile density error bounds)**.** *Let $(Z_i)_{i=1}^n \in \mathbb{R}$ be $n > 1$ i.i.d. samples of a distribution with twice differentiable CDF F such that its quantile function F^{-1} is also twice differentiable on $(0, 1)$ with first derivative q and second derivative q'.*

For any $s \in (0, 1/2)$, assume there is a constant $C_s > 0$ such that $q(t) \vee |q'(t)| < C_s$ for any $t \in [s, 1-s]$. Call $(Z_{(i)})_{i=1}^n$ the order statistics of the sample $(Z_i)_{i=1}^n$ and let:

$$\hat{q}(t) = \sum_{i=2}^{n} \left(Z_{(i)} - Z_{(i-1)}\right) K_h\left(t - \frac{i-1}{n}\right),$$

where $h > 1/n$ and $K_h(\cdot) \doteq h^{-1}K(\cdot/h)$ with K a L-Lipschitz kernel supported on $[-1, 1]$ such that $\int K(u)du = 1$. Finally, call also $\|K\|_\infty = \sup_{u\in[-1,1]} |K(u)|$. For $\delta \in (0, 1)$ and $s \in (0, 1/2)$, whenever n and h satisfy the following condition:

$$h + \sqrt{\frac{3\log(6n/\delta)}{nh-1}} + \frac{1}{n} \leq \frac{s}{2}, \tag{47}$$

we have that with probability at least $1 - \delta$:

$$\sup_{t\in[s,1-s]} |\hat{q}(t) - q(t)| \leq \frac{A}{h}\left(2h + \sqrt{\frac{\log(6n/\delta)}{2n}}\right)^2 + B_{n,h}\sqrt{\frac{3\log(6n/\delta)}{nh-1}} + \frac{C}{nh}, \tag{48}$$

where $A \doteq 2C_{s/2}\|K\|_\infty$, $B_{n,h} \doteq C_s\left(12\|K\|_\infty + \frac{L}{nh^{3/2}}\right)$ and $C \doteq 11C_sL$.

Proof. First note that $(Z_{(i)})_{i=1}^n \overset{d}{=} (F^{-1}(U_{(i)}))_{i=1}^n$ where $U_{(i)}$ is the i^{th} order statistic of n i.i.d. uniform variables on $[0, 1]$. It is also well-known that $(U_{(i)})_{i=1}^n \overset{d}{=} (S_i/S_{n+1})_{i=1}^n$ where $S_i = \sum_{j=1}^i \xi_j$ and $(\xi_j)_{j=1}^{n+1} \overset{i.i.d.}{\sim} \exp(1)$ (see Section 2 from Del Barrio et al. (2005) for instance). As we are only interested in bounds in probability, we can actually replace $Z_{(i)}$ by $F^{-1}(U_{(i)})$ and $U_{(i)}$ by S_i/S_{n+1} in our analysis. Finally, note that we have $U_{(i)} \sim \text{Beta}(i, n + 1 - i)$.

Let $s \in (0, 1/2)$ and take $t \in [s, 1-s]$. We are first going to rewrite $\hat{q}(t)$ in a more useful way. Since K is supported on $[-1, 1]$, we have:

$$\forall i \notin [\,\lceil n(t-h)\rceil + 1, \lfloor n(t+h)\rfloor + 1\,], \quad K_h\left(t - \frac{i-1}{n}\right) = 0.$$

Under condition (47) we have $[\,\lceil n(t-h)\rceil + 1, \lfloor n(t+h)\rfloor + 1\,] \subseteq [2, n]$. Therefore,

$$\hat{q}(t) = \sum_{i=\lceil n(t-h)\rceil+1}^{\lfloor n(t+h)\rfloor+1} \left(F^{-1}(U_{(i)}) - F^{-1}(U_{(i-1)})\right) K_h\left(t - \frac{i-1}{n}\right).$$

Finally, by the mean value theorem, we get that there exist variables $\kappa_i \in [U_{(i-1)}, U_{(i)}]$ such that $F^{-1}(U_{(i)}) - F^{-1}(U_{(i-1)}) = (U_{(i)} - U_{(i-1)})q(\kappa_i)$. Hence,

$$\hat{q}(t) = \sum_{i=\lceil n(t-h)\rceil+1}^{\lfloor n(t+h)\rfloor+1} \left(U_{(i)} - U_{(i-1)}\right) q(\kappa_i) K_h\left(t - \frac{i-1}{n}\right).$$

As a consequence, if we define the following three functions depending on t:

$$A_{n,h}(t) \doteq \left| \sum_{i=\lceil n(t-h)\rceil+1}^{\lfloor n(t+h)\rfloor+1} \left(U_{(i)} - U_{(i-1)}\right) (q(\kappa_i) - q(t)) K_h\left(t - \frac{i-1}{n}\right) \right|,$$
$$B_{n,h}(t) \doteq \left| \sum_{i=\lceil n(t-h)\rceil+1}^{\lfloor n(t+h)\rfloor+1} q(t)\left(U_{(i)} - U_{(i-1)} - \frac{1}{n+1}\right) K_h\left(t - \frac{i-1}{n}\right) \right|,$$
$$C_{n,h}(t) \doteq \left| q(t) - \sum_{i=\lceil n(t-h)\rceil+1}^{\lfloor n(t+h)\rfloor+1} \frac{q(t)}{n+1} K_h\left(t - \frac{i-1}{n}\right) \right|,$$

we can then obviously bound the absolute error as follows:

$$|\hat{q}(t) - q(t)| \leq A_{n,h}(t) + B_{n,h}(t) + C_{n,h}(t).$$

Bounding $A_{n,h}(t)$. We first provide a high probability bound for $A_{n,h}(t)$, uniformly for $t \in [s, 1-s]$. From Marchal and Arbel (2017), Theorem 1, a Beta(α, β) distribution is sub-Gaussian with parameter $1/4(\alpha+\beta+1)$. Therefore (see Wainwright (2019), Chapter 2):

$$\forall i \in [n], \forall \epsilon > 0, \quad \mathbb{P}\left(\left|U_{(i)} - \frac{i}{n+1}\right| \geq \epsilon\right) \leq 2\exp\left(-2\epsilon^2(n+2)\right).$$

Using a union bound we get that for any $\delta \in (0,1)$, we have with probability at least $1 - \delta/3$:

$$\forall i \in [n], \quad \left|U_{(i)} - \frac{i}{n+1}\right| < \sqrt{\frac{\log(6n/\delta)}{2(n+2)}}. \tag{49}$$

When inequality (49) and condition (47) are true, for $i \in [\,\lceil n(t-h)\rceil + 1, \lfloor n(t+h)\rfloor + 1\,]$, we have:

$$\begin{aligned}
\kappa_i \in \left[U_{(\lceil n(t-h)\rceil)}, U_{(\lfloor n(t+h)\rfloor+1)}\right] &\subseteq \left[\frac{\lceil n(t-h)\rceil}{n+1} - \sqrt{\frac{\log(6n/\delta)}{2(n+2)}}, \frac{\lfloor n(t+h)\rfloor+1}{n+1} + \sqrt{\frac{\log(6n/\delta)}{2(n+2)}}\right] \\
&\subseteq \left[t - h - \sqrt{\frac{\log(6n/\delta)}{2(n+2)}} - \frac{1}{n+1}, t + h + \sqrt{\frac{\log(6n/\delta)}{2(n+2)}} + \frac{1}{n+1}\right] \\
&\subseteq \left[\frac{s}{2}, 1 - \frac{s}{2}\right].
\end{aligned}$$

Therefore, with probability at least $1 - \delta/3$:

$$\begin{aligned}
\forall t \in [s, 1-s], \quad A_{n,h}(t) &\leq \frac{\|K\|_\infty}{h} \sum_{i=\lceil n(t-h)\rceil+1}^{\lfloor n(t+h)\rfloor+1} \left(U_{(i)} - U_{(i-1)}\right) \sup_{u \in [s/2, 1-s/2]} |q'(u)|\,|\kappa_i - t| \\
&\leq \frac{C_{s/2}\|K\|_\infty}{h} \left(U_{(\lceil n(t-h)\rceil)} - U_{(\lfloor n(t+h)\rfloor+1)}\right)\left(h + \sqrt{\frac{\log(6n/\delta)}{2(n+2)}} + \frac{1}{n+1}\right) \\
&\leq \frac{2C_{s/2}\|K\|_\infty}{h}\left(h + \sqrt{\frac{\log(6n/\delta)}{2(n+2)}} + \frac{1}{n+1}\right)^2 \leq \frac{A}{h}\left(2h + \sqrt{\frac{\log(6n/\delta)}{2n}}\right)^2.
\end{aligned}$$

Bounding $B_{n,h}(t)$. Notice that $2nh-2 \leq \lfloor n(t+h) \rfloor - \lceil n(t-h) \rceil \leq 2nh$. Thus,

$$\begin{aligned} B_{n,h}(t) &= |q(t)| \left| \sum_{i=\lceil n(t-h) \rceil + 1}^{\lfloor n(t+h) \rfloor + 1} \left(U_{(i)} - U_{(i-1)} - \frac{1}{n+1} \right) \left(K_h \left(\frac{\lfloor nt \rfloor}{n} - \frac{i-1}{n} \right) \right. \right. \\ &\qquad \left. \left. + K_h \left(t - \frac{i-1}{n} \right) - K_h \left(\frac{\lfloor nt \rfloor}{n} - \frac{i-1}{n} \right) \right) \right| \\ &\leq C_s \underbrace{\left| \sum_{i=\lceil n(t-h) \rceil + 1}^{\lfloor n(t+h) \rfloor + 1} \left(U_{(i)} - U_{(i-1)} - \frac{1}{n+1} \right) K_h \left(\frac{\lfloor nt \rfloor}{n} - \frac{i-1}{n} \right) \right|}_{\doteq \textcircled{1}} \\ &\qquad + \frac{C_s L}{nh^2} \underbrace{\left(U_{(\lfloor n(t+h) \rfloor + 1)} - U_{(\lceil n(t-h) \rceil)} + \frac{2nh+1}{n+1} \right)}_{\doteq \textcircled{2}}. \end{aligned} \tag{50}$$

Starting first with the second term in the last inequality above, from (49) we know that we can bound it with probability at least $1 - \delta/3$ as follows:

$$\forall t \in [s, 1-s], \quad \textcircled{2} \leq 2h + \sqrt{\frac{2\log(6n/\delta)}{n+2}} + \frac{2nh+3}{n+1} \leq 7h + \sqrt{\frac{2\log(6n/\delta)}{n+2}}. \tag{51}$$

Now turning to the first term, we also have:

$$\begin{aligned} \textcircled{1} &= \left| \sum_{i=\lceil n(t-h) \rceil + 1}^{\lfloor n(t+h) \rfloor + 1} \left(\frac{\xi_i}{S_{n+1}} - \frac{1}{n+1} \right) K_h \left(\frac{\lfloor nt \rfloor}{n} - \frac{i-1}{n} \right) \right| \\ &\leq \frac{n+1}{S_{n+1}} \cdot \left[\frac{1}{(n+1)h} \left| \sum_{i=\lceil n(t-h) \rceil + 1}^{\lfloor n(t+h) \rfloor + 1} (\xi_i - 1) K \left(\frac{\lfloor nt \rfloor}{nh} - \frac{i-1}{nh} \right) \right| \right. \\ &\qquad \left. + \frac{1}{(n+1)h} \left| \sum_{i=\lceil n(t-h) \rceil + 1}^{\lfloor n(t+h) \rfloor + 1} \left(\frac{S_{n+1}}{n+1} - 1 \right) K \left(\frac{\lfloor nt \rfloor}{nh} - \frac{i-1}{nh} \right) \right| \right] \\ &\leq \frac{n+1}{S_{n+1}} \cdot \left[\frac{1}{(n+1)h} \left| \sum_{i=\lceil n(t-h) \rceil + 1}^{\lfloor n(t+h) \rfloor + 1} (\xi_i - 1) K \left(\frac{\lfloor nt \rfloor}{nh} - \frac{i-1}{nh} \right) \right| + \frac{2nh+1}{(n+1)h} \left| \frac{S_{n+1}}{n+1} - 1 \right| \|K\|_\infty \right]. \end{aligned} \tag{52}$$

And we can in turn bound the last two terms using concentration bounds for sub-exponential variables (Wainwright, 2019, eq. (2.20)). In particular, for any $\alpha \in \mathbb{R}$, $\alpha(\xi_i - 1)$ is sub-exponential with parameters $(2\alpha, 2\alpha)$, which means that for any $i, j, k \in [n]$ s.t. $i + 2nh - 2 \leq j$:

$$\mathbb{P}\left(\left| \frac{1}{j-i+1} \sum_{l=i}^{j} (\xi_l - 1) K \left(\frac{k}{nh} - \frac{l-1}{nh} \right) \right| \geq \epsilon \right) \leq 2\exp(-(j-i+1)\epsilon^2/8\|K\|_\infty),$$

as long as $0 \leq \epsilon \leq 2\|K\|_\infty$. Therefore by a union bound we get that with probability at least $1 - \delta/3$:

$$\forall i, j, k \in [n] \text{ s.t. } i + 2nh - 2 \leq j, \left| \frac{1}{j-i+1} \sum_{l=i}^{j} (\xi_l - 1) K \left(\frac{k}{nh} - \frac{l-1}{nh} \right) \right| < 2\sqrt{\frac{3\log(6n/\delta)}{nh-1}} \|K\|_\infty, \tag{53}$$

whenever $2\|K\|_\infty \sqrt{\frac{3\log(6n/\delta)}{nh-1}} \leq 2\|K\|_\infty$, which is true under condition (47). Recall again that $2nh - 2 \leq \lfloor n(t+h) \rfloor - \lceil n(t-h) \rceil \leq 2nh$ and that $nh > 1$. Therefore when (53) is true, we have:

$$\forall t \in [s, 1-s], \quad \frac{1}{(n+1)h} \left| \sum_{i=\lceil n(t-h) \rceil + 1}^{\lfloor n(t+h) \rfloor + 1} (\xi_i - 1) K \left(\frac{\lfloor nt \rfloor}{nh} - \frac{i-1}{nh} \right) \right| \leq 6\sqrt{\frac{3\log(6n/\delta)}{nh-1}} \|K\|_\infty.$$

Similarly to (53), we have with probability at least $1-\delta/3$:

$$\left|\frac{S_{n+1}}{n+1}-1\right| \leq 2\sqrt{\frac{2\log(6/\delta)}{n+1}} \leq \sqrt{\frac{3\log(6n/\delta)}{nh-1}} \leq 1/4, \tag{54}$$

where the last inequality is implied by condition (47) since $h \leq s/2 \leq 1/4$.

By a union bound over equations (53) and (54), from inequality (52) we therefore have with probability at least $1-2\delta/3$:

$$\forall t \in [s, 1-s], \quad \textcircled{1} \leq 12\|K\|_\infty \sqrt{\frac{3\log(6n/\delta)}{nh-1}}.$$

And finally with a union bound over the three events (49), (53) and (54) we get that, with probability at least $1-\delta$, the uniform bound we derived for $A_{n,h}(t)$ holds as well as the following, by (50):

$$\begin{aligned}
\forall t \in [s, 1-s], \quad B_{n,h}(t) &\leq C_s \cdot \textcircled{1} + \frac{C_s L}{nh^2} \cdot \textcircled{2} \\
&\leq 12 C_s \|K\|_\infty \sqrt{\frac{3\log(6n/\delta)}{nh-1}} + \frac{7C_s L}{nh} + \frac{C_s L}{nh^{3/2}}\sqrt{\frac{3\log(6n/\delta)}{nh-1}} \\
&\leq B_{n,h}\sqrt{\frac{3\log(6n/\delta)}{nh-1}} + \frac{7C_s L}{nh}.
\end{aligned}$$

Bounding $C_{n,h}(t)$. Recall that $K_h(t-(i-1)/n) = 0$ whenever $i \notin [\,\lceil n(t-h)\rceil + 1, \lfloor n(t+h)\rfloor + 1\,]$. Therefore we have:

$$C_{n,h}(t) = |q(t)|\left|1 - \frac{1}{n+1}\sum_{i=-\infty}^{+\infty} K_h\left(t-\frac{i}{n}\right)\right| \leq C_s \frac{n}{n+1}\left(\frac{1}{n} + \left|1 - \frac{1}{n}\sum_{i=-\infty}^{+\infty} K_h\left(t-\frac{i}{n}\right)\right|\right).$$

Recall that by assumption $\int K_h(u)du = 1$, hence:

$$\begin{aligned}
\left|1 - \frac{1}{n}\sum_{i=-\infty}^{+\infty} K_h\left(t-\frac{i}{n}\right)\right| &\leq \sum_{i=-\infty}^{+\infty}\int_{t-(i+1)/n}^{t-i/n} |K_h(u) - K_h(t-i/n)|\,du \\
&\leq \sum_{i=\lceil n(t-h)\rceil - 1}^{\lfloor n(t+h)\rfloor} \frac{L}{n^2h^2} \leq \frac{(2nh+2)L}{n^2h^2} \leq \frac{4L}{nh}.
\end{aligned}$$

Therefore,

$$\forall t \in [s, 1-s], \quad C_{n,h}(t) \leq \frac{4C_s L}{nh}.$$

□

E.8 Supporting Lemmas

Note that for the following lemmas, we are going to use the notations from section E.3 without necessarily re-introducing them.

Lemma 8. *Consider any $g, g' \in C_0(\mathcal{X}, ||\cdot||_\infty)$. Under condition (1) of Assumption 4, we have:*

$$\left|\sqrt{WV_{\boldsymbol{w}}(\boldsymbol{\nu}(g))} - \sqrt{WV_{\boldsymbol{w}}(\boldsymbol{\nu}(g'))}\right| \leq \|g - g'\|_\infty.$$

The result above holds also if we replace $\boldsymbol{\nu}(g), \boldsymbol{\nu}(g')$ by their empirical counterparts $\hat{\boldsymbol{\nu}}(g)$ and $\hat{\boldsymbol{\nu}}(g')$.

Proof. Using the notation of Section 5, we call $F^{-1}_{g,e}$ the quantile function of $\nu_e(g)$ for $e \in [E]$, and $F^{-1}_{g',e}$ is defined similarly for g'. Then, by equation 11:

$$\mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g)) = \int_0^1 \sum_{e=1}^E w_e \left(F^{-1}_{g,e}(t) - \sum_{e'=1}^E w_{e'} F^{-1}_{g,e'}\right)^2 dt.$$

For short, call $G_g(t,e) = F^{-1}_{g,e}(t) - \sum_{e'=1}^{E} w_{e'} F^{-1}_{g,e'}(t)$. We have that:

$$\mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g)) = \int G_g^2(t,e) d\lambda \otimes p_{\boldsymbol{w}}(t,e) = \|G_g\|^2_{\mathrm{L}_2([0,1]\times[E],\lambda\otimes p_{\boldsymbol{w}})},$$

where λ is the Lebesgue measure on $[0,1]$ and $p_{\boldsymbol{w}}$ is the probability measure on $[E]$ with probabilities $\boldsymbol{w}$. Therefore, by the triangular inequality:

$$\left|\sqrt{\mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g))} - \sqrt{\mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g'))}\right| \leq \|G_g - G_{g'}\|_{\mathrm{L}_2([0,1]\times[E],\lambda\otimes p_{\boldsymbol{w}})}.$$

Note that, because $var(X) \leq \mathbb{E}[X^2]$ for any real variable X, we have:

$$\begin{aligned}\sum_{e=1}^{E} w_e \left(G_g(t,e) - G_{g'}(t,e)\right)^2 &= \sum_{e=1}^{E} w_e \left(F^{-1}_{g,e}(t) - F^{-1}_{g',e}(t) - \sum_{e'=1}^{E} w_{e'}\left(F^{-1}_{g,e'}(t) - F^{-1}_{g',e'}(t)\right)\right)^2 \\ &\leq \sum_{e=1}^{E} w_e \left(F^{-1}_{g,e}(t) - F^{-1}_{g',e}(t)\right)^2.\end{aligned}$$

Therefore we get that:

$$\begin{aligned}\left|\sqrt{\mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g))} - \sqrt{\mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g'))}\right|^2 \leq \|G_g - G_{g'}\|^2_{\mathrm{L}_2([0,1]\times[E],\lambda\otimes p_{\boldsymbol{w}})} &\leq \int_0^1 \sum_{e=1}^{E} w_e \left(F^{-1}_{g,e}(t) - F^{-1}_{g',e}(t)\right)^2 dt \\ &\leq \sum_{e=1}^{E} w_e W_2^2(\nu_e(g), \nu_e(g')) \leq \sum_{e=1}^{E} w_e \mathbb{E}\left[\left(g(X^e) - g'(X^e)\right)^2\right] \leq \|g - g'\|^2_\infty,\end{aligned}$$

where in the third inequality (which is an equality) we used the explicit form of the Wasserstein distance for measures defined on $\mathbb{R}$ (see remarks 2.30 in Peyré et al. (2019) and Theorem 2.18 in Villani (2003)); and the fourth inequality is a direct consequence of the definition of the Wasserstein distance. The last inequality concludes our proof. □

Lemma 9. *Fix $\lambda \in (0, E^{-1}]$ and let $\Lambda_\lambda = \left\{\boldsymbol{w} = (w_e)_{e=1}^E : \forall e \in [E], w_e \in [\lambda, 1] \text{ and } \sum_{e=1}^E = 1\right\}$.*

For any class of functions $\mathcal{G}$, if for some $\boldsymbol{w}^0 \in \Lambda$ we have $\Gamma_{\boldsymbol{w}^0}(\mathcal{G}) > 0$ then there exists $\gamma > 0$ such that for any $\boldsymbol{w} \in \Lambda_\lambda$, we have $\Gamma_{\boldsymbol{w}}(\mathcal{G}) \geq \gamma$.

Conversely, if for some $\epsilon > 0$ there exists a $\boldsymbol{w} \in \Lambda_\lambda$ such that $\Gamma_{\boldsymbol{w}}(\mathcal{G}) \leq \epsilon$, then for any $\boldsymbol{w}^0 \in \Lambda$, we have $\Gamma_{\boldsymbol{w}^0}(\mathcal{G}) \leq \epsilon/\lambda$.

Proof. For any $\boldsymbol{w}^0 \in \Lambda$ and $\boldsymbol{w} \in \Lambda_\lambda$ it is easy to see that:

$$\forall g \in \mathcal{G}, \quad \mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g)) \geq \lambda \mathrm{WV}_{\boldsymbol{w}^0}(\boldsymbol{\nu}(g)).$$

Taking the infimum over $\mathcal{G}$ on both sides directly yields the result of this lemma. □

Lemma 10. *Under $\tilde{H}_0(\mathcal{E})$ and Assumption 4, for any $\delta \in (0, 1/2)$ the set $\mathcal{G}_0^\delta$ is a non-empty compact subset of $C_0(\mathcal{X}, ||\cdot||_\infty)$.*

Proof. The fact that $\mathcal{G}_0^\delta$ is compact in $C_0(\mathcal{X}, ||\cdot||_\infty)$ is direct from the continuity of $g \mapsto \mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g))$ proved in Lemma 8 and that $\mathcal{G}^\delta$ is itself compact (see Remark 2). Therefore, we just have to show that it is non-empty.

First, recall from Remark 2 that $W^{2,d}(\mathcal{X})$ is norm equivalent to the RKHS generated by the Matérn kernel $k_{d-p/2,h}$ of degree $d - p/2$ for any scale $h > 0$. Hence, there exists a ball $B'_{R'_\delta}$ of radius R'_δ centered at the origin in this RKHS such that: $\mathcal{G}^\delta \subseteq \overline{B'_{R'_\delta}}$. Therefore, by Lemma 26.10 from Shalev-Shwartz and Ben-David (2014):

$$\mathcal{R}_{n_e}(\mathcal{G}^\delta) \leq \mathcal{R}_{n_e}(\overline{B'_{R'_\delta}}) = \mathcal{R}_{n_e}(B'_{R'_\delta}) \leq \frac{R'_\delta}{\sqrt{n_e}} \sup_{x\in\mathcal{X}} \sqrt{k_{d-p/2,h}(x,x)}.$$

Note that for $\mathcal{G} = \mathcal{G}^\delta$ the variables $(Z_e)_{e=1}^E$ from Theorem 1 are sub-Gaussian since $\mathcal{G}^\delta$ is a bounded subset of $C_0(\mathcal{X}, ||\cdot||_\infty)$ (see Section D.1). Using Theorem 1 and the fact that $w_e = n_e/n$, we have that with probability at least $1 - \delta/2$:

$$\forall g \in \mathcal{G}^\delta, \quad |\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(g)) - \mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g))| \leq c_{\delta/2} \frac{\log^2(n)}{\sqrt{n}}, \tag{55}$$

for some constant $c_{\delta/2}$ that depends only on δ.

We are going to prove that $\mathcal{G}_0^\delta$ is non-empty by contradiction. Assume $\mathcal{G}_0^\delta = \emptyset$. Since $\mathcal{G}^\delta$ is compact and $g \mapsto \mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g))$ is continuous, it means there exists $\gamma > 0$ such that $\forall g \in \mathcal{G}^\delta$ we have $\mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g)) \geq \gamma$ (by Lemma 9 and condition (3) of Assumption 4, this γ can be chosen independently of $\boldsymbol{w}$). Also, since we are under $\tilde{H}_0(\mathcal{E})$ and because the sets $\mathcal{F}'_n$ are non-decreasing and $\mathcal{F}' = \overline{\bigcup_n \mathcal{F}'_n}$, there exist a function $f \in \mathcal{F}'$ and a constant $c_\gamma > 0$ such that:

$$\forall n \geq c_\gamma, \quad f \in \mathcal{F}'_n \ \text{ and } \ \mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(f)) \leq \frac{\gamma}{2},$$

(again by Lemma 9, f can be chosen such that this inequality holds for all $\boldsymbol{w} \in \Lambda_\lambda$).

Using Theorem 1 again, one can prove that with probability at least $1 - \delta/2$:

$$|\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(f)) - \mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(f))| \leq c'_{\delta/2} \frac{\log^2(n)}{\sqrt{n}}, \tag{56}$$

for some constant $c'_{\delta/2}$ that depends on δ.

Combining (55) and (56) we get that there exists a constant $c = c(c_{\delta/2}, c_\gamma, c'_{\delta/2})$ such that for any $n \geq c$, we have with probability at least $1 - \delta$:

$$\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{G}^\delta) > \frac{3}{4}\gamma \quad \text{and} \quad \mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(f)) < \frac{3}{4}\gamma,$$

and therefore with probability at least $1 - \delta$ we must have $\hat{f}_n \notin \mathcal{G}^\delta$, a contradiction with Assumption 4 condition (4) since $\delta < 1/2$. □

Lemma 11. *Under conditions (6a) and (6b) from Assumption 4 we have that, for all $s \in (0, 1/2)$ and $\delta \in (0, 1)$, there exists a constant $C_{\delta,s} > 0$ such that:*

$$\forall f \in \mathcal{G}^\delta, \forall e \in [E], \forall t \in [s, 1-s], \ q_f^e(t) \vee |q_f^{e\prime}(t)| < C_{\delta,s}.$$

Proof. Take $f \in \mathcal{G}^\delta$ and $e \in [E]$, from (6a) we have:

$$\forall t \in [s, 1-s], \quad \frac{s}{2}|q_f^{e\prime}(t)|/q_f^e(t) < M_{\delta,s} \quad \Leftrightarrow \quad |q_f^{e\prime}(t)| < \frac{2M_{\delta,s}}{s} q_f^e(t). \tag{57}$$

Call $m_{s,f}^e \doteq \max_{t \in [s,1-s]} q_f^e(t)$ achieved at some $t_{s,f}^e \in [s, 1-s]$ by continuity. This implies that for any $t \in [s, 1-s]$ we have $|q_f^{e\prime}(t)| < 2M_{\delta,s} m_{s,f}^e$.

Let $\epsilon > 0$ such that $2\epsilon(M_{\delta,s} \vee 1)/(s \wedge (1/2 - s)) = 1/2 \ \Leftrightarrow \ \epsilon = (s \wedge (1/2 - s))/4(M_{\delta,s} \vee 1)$. Then for $t \in [t_{s,f}^e - \epsilon, t_{s,f}^e + \epsilon] \cap [s, 1-s]$ we have:

$$|q_f^e(t) - q_f^e(t_{s,f}^e)| = \left| \int_{t_{s,f}^e}^t q_f^{e\prime}(u) du \right| \leq \epsilon \frac{2M_{\delta,s}}{s} m_{s,f}^e \leq \frac{m_{s,f}^e}{2}.$$

Thus this means that $q_f^e(t) \geq \frac{m_{s,f}^e}{2}$ for $t \in [t_{s,f}^e - \epsilon, t_{s,f}^e + \epsilon] \cap [s, 1-s]$ and that we have from condition (6b) of Assumption 4:

$$M'_{\delta,s} > \int_s^{1-s} t(1-t) q_f^e(t)^2 dt \geq \frac{s}{2}\epsilon \left(\frac{m_{s,f}^e}{2}\right)^2 = \frac{s(s \wedge (1/2-s))}{32(M_{\delta,s} \vee 1)} {m_{s,f}^e}^2.$$

Therefore $m_{s,f}^e \leq \sqrt{\frac{32(M_{\delta,s} \vee 1) M'_{\delta,s}}{s(s \wedge (1/2-s))}}$, a bound which does not depend on either f or e. Combining with inequality (57), we get our result. □

Lemma 12. *Let $\delta \in (0,1)$. For any $e \in [E]$ and functions g_1 and g_2 in $\mathcal{G}^\delta$, define for short q_1^e and q_2^e the respective quantile densities of $\nu_e(g_1)$ and $\nu_e(g_2)$. Under Assumption 4, we have for $s \in (0,1/2)$:*

$$\int_s^{1-s} (q_1^e(t) - q_2^e(t))^2 dt \leq A_{\delta,s} \left(\|g_1 - g_2\|_\infty^{2/3} \vee \|g_1 - g_2\|_\infty^2 \right),$$

where $A_{\delta,s} \doteq \frac{10 C_{\delta,s}^2}{\pi^2}(1 - 3s + 2s^2) + \frac{108(1+\pi^2)}{1-2s}$ and $C_{\delta,s}$ is the constant derived in Lemma 11.

Proof. Set $s \in (0,1/2)$. Under Assumption 4, by Lemma 11 we have that there exists a constant $C_{\delta,s}$ such that:

$$\forall g \in \mathcal{G}^\delta, \forall e \in [E], \forall t \in [s, 1-s],\ q_g^e(t) \vee |q_g^{e\prime}(t)| < C_{\delta,s}. \tag{58}$$

Let $g_1, g_2 \in \mathcal{G}^\delta$ and $e \in [E]$. Call $F_{1,e}^{-1}$ and $F_{2,e}^{-1}$ the quantile functions of $\nu_e(g_1)$ and $\nu_e(g_2)$ respectively. We have by the definition of the infinite Wasserstein distance W_∞ (see section 5.5.1 from Santambrogio (2015) for its definition):

$$\|F_{1,e}^{-1} - F_{2,e}^{-1}\|_\infty = W_\infty(\nu_e(g_1), \nu_e(g_2)) \leq \text{ess sup}|Y^e - g_1(X^e) - Y^e + g_2(X^e)| \leq \epsilon \doteq \|g_1 - g_2\|_\infty,$$

where "ess sup" refers to the essential supremum. Therefore, in particular for any $t \in [s, 1-s]$, $|F_{1,e}^{-1}(t) - F_{2,e}^{-1}(t)| \leq \epsilon$. Now consider the sequence of Fourier bases in $\mathrm{L}_2([s,1-s])$:

$$\forall k \in \mathbb{Z}, \quad \phi_k(t) \doteq \frac{e^{i2\pi k(t-1/2)/(1-2s)}}{\sqrt{1-2s}},$$

and the Fourier coefficients of $F_{j,e}^{-1}$ for $j \in \{1,2\}$ are:

$$\forall k \in \mathbb{Z}, \quad \alpha_{j,e}^k \doteq \int_s^{1-s} F_{j,e}^{-1}(t)\overline{\phi_k(t)}dt.$$

Next, notice that because of (58) we have that q_1^e and q_2^e are in $\mathrm{L}_2([s,1-s])$; by integration by parts we can express the Fourier coefficients of q_j^e for $j \in \{1,2\}$ as follows:

$$\begin{aligned}
\forall k \in \mathbb{Z}, \quad \beta_{j,e}^k &\doteq \int_s^{1-s} q_j^e(t)\overline{\phi_k(t)}dt \\
&= \left[F_{j,e}^{-1}(t)\overline{\phi_k(t)}\right]_s^{1-s} + \frac{i2\pi k}{1-2s} \cdot \int_s^{1-s} F_{j,e}^{-1}(t)\overline{\phi_k(t)}dt \\
&= \frac{(-1)^k}{\sqrt{1-2s}} \cdot \left(F_{j,e}^{-1}(1-s) - F_{j,e}^{-1}(s)\right) + \frac{i2\pi k}{1-2s} \cdot \alpha_{j,e}^k.
\end{aligned}$$

Therefore, by Parseval's formula we have for any $K \in \mathbb{N}$:

$$\begin{aligned}
\int_s^{1-s} (q_1^e(t) - q_2^e(t))^2 dt &= \sum_{k\in\mathbb{Z}} |\beta_{1,e}^k - \beta_{2,e}^k|^2 = \sum_{|k|\geq K} |\beta_{1,e}^k - \beta_{2,e}^k|^2 \\
&+ 3 \sum_{|k|<K} \left[\frac{\left(F_{1,e}^{-1}(s) - F_{2,e}^{-1}(s)\right)^2}{1-2s} + \frac{\left(F_{1,e}^{-1}(1-s) - F_{2,e}^{-1}(1-s)\right)^2}{1-2s} + \left(\frac{2\pi k}{1-2s}\right)^2 \left|\alpha_{1,e}^k - \alpha_{2,e}^k\right|^2 \right] \\
&\leq \sum_{|k|\geq K} |\beta_{1,e}^k - \beta_{2,e}^k|^2 + \frac{12K\epsilon^2}{1-2s} + 3\left(\frac{2\pi K}{1-2s}\right)^2 \cdot \int_s^{1-s} \left(F_{1,e}^{-1}(t) - F_{2,e}^{-1}(t)\right)^2 dt \\
&\leq \sum_{|k|\geq K} |\beta_{1,e}^k - \beta_{2,e}^k|^2 + \frac{12(1+\pi^2)K^2\epsilon^2}{1-2s}.
\end{aligned} \tag{59}$$

What's left is to bound $\sum_{|k|\geq K} |\beta_{1,e}^k - \beta_{2,e}^k|^2$. Now consider the derivative $q_j^{e\prime}$ of q_j^e for $j \in \{1,2\}$, which is also in $\mathrm{L}_2([s,1-s])$ because of (58). It has the following Fourier coefficients:

$$\begin{aligned}\forall k \in \mathbb{Z}, \quad \gamma_{j,e}^k &\doteq \int_s^{1-s} q_j^{e\prime}(t)\overline{\phi_k(t)}dt \\ &= \left[q_j^e(t)\overline{\phi_k(t)}\right]_s^{1-s} + \frac{i2\pi k}{1-2s} \cdot \int_s^{1-s} q_j^e(t)\overline{\phi_k(t)}dt \\ &= \frac{(-1)^k}{\sqrt{1-2s}} \cdot \left(q_j^e(1-s) - q_j^e(s)\right) + \frac{i2\pi k}{1-2s} \cdot \beta_{j,e}^k.\end{aligned}$$

Hence, using again bound (58) we obtain for any $k \in \mathbb{Z}$:

$$\begin{aligned}|\beta_{1,e}^k - \beta_{2,e}^k|^2 &= \left(\frac{1-2s}{2\pi k}\right)^2 \left|\gamma_{1,e}^k - \gamma_{2,e}^k + \frac{(-1)^{k+1}}{\sqrt{1-2s}}\left(q_1^e(1-s) - q_2^e(1-s) + q_1^e(s) - q_2^e(s)\right)\right|^2 \\ &\leq 5\left(\frac{1-2s}{2\pi k}\right)^2 \left(|\gamma_{1,e}^k - \gamma_{2,e}^k|^2 + \frac{4C_{\delta,s}^2}{1-2s}\right) \\ &\leq 5\left(\frac{1-2s}{2\pi k}\right)^2 \left(2\int_s^{1-s} |q_1^{e\prime}(t)|^2 dt + 2\int_s^{1-s} |q_2^{e\prime}(t)|^2 dt + \frac{4C_{\delta,s}^2}{1-2s}\right) \\ &\leq 5\left(\frac{1-2s}{2\pi k}\right)^2 \left(4C_{\delta,s}^2 + \frac{4C_{\delta,s}^2}{1-2s}\right) \\ &\leq 20\left(\frac{1-2s}{2\pi k}\right)^2 C_{\delta,s}^2 \frac{1-s}{1-2s} = \frac{5}{\pi^2 k^2}(1-3s+2s^2)C_{\delta,s}^2.\end{aligned}$$

From inequality (59) we thus get:

$$\int_s^{1-s} (q_1^e(t) - q_2^e(t))^2 dt \leq \frac{10C_{\delta,s}^2}{\pi^2}(1-3s+2s^2) \cdot \sum_{k\geq K} \frac{1}{k^2} + \frac{12(1+\pi^2)K^2\epsilon^2}{1-2s}.$$

This means that for any $x \geq 1$, we have:

$$\begin{aligned}\int_s^{1-s} (q_1^e(t) - q_2^e(t))^2 dt &\leq \frac{10C_{\delta,s}^2}{\pi^2}(1-3s+2s^2)\int_x^\infty \frac{1}{u^2}du + \frac{12(1+\pi^2)\epsilon^2}{1-2s}(x+2)^2 \\ &\leq \frac{10C_{\delta,s}^2}{\pi^2}(1-3s+2s^2)\frac{1}{x} + \frac{108(1+\pi^2)\epsilon^2}{1-2s}x^2.\end{aligned}$$

If $\epsilon \leq 1$, we set $x = \epsilon^{-2/3}$ and we get:

$$\int_s^{1-s} (q_1^e(t) - q_2^e(t))^2 dt \leq \left(\frac{10C_{\delta,s}^2}{\pi^2}(1-3s+2s^2) + \frac{108(1+\pi^2)}{1-2s}\right)\epsilon^{2/3}.$$

Otherwise, if $\epsilon > 1$ we set $x = 1$ and we finally obtain:

$$\int_s^{1-s} (q_1^e(t) - q_2^e(t))^2 dt \leq \left(\frac{10C_{\delta,s}^2}{\pi^2}(1-3s+2s^2) + \frac{108(1+\pi^2)}{1-2s}\right)\epsilon^2.$$

□

Lemma 13. *Consider X, Y two real-valued variables. Let $\alpha, \alpha_1, \alpha_2 \in (0,1)$ such that $\alpha = \alpha_1 + \alpha_2$. If we define $q_{1-\alpha}(Z)$ as the $(1-\alpha)$-quantile of any real-valued variable Z, we have:*

$$q_{1-\alpha}(X+Y) \leq q_{1-\alpha_1}(X) + q_{1-\alpha_2}(Y).$$

Proof. For any $a, b \in \mathbb{R}$, we have $\mathbb{P}(X + Y > a + b) \leq \mathbb{P}(X > a) + \mathbb{P}(Y > b)$. If we set $a = q_{1-\alpha_1}(X)$ and $b = q_{1-\alpha_2}(Y)$, we thus get that $\mathbb{P}(X + Y > a + b) \leq \alpha_1 + \alpha_2 = \alpha$. Therefore, $q_{1-\alpha}(X + Y) \leq a + b$. □

Lemma 14. *Let $s \in [0, 1/2)$ and for any quantile density function q (potentially empirical) that satisfies condition (iii) from Assumption 3 we define the following variable:*

$$T_s(q) \doteq \sum_{e=1}^{E-1} \int_s^{1-s} B_e^2(t) q^2(t) dt,$$

where $(B_e(t))_{e=1}^{E-1}$ are $E-1$ independent Brownian bridges. Call also $q_{1-\alpha}(T_s(q))$ the $(1-\alpha)$-quantile of $T_s(q)$. For any quantile densities q_1, q_2 that satisfy (iii) of Assumption 3, we have:

$$\left(\mathbb{E}\left[\left|\sqrt{T_s(q_1)} - \sqrt{T_s(q_2)}\right|\right]\right)^2 \leq (E-1) \cdot \int_s^{1-s} t(1-t)(q_1(t) - q_2(t))^2 dt.$$

And for any $0 < \alpha < \alpha' < 1$, we have:

$$\sqrt{q_{1-\alpha}(T_s(q_1))} \geq \sqrt{q_{1-\alpha'}(T_s(q_2))} - \frac{(E-1)^{1/2}}{\alpha' - \alpha} \left(\int_s^{1-s} t(1-t)(q_1(t) - q_2(t))^2 dt\right)^{1/2}.$$

Proof. Take $s \in [0, 1/2)$ and q_1, q_2 two quantile densities that satisfy condition (iii) from Assumption 3. Notice that, for $i \in \{1, 2\}$, we can rewrite $T_s(q_i)$ as follows:

$$T_s(q_i) \doteq \sum_{e=1}^{E-1} \int_s^{1-s} B_e^2(t) q_i^2(t) dt = \|B_e(t) q_i(t)\|^2_{\mathrm{L}_2([E-1]\times[s,1-s],\, \mu\otimes\lambda)},$$

where $B_e(t)q_i(t)$ is treated as a function of both e and t, and we denote the counting by μ and the Lebesgue measure by λ. Therefore, the triangular inequality applies:

$$\left|\sqrt{T_s(q_1)} - \sqrt{T_s(q_2)}\right| \leq \sqrt{T_s(q_1 - q_2)}.$$

Using this fact and Jensen's inequality we get:

$$\begin{aligned}
\left(\mathbb{E}\left[\left|\sqrt{T_s(q_1)} - \sqrt{T_s(q_2)}\right|\right]\right)^2 &\leq \left(\mathbb{E}\left[\sqrt{T_s(q_1 - q_2)}\right]\right)^2 \\
&\leq \mathbb{E}\left[\sum_{e=1}^{E-1} \int_s^{1-s} B_e^2(t)(q_1(t) - q_2(t))^2 dt\right] \\
&\leq (E-1) \int_s^{1-s} t(1-t)(q_1(t) - q_2(t))^2 dt.
\end{aligned}$$

Now let's turn to the lower bound for $q_{1-\alpha}(T_s(q_1))$. Let $0 < \alpha < \alpha' < 1$, by Lemma 13 we have that:

$$\begin{aligned}
\sqrt{q_{1-\alpha'}(T_s(q_2))} = q_{1-\alpha'}\left(\sqrt{T_s(q_2)}\right) &= q_{1-\alpha'}\left(\sqrt{T_s(q_2)} - \sqrt{T_s(q_1)} + \sqrt{T_s(q_1)}\right) \\
&\leq q_{1-\alpha}\left(\sqrt{T_s(q_1)}\right) + q_{1-(\alpha'-\alpha)}\left(\sqrt{T_s(q_2)} - \sqrt{T_s(q_1)}\right) \\
&\leq \sqrt{q_{1-\alpha}\left(T_s(q_1)\right)} + q_{1-(\alpha'-\alpha)}\left(\left|\sqrt{T_s(q_2)} - \sqrt{T_s(q_1)}\right|\right).
\end{aligned}$$

Furthermore, it is easy to see that:

$$\begin{aligned}
q_{1-(\alpha'-\alpha)}\left(\left|\sqrt{T_s(q_2)} - \sqrt{T_s(q_1)}\right|\right) &\leq \frac{\mathbb{E}\left[\left|\sqrt{T_s(q_1)} - \sqrt{T_s(q_2)}\right|\right]}{\alpha' - \alpha} \\
&\leq \frac{(E-1)^{1/2}}{\alpha' - \alpha} \left(\int_s^{1-s} t(1-t)(q_1(t) - q_2(t))^2 dt\right)^{1/2}.
\end{aligned}$$

□

Lemma 15. *Assume Assumption 4 and fix $\delta \in (0, 1/2)$. For any $s \in [0, 1/2)$ and quantile density q that satisfies condition (iii) from Assumption 3, we let $T_s(q)$ be the variable introduced in Lemma 14. For any $\alpha \in (0, 1)$, define:*

$$t^\delta_{s,\alpha} \doteq \inf_{f \in \mathcal{G}^\delta_0} \left(q_{1-\alpha}(T_s(q_f))\right),$$

where q_f is the quantile density of $\nu_e(f)$ – note that since $f \in \mathcal{G}^\delta_0$, the $\nu_e(f)$'s are actually identical. Then, for any $s \in [0, 1/2)$ and $\alpha \in (0, 1)$, the infimum for $t^\delta_{s,\alpha}$ is attained by a function in $\mathcal{G}^\delta_0$, that is:

$$\forall s \in [0, 1/2), \forall \alpha \in (0, 1), \quad \exists f \in \mathcal{G}^\delta_0 \quad s.t. \quad q_{1-\alpha}(T_s(q_f)) = t^\delta_{s,\alpha}.$$

Finally, for any $\alpha \in (0, 1)$ we also have:

$$\forall \epsilon > 0, \quad \exists (\alpha', s) \in (\alpha, 1) \times (0, 1/2) \quad s.t. \quad t^\delta_{s,\alpha'} + \epsilon \geq t^\delta_{0,\alpha}.$$

Proof. Recall that by Lemma 10, $\mathcal{G}^\delta_0$ is a non-empty compact subset of $C_0(\mathcal{X}, ||\cdot||_\infty)$. Let $\alpha \in (0, 1)$ and first consider the case where $s \in (0, 1/2)$ – that is $s > 0$. Notice that the function:

$$Q_{s,\alpha} : \quad f \in \left(\mathcal{G}^\delta_0, \|\cdot\|_\infty\right) \mapsto q_{1-\alpha}(T_s(q_f))$$

is in fact continuous when $s > 0$. Indeed, let $f \in \mathcal{G}^\delta_0$ and $(f_k)_{k\geq 1} \in \mathcal{G}^\delta_0$ such that $\|f - f_k\|_\infty \to_{k\to\infty} 0$, then by Lemma 12 and Lemma 14 we have:

$$\begin{aligned}\left(\mathbb{E}\left[\left|\sqrt{T_s(q_f)} - \sqrt{T_s(q_{f_k})}\right|\right]\right)^2 &\leq (E-1)\int_s^{1-s} \left(q_f(t) - q_{f_k}(t)\right)^2 dt \\ &\leq (E-1)\cdot A_{\delta,s}\left(\|f - f_k\|^{2/3}_\infty \vee \|f - f_k\|^2_\infty\right) \xrightarrow[k\to\infty]{} 0.\end{aligned}$$

This proves at least that $T_s(q_{f_k})$ converges in probability (thus in distribution too) toward $T_s(q_f)$ as k goes to infinity. Therefore, since $T_s(q_f)$ is a continuous variable:

$$q_{1-\alpha}(T_s(q_{f_k})) \xrightarrow[k\to\infty]{} q_{1-\alpha}(T_s(q_f)).$$

Hence, the above map $Q_{s,\alpha}$ is continuous when $s > 0$. Since $\mathcal{G}^\delta_0$ is compact, it implies that $t^\delta_{s,\alpha}$ is attained by some function $f^\delta_{s,\alpha}$ in $\mathcal{G}^\delta_0$ for $s > 0$.

Now, let's turn to the case where $s = 0$. Let $(s_k)_{k\geq 1} \in (0, 1/2)$ be a decreasing sequence that converges to 0, and for short let $f_k \doteq f^\delta_{s_k,\alpha}$ the minimizer for $t^\delta_{s_k,\alpha}$. Since $\mathcal{G}^\delta_0$ is compact, up to extraction we can consider that $(f_k)_{k\geq 1}$ converges w.r.t. $\|\cdot\|_\infty$ to a function $f^\delta_{0,\alpha}$ in $\mathcal{G}^\delta_0$. We are going to show that $f^\delta_{0,\alpha}$ is indeed a minimizer for $t^\delta_{0,\alpha}$. For simplicity, call $q_k \doteq q_{f_k}$ and $q_0 \doteq q_{f^\delta_{0,\alpha}}$ and note that $T_{s_k}(q_k) = T_0(q_k \cdot \mathbb{1}_{\cdot\in[s_k, 1-s_k]})$.

Fix $K \geq 1$, and by Lemma 14 for any $k \geq K$ we have:

$$\begin{aligned}\left(\mathbb{E}\left[\left|\sqrt{T_{s_k}(q_k)} - \sqrt{T_0(q_0)}\right|\right]\right)^2 &\leq (E-1)\int_0^1 (1-t)t\left(q_k(t)\mathbb{1}_{t\in[s_k,1-s_k]} - q_0(t)\right)^2 dt \\ \leq 2(E-1)&\left[\int_0^{s_K} t(1-t)\left(q^2_k(t) + q^2_0(t)\right)dt + \int_{1-s_K}^1 t(1-t)\left(q^2_k(t) + q^2_0(t)\right)dt\right] \\ &+ (E-1)\int_{s_K}^{1-s_K} (q_k(t) - q_0(t))^2 dt.\end{aligned}$$

By condition (6c) of Assumption 4 we have that for any $\epsilon > 0$ we can choose K large enough such that, for any $k \geq K$:

$$2(E-1)\left[\int_0^{s_K} t(1-t)\left(q^2_k(t) + q^2_0(t)\right)dt + \int_{1-s_K}^1 t(1-t)\left(q^2_k(t) + q^2_0(t)\right)dt\right] \leq \epsilon.$$

Also notice that for a fixed K, by Lemma 12 the term $\int_{s_K}^{1-s_K}(q_k(t) - q_0(t))^2 dt$ goes to 0 as $k \to \infty$. It means that:

$$\forall \epsilon > 0, \quad \limsup_{k\to\infty}\left(\mathbb{E}\left[\left|\sqrt{T_{s_k}(q_k)} - \sqrt{T_0(q_0)}\right|\right]\right)^2 \leq \epsilon.$$

Because ϵ is chosen arbitrarily, this implies that $T_{s_k}(q_k)$ converges in distribution toward $T_0(q_0)$. Since $\forall k, t^\delta_{s_k,\alpha} \leq t^\delta_{0,\alpha}$, we then have:

$$t^\delta_{0,\alpha} \leq q_{1-\alpha}(T_0(q_0)) = \lim_{k\to\infty} q_{1-\alpha}(T_{s_k}(q_k)) = \lim_{k\to\infty} t^\delta_{s_k,\alpha} \leq t^\delta_{0,\alpha}.$$

Hence, $f^\delta_{0,\alpha}$ is a minimizer for $t^\delta_{0,\alpha}$. Note we also proved that, for a fix $\alpha \in (0,1)$:

$$\forall \epsilon > 0, \quad \exists s \in (0, 1/2) \quad \text{s.t.} \quad t^\delta_{0,\alpha} \leq t^\delta_{s,\alpha} + \frac{\epsilon}{2}.$$

To prove the last point we just need to show that we can also find $\alpha' > \alpha$ such that $t^\delta_{s,\alpha} \leq t^\delta_{s,\alpha'} + \frac{\epsilon}{2}$. We are going to use the same approach as above: Let $(\alpha_k)_{k\geq 1}$ be a decreasing sequence in $(0,1)$ that converges to α (meaning in particular that $\alpha_k \geq \alpha$ for all k); call f_k the minimizer for t^δ_{s,α_k}. By compactness of $\mathcal{G}^\delta_0$, up to extraction we can consider that f_k converges to some $f_0 \in \mathcal{G}^\delta_0$. Let $K \geq 1$ and take any $k \geq K$, by continuity of Q_{s,α_K} we have:

$$t^\delta_{s,\alpha} \geq t^\delta_{s,\alpha_k} = Q_{s,\alpha_k}(f_k) \geq Q_{s,\alpha_K}(f_k) \xrightarrow[k\to\infty]{} Q_{s,\alpha_K}(f_0) = q_{1-\alpha_K}(T_s(q_{f_0})) \xrightarrow[K\to\infty]{} q_{1-\alpha}(T_s(q_{f_0})) \geq t^\delta_{s,\alpha}.$$

This proves that $t^\delta_{s,\alpha_k} \xrightarrow[k\to\infty]{} t^\delta_{s,\alpha}$ and hence that for k large enough $t^\delta_{s,\alpha} \leq t^\delta_{s,\alpha_k} + \frac{\epsilon}{2}$. □

Lemma 16. *Assume Assumption 4 and let $\delta \in (0, 1/2)$. For any $\epsilon > 0$, there exists a constant $c_{\delta,\epsilon}$ such that for any $n \geq c_{\delta,\epsilon}$, we have with probability at least $1 - 2\delta$ that simultaneously:*

$$\hat{f}_n \in \mathcal{G}^\delta \quad \text{and} \quad \exists f \in \mathcal{G}^\delta_0 \quad \text{s.t.} \quad \|\hat{f}_n - f\|_\infty < \epsilon.$$

Proof. Set $\epsilon > 0$. We know already that, under Assumption 4, when $n \geq c_\delta$ for some constant c_δ, with probability at least $1 - \delta$ we have $\hat{f}_n \in \mathcal{G}^\delta$. Also, based on the proof of Lemma 10, there is a constant C_δ such that with probability at least $1 - \delta$:

$$\forall g \in \mathcal{G}^\delta, \quad |\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(g)) - \mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g))| \leq C_\delta \frac{\log^2(n)}{\sqrt{n}}. \tag{60}$$

Recall that by Lemma 10, $\mathcal{G}^\delta_0$ is a non-empty compact subset of $\mathcal{G}^\delta$. Consider $\mathcal{G}^\delta_\epsilon \doteq \{g \in \mathcal{G}^\delta : \ d(g, \mathcal{G}^\delta_0) \geq \epsilon\}$. $\mathcal{G}^\delta_\epsilon$ is a closed subset of $\mathcal{G}^\delta$, which is compact, hence $\mathcal{G}^\delta_\epsilon$ is also compact. By the continuity of the Wasserstein variance w.r.t. g implied by Lemma 8 we can find a constant $\gamma > 0$ such that for any $g \in \mathcal{G}^\delta_\epsilon$, we have $\mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g)) \geq \gamma$. Note that Lemma 9 tells us that γ can be chosen independently of $\boldsymbol{w}$ under Assumption 4. Therefore, there is a constant $c_{\delta,\epsilon} > 0$ such that $\forall n \geq c_{\delta,\epsilon}$, under the event of equation (60) we have:

$$\forall g \in \mathcal{G}^\delta_0, \quad \mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(g)) < \gamma/2 \quad \text{and} \quad \forall g \in \mathcal{G}^\delta_\epsilon, \quad \mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(g)) > \gamma/2.$$

Intersecting with the event that $\hat{f}_n \in \mathcal{G}^\delta$, and using a union bound, we get that whenever $n \geq \max(c_\delta, c_{\delta,\epsilon})$ with probability at least $1 - 2\delta$, $\hat{f}_n$ is in $\mathcal{G}^\delta$ but cannot be in $\mathcal{G}^\delta_\epsilon$. This means that $d(\hat{f}_n, \mathcal{G}^\delta_0) < \epsilon$. □

Lemma 17. *Let $f, g \in C_0(\mathcal{X}, ||\cdot||_\infty)$ and $e \in [E]$. Consider $\hat{q}^e_f$ and $\hat{q}^e_g$ the kernel quantile density estimators at respectively f and g as defined from Definition 4, where the kernel K is L-Lipschitz. We also set the bandwidth at $h_e = \beta n_e^{-1/3}$ for some constant $\beta > 0$. For any $s \in (0, 1/2)$, as long as $n_e > \frac{\beta+1}{s} \vee \beta^{-1/2}$ we have:*

$$\forall t \in [s, 1-s], \quad \left|\hat{q}^e_f(t) - \hat{q}^e_g(t)\right| \leq \frac{4L}{h_e}\|f - g\|_\infty.$$

Proof. First, since K is supported on $[-1, 1]$, notice that as long as $n_e^{1/3} > (\beta+1)/s$, we have $K_{h_e}(t-(n_e-1)/n_e) = K_{h_e}(t - 1/n_e) = 0$ for $t \in [s, 1-s]$. Therefore, we can rewrite $\hat{q}^e_f(t)$ (and $\hat{q}^e_g(t)$ similarly) as follows:

$$\forall t \in [s, 1-s], \quad \hat{q}^e_f(t) = \sum_{i=2}^{n_e-1} \epsilon^e_{(i)}(f) \cdot \left(K_{h_e}\left(t - \frac{i-1}{n_e}\right) - K_{h_e}\left(t - \frac{i}{n_e}\right)\right).$$

Because K is supported on $[-1,1]$ and is L-Lipschitz, we get that:

$$\begin{aligned}\forall t \in [s, 1-s], \quad \left|\hat{q}^e_f(t) - \hat{q}^e_g(t)\right| &\leq \sum_{i=2}^{n_e-1} \left|\epsilon^e_{(i)}(f) - \epsilon^e_{(i)}(g)\right| \cdot \left|K_{h_e}\left(t - \frac{i-1}{n_e}\right) - K_{h_e}\left(t - \frac{i}{n_e}\right)\right| \\ &\leq \sum_{i=\lceil (t-h_e)n_e \rceil}^{\lfloor (t+h_e)n_e \rfloor + 1} \left|\epsilon^e_{(i)}(f) - \epsilon^e_{(i)}(g)\right| \cdot \left|K_{h_e}\left(t - \frac{i-1}{n_e}\right) - K_{h_e}\left(t - \frac{i}{n_e}\right)\right| \\ &\leq \frac{2n_e h_e + 2}{n_e h_e^2} L \max_{i \in [n_e]} \left|\epsilon^e_{(i)}(f) - \epsilon^e_{(i)}(g)\right| \\ &\leq \frac{4L}{h_e} \max_{i \in [n_e]} \left|\epsilon^e_{(i)}(f) - \epsilon^e_{(i)}(g)\right|,\end{aligned}$$

where we used in the last inequality the fact that $n_e^{1/3} > \beta^{-1/2} \Leftrightarrow n_e h_e > 1$. Call $\epsilon^e_i(f) \doteq y^e_i - f(x^e_i)$ and $\epsilon^e_i(g) \doteq y^e_i - g(x^e_i)$ the unordered residuals. Finally, we have:

$$\max_{i \in [n_e]} \left|\epsilon^e_{(i)}(f) - \epsilon^e_{(i)}(g)\right| = W_\infty(\hat{\nu}(f), \hat{\nu}(g)) \leq \max_{i \in [n_e]} |\epsilon^e_i(f) - \epsilon^e_i(g)| \leq \|f - g\|_\infty,$$

where the definition of W_∞ can be found in Section 5.5.1 of Santambrogio (2015). This concludes the proof. □

Lemma 18. *Assume that for any $\mathcal{F}' \in \{\mathcal{F}_{-k},\ k \in [p]\}$, Assumption 4 holds with $\mathcal{F}'_n = \mathcal{F}', \forall n$ (as it is summarized in Assumption 2). Then under Assumption 1, if f^* is the unique function in $\overline{\mathcal{F}}$ such that $WV_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(f^*)) = 0$ then $\tilde{S}(\mathcal{E}) = S^*$.*

Proof. The fact that $\Gamma_{\boldsymbol{w}}(\mathcal{F}_{-k}) = 0$ for $k \notin S^*$ is obvious from Assumption 1, so we just need to prove that $\Gamma_{\boldsymbol{w}}(\mathcal{F}_{-k}) > 0$ for $k \in S^*$. Let's do it by contradiction, and assume there is a $k \in S^*$ such that $\Gamma_{\boldsymbol{w}}(\mathcal{F}_{-k}) = 0$. Let $\delta \in (0, 1/2)$, and $\mathcal{G}^\delta_k$ the corresponding set from Assumption 4 for $\mathcal{F}' = \mathcal{F}_{-k}$ and $\hat{f}^k_n$ the minimizer of $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_{-k})$.

Notice that because f^* is the unique function in $\overline{\mathcal{F}}$ such that $\mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(f^*)) = 0$, it implies that $\mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g)) > 0$ for any $g \in \mathcal{G}^\delta_k$. Then, based on the proof of Lemma 10 we can conclude that there is a constant $\gamma > 0$ (independent of $\boldsymbol{w}$) such that for any $g \in \mathcal{G}^\delta_k$, $\mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(g)) \geq \gamma$ and, since $\Gamma_{\boldsymbol{w}}(\mathcal{F}_{-k}) = 0$, there exits a function $f \in \mathcal{F}_{-k}$ such that $\mathrm{WV}_{\boldsymbol{w}}(\boldsymbol{\nu}(f)) \leq \gamma/2$. Furthermore, using equations (55) and (56) from the proof of Lemma 10, we get that for n large enough, with probability at least $1 - \delta$, $\hat{f}^k_n$ must be outside of $\mathcal{G}^\delta_k$. A contradiction with condition (4) of Assumption 4. □

F ADDITIONAL IMPLEMENTATION DETAILS

F.1 Optimization Details

For any function f_θ parametrized by $\theta \in \Theta$ and any $e \in [E]$, denote $(\epsilon^e_{(i)}(f_\theta))_{i=1}^{n_e}$ as the residuals obtained with f_θ in environment e sorted in increasing order as in Definition 4. Moreover, define $\Pi \doteq \{i/n_e : e \in [E],\ i \in [n_e]\}$. We can also rewrite this set as $\Pi = \{\pi_1, \ldots, \pi_L\}$ where $L \doteq |\Pi|$ and the π_ℓ's are the elements of Π sorted in increasing order; we also set $\pi_0 \doteq 0$. Furthermore note that for $e \in [E]$, the quantile function of $\hat{\nu}_e(f_\theta)$ can be written as:

$$F^{-1}_e(t) = \epsilon^e_{(1)}(f_\theta) \mathbb{1}\left\{t \in \left[0,\ \frac{i}{n_e}\right]\right\} + \sum_{i=2}^{n_e} \epsilon^e_{(i)}(f_\theta) \mathbb{1}\left\{t \in \left(\frac{i-1}{n_e},\ \frac{i}{n_e}\right]\right\}.$$

Therefore, for the empirical distributions $\hat{\boldsymbol{\nu}}(f_\theta)$ the closed form of the Wasserstein variance from Equation (11) becomes:

$$\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(f_\theta)) = \sum_{\ell=1}^{L} \left(\sum_{e=1}^{E} w_e \left(\epsilon^e_{(\lceil \pi_\ell n_e \rceil)}(f_\theta) - \sum_{e'=1}^{E} w_{e'} \epsilon^{e'}_{(\lceil \pi_\ell n_{e'} \rceil)}(f_\theta)\right)^2\right) \cdot (\pi_\ell - \pi_{\ell-1}). \tag{61}$$

Let's call $j = j(\theta; e, \ell)$ the index of the observation in environment e that corresponds to the $\lceil \pi_\ell n_e \rceil$th ordered residual $\epsilon^e_{(\lceil \pi_\ell n_e \rceil)}(f_\theta)$, that is $\epsilon^e_{(\lceil \pi_\ell n_e \rceil)}(f_\theta) = y^e_j - f_\theta(x^e_j)$. If in a neighborhood of θ the order of the residuals doesn't change, i.e. j is locally independent of θ in that neighborhood, then $\epsilon^e_{(\lceil \pi_\ell n_e \rceil)}(f_\theta)$ is differentiable w.r.t. θ

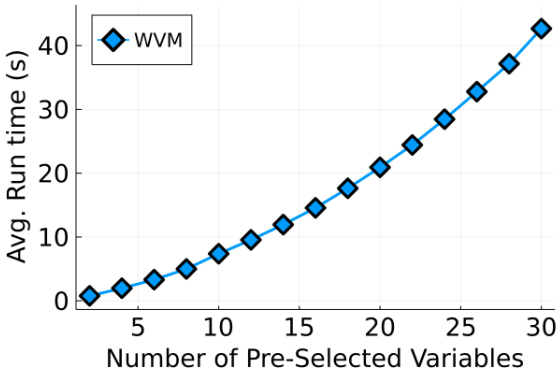


Figure 6: Average run time of WVM in seconds over the 100 simulations from Section 6 for different number of pre-selected variables. This is the same data from Figure 2a, but displayed without the log-scale axis.

at that point, and we have $\nabla_\theta \epsilon^e_{(\lceil \pi_\ell n_e \rceil)}(f_\theta) = -\nabla_\theta f_\theta(x^e_j)$. Hence, whenever the order of the residuals in each environment remain unchanged in a neighborhood of θ, then the gradient of (61) at this θ is:

$$\nabla_\theta \mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(f_\theta)) = \sum_{e=1}^{E}\sum_{\ell=1}^{L} 2(\pi_\ell - \pi_{\ell-1}) w_e \nabla_\theta f_\theta\left(x^e_{j(\theta;e,\ell)}\right)\left(\sum_{e'=1}^{E} w_{e'} \epsilon^{e'}_{(\lceil \pi_\ell n_{e'} \rceil)}(f_\theta) - \epsilon^e_{(\lceil \pi_\ell n_e \rceil)}(f_\theta)\right).$$

At points where $\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}}(f_\theta))$ is not differentiable, the above expression is still a supergradient. Note also that this gradient can be computed efficiently; it requires only sorting and matrix multiplications. Finally, the optimization we use in our experiments is L-BFGS with fixed memory $m = 50$.

Remark 5. *The problem of minimizing the Wasserstein variance to compute the statistics $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_{-k})$ is in general non-convex. However, we haven't found in our experiments any examples where the optimization reached a "bad" local minimum. We believe that finding the minimal Wasserstein variance values like $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_{-k})$ shouldn't be hard in general, at least for classes of linear functions; but of course a deeper study of the landscape of the Wasserstein variance needs to be done to be able to answer this question formally. Furthermore, the optimization procedure behind each of the p tests WVM needs to perform also scales with the number of predictors. For example, L-BFGS scales linearly in the number of predictors yielding an overall complexity for WVM to be $\mathcal{O}(p^2)$ – see Figure 6. As we saw in Section 6, compared to the exponential scaling of ICP the quadratic scaling of WVM is modest.*

F.2 Approximation of the Asymptotic Distribution

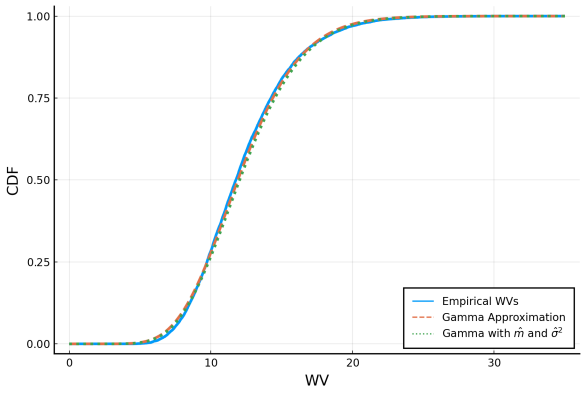


Figure 7: Comparison of CDFs of the Wasserstein Variance and its approximations with a Gamma distribution.

To show that the Gamma distribution can well approximate the asymptotic distribution of the Wasserstein Variance under $\tilde{H}_0(\mathcal{E})$ from (8) and (10), we plot in Figure 7 the empirical CDF for $\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}})$ compared to a Gamma distribution with the same mean and variance as well as the Gamma approximation introduced in Proposition 2. More precisely, we sampled in each environment ($E = 5$) 500 i.i.d. $\mathcal{N}(0, 1)$ observations, so that

the distributions across environments are identical; the resulting empirical distributions are denoted by $\hat{\boldsymbol{\nu}}$. We repeat this process $10,000$ times and compute $\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}})$ at each iteration (we set the weights at $1/E$) to generate the empirical CDF of $\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}})$ under that setting ("Empirical WVs" in Figure 7). Moreover, we compute the empirical mean and variance of $\mathrm{WV}_{\boldsymbol{w}}(\hat{\boldsymbol{\nu}})$ over these $10,000$ simulations and consider a Gamma distribution with the same mean and variance ("Gamma with Empirical Params" in Figure 7). Lastly, we use the approximation with the kernel quantile density estimators proposed in Proposition 2 ("Quantile KDE" in Figure 7), increasing the sample size to $5,000$ samples for each environment to make sure the kernel estimator converged. In addition to Figure 7, we refer to Figure 1 of Gretton et al. (2007) to show a further comparison between the CDFs of a generalized χ^2-distribution and its approximation with a Gamma distribution with the same mean and variance.

Finally, we present below a proof of Proposition 2 that characterizes the mean and variance of the random variable we use to construct our test in terms of integrals of the covariance $\eta(t,s)$ of the Brownian bridge.

Proof of Proposition (2). For any quantile density q that satisfies condition (iii) from Assumption 3, we can define the variable:

$$T_0(q) \doteq \sum_{e=1}^{E-1} \int_0^1 B_e^2(t) q^2(t) dt$$

where $(B_e(t))_{e=1}^{E-1}$ are $(E-1)$ independent Brownian bridges. Note that the variable in Equation (10) is simply $n^{-1}T_0(\hat{q})$. So we just need to compute the expectation and variance of $T_0(q)$ in order to prove Proposition 2. First the expectation; by switching the integral sign with the expectation, we get:

$$\mathbb{E}[T_0(q)] = \sum_{e=1}^{E-1} \int_0^1 \mathbb{E}\left[B_e^2(t) q^2(t)\right] dt = (E-1) \int_0^1 t(1-t) q^2(t) dt = (E-1) \int_0^1 \eta(t,t) q^2(t) dt.$$

Then, by independence of the Brownian bridges the variance is:

$$\mathrm{Var}[T_0(q)] = \sum_{e=1}^{E-1} \mathrm{Var}\left[\int_0^1 B_e^2(t) q^2(t) dt\right] = (E-1)\left(\mathbb{E}\left[\left(\int_0^1 B_1^2(t) q^2(t) dt\right)^2\right] - \left(\int_0^1 \eta(t,t) q^2(t) dt\right)^2\right).$$

Furthermore notice that:

$$\begin{aligned}
\mathbb{E}\left[\left(\int_0^1 B_1^2(t) q^2(t) dt\right)^2\right] &= \mathbb{E}\left[\left(\int_0^1 B_1^2(t) q^2(t) dt\right) \cdot \left(\int_0^1 B_1^2(s) q^2(s) ds\right)\right] \\
&= \mathbb{E}\left[\int_0^1 \int_0^1 B_1^2(t) B_1^2(s) q^2(t) q^2(s) dt ds\right] \\
&= \int_0^1 \int_0^1 \mathbb{E}[B_1^2(t) B_1^2(s)] q^2(t) q^2(s) dt ds \\
&= \int_0^1 \int_0^1 \left(\mathrm{Var}[B_1(t)]\mathrm{Var}[B_1(s)] + 2\mathrm{cov}(B_1(t), B_1(s))^2\right) q^2(t) q^2(s) dt ds \\
&= \int_0^1 \int_0^1 \left(\eta(t,t)\eta(s,s) + 2\eta^2(s,t)\right) q^2(t) q^2(s) dt ds \\
&= 2\int_0^1 \int_0^1 \eta^2(s,t) q^2(t) q^2(s) dt ds + \left(\int_0^1 \eta(t,t) q^2(t) dt\right)^2
\end{aligned}$$

Therefore:

$$\mathrm{Var}[T_0(q)] = 2\int_0^1 \int_0^1 \eta^2(s,t) q^2(t) q^2(s) dt ds$$

□

F.3 Bootstrap Approximation

Our bootstrap heuristic proceeds as follows:

1. Generate bootstrap samples in each environment by drawing with replacement n_e samples from the existing observations.
2. Based on the resulting bootstrap samples, compute $\hat{\Gamma}(\mathcal{F})$.
3. Repeat the above process $B = 50$ times, and compute the average $\hat{m}$ and variance $\hat{\sigma}^2$ for $\hat{\Gamma}(\mathcal{F})$ over these B simulations.
4. Use a Gamma distribution with mean $\hat{m}$ and variance $\hat{\sigma}^2$ for setting the thresholds.

Note that we saw in practice that increasing the number of bootstrap iterations B over 50 did not increase performance in our simulations.

G DETAILS ON THE SIMULATIONS AND ADDITIONAL EXPERIMENTS

In this section we provide more details on the experiments carried out in Section 6, as well as other additional experiments.

G.1 Additional Details on the Simulations from Section 6

For each of the simulated graphs, we consider $p+1$ variables that we randomly permute to determine their causal order; the 21st variable in this permutation is declared as the target variable, so that it has 20 non-descendants and 30 non-ancestors. Then for each pair of variables with probability k/p where $k = 12$ is the average degree, we connect them with an arrow (the direction of which is determined by their causal orders). For the target, we drop any of its incoming arrows generated by the above procedure, and instead we randomly select a subset S^* (where $|S^*| = 6$) among its 20 non-descendants as its parents and we connect the target to them with incoming arrows.

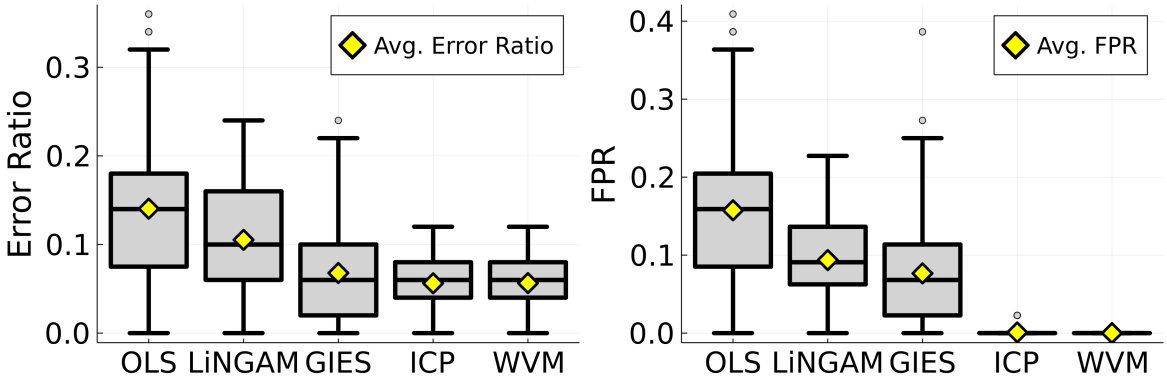
Once the structure of the graph is drawn, we generate the linear Gaussian structural equations for each of the nodes in the graph as follows: The Gaussian noises each have mean zero and their variances are sampled uniformly and independently for each graph from $[0.3^2, 1]$. We uniformly sample the linear coefficients for the parents of the node in absolute value independently for each node in the graph from [0.2, 1], and we switch their signs with probability 1/2. We normalize the coefficients so that the linear function of the parents has variance 1; we do so to avoid having extreme variances for variables that appear at the bottom of the graph.

At each intervened node, with probability 2/3, we scale the noise by a random scaling factor uniformly distributed on $[lb, ub]$, where lb and ub are chosen uniformly at random from [0.5, 5] with $lb < ub$. However, with probability 1/3, the scaling factor is chosen to be a constant, equal to the mid-point between lb and ub. Furthermore, the mechanistic intervention only changes the coefficients with probability 1/3 by adding a standard normal noise to them, otherwise the coefficients remain unchanged. Lastly, we added two additional constraints: we let all variables (except the target) be intervened on at least once and each environment with its consecutive environment share 40% of the variables. More precisely, in each environment we intervene on 65% of the predictors, chosen at random (note however that the scale of each of these interventions based on the choice of lb and ub above can often be negligible or not statistically significant w.r.t. the sample size).

Finally, we use in our experiments the R package *pcalg* (Kalisch et al., 2012) for GIES and LiNGAM, and for ICP we use the R package *InvariantCausalPrediction* (Peters et al., 2016). All experiments were run on a laptop with a quadcore 2.7 GHz Intel Core i7 processor.

G.2 Additional Linear Experiments

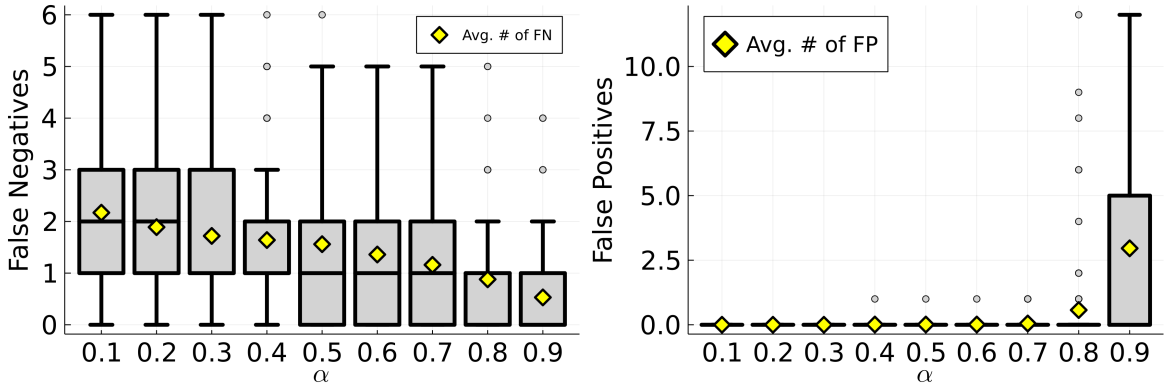
Mixed Noise Distributions. In addition to the main simulations performed in Section 6, we investigate how using a variety of distributions would affect the error ratio and the false positive rate for the various methods, in particular how the introduction of heavy-tailed distributions would affect the results. Concretely, we sample uniformly for each variable a noise distribution from either: a standard normal, a standardized Student's



(a) (b)

Figure 8: Error ratio (a) and False Positive Rate (b) of various causal discovery methods with noise distributions sampled from either a Normal distribution, some Student's T-Distributions or a uniform distribution.

T-distribution (mean zero, unit variance) with degrees of freedom equal to 3, 5, 10, 20, or 50, and a mean zero, unit variance uniform distribution. All other parameters are the same as in the main simulations from Section 6. As is seen in Figures 8a and 8b, the results are not dramatically different from the results using only a standard normal distribution as the noise distribution for each variable. The only noticeable differences are seen with GIES having a worse FPR (and thus worse Error Ratio) and WVM having a slightly worsened error ratio now comparable to ICP. A possible reason why WVM might behave slightly worse under that setting could be the fact that the asymptotic result from (8) doesn't hold for heavy-tailed distributions (see Remark 1 in Section E.1 for further details, and a simple solution to this issue).



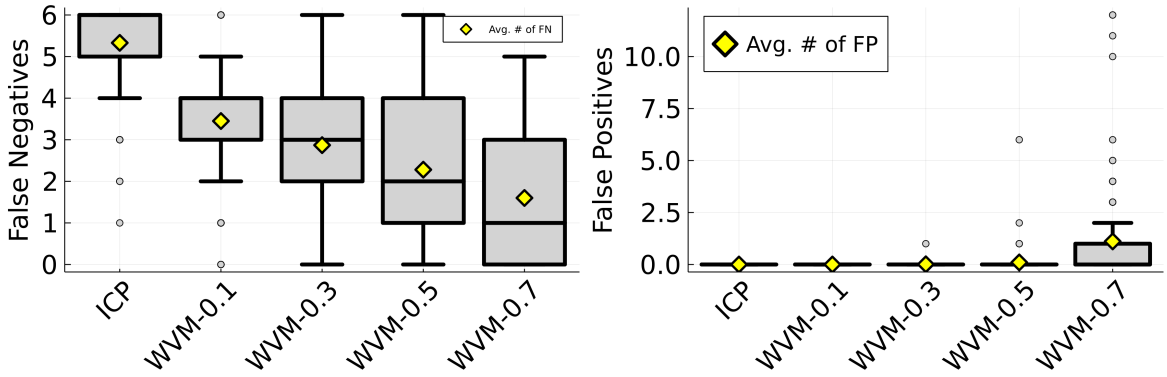
(a) (b)

Figure 9: Numbers of false negatives (a) and false positives (b) vs varying levels of significance (α) for WVM.

Varying α for WVM. Since the asymptotic distribution from Theorem 2 and our bootstrap heuristic can sometimes lead to thresholds that are too conservative in finite samples, we investigate here the use of higher confidence levels α for the WVM algorithm. More precisely, using the same simulations with linear Gaussian SCMs as in Section 6, which are also described in Section G.1, we display in Figures 9a and 9b the distributions of the numbers of false negatives and false positives across the 100 generated data-sets for different values of α from 0.1 to 0.9. It turns out that taking a higher confidence level α can in fact significantly improve the performance of WVM. In particular in this experiment, using $\alpha = 0.7$ halves the average number of false negatives compared to $\alpha = 0.1$, while maintaining the average number of false positives quasi-identical.

A reason why using a higher α can be beneficial in some cases is that the statistical test developed in Theorem 2 was derived to be consistent and of asymptotic level α for rather general classes of functions $\mathcal{F}$; it's possible that for specific classes of functions (e.g. the class of linear functions) the distribution of the variable in (10) could be too conservative in the sense that it would only be an upper bound on the true asymptotic distribution of

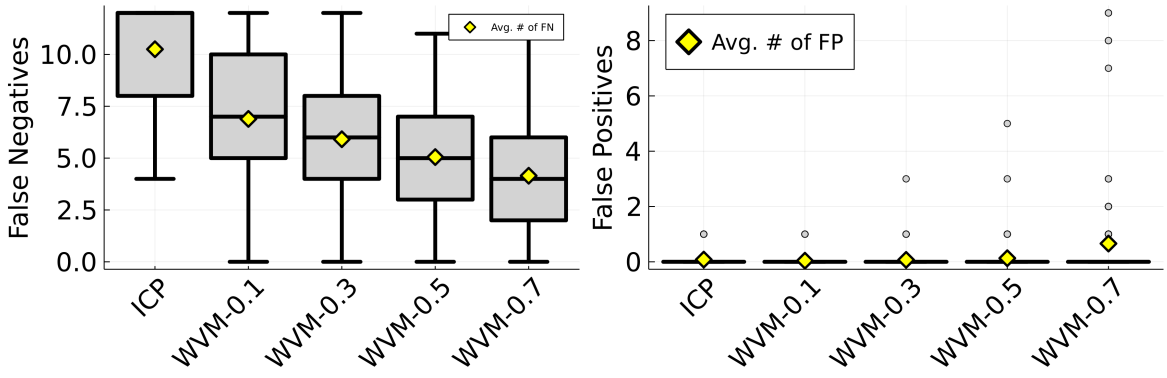
$\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_{-k})$ under $\tilde{H}_{0,k}(\mathcal{E})$. Therefore, it could be of interest to investigate the asymptotic distribution of $\hat{\Gamma}_{\boldsymbol{w}}(\mathcal{F}_{-k})$ for more specific classes of functions. As discussed in Section 7, we leave such considerations for future research.



(a) (b)

Figure 10: Number of false negatives (a) and number of false positives (b) for ICP and WVM with varying levels of α for a smaller sample size simulation with $n_e = 100$.

Smaller Sample Size. We now investigate the performance of WVM vs ICP for small sample sizes. We use again the simulations from Sections 6 and G.1 with the only difference that we now set $n_e = 100$ in each environment – all other parameters are kept identical. We display our results in Figures 10a and 10b, where we also choose $\alpha \in [0.1, 0.3, 0.5, 0.7]$ for WVM and $\alpha = 0.1$ for ICP – we observed in practice that using different confidence levels α for ICP does not improve its performance significantly. In this setting, WVM significantly outperforms ICP in terms of number of false negatives, while maintaining a number of false positives equal to zero as ICP (or at least close to zero on average for some higher αs); the difference between ICP and WVM is even more pronounced when α is larger than 0.1 for WVM.

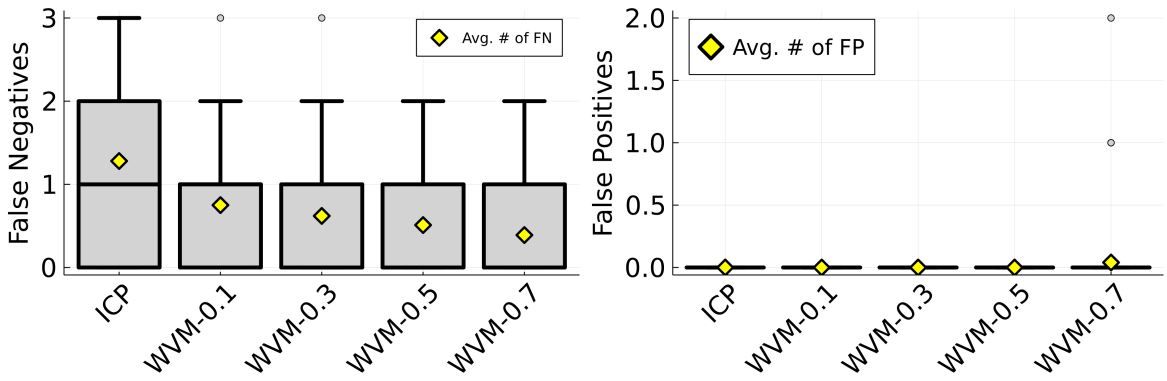


(a) (b)

Figure 11: Number of false negatives (a) and number of false positives (b) for ICP and WVM with varying levels of α for a simulation with a larger number of direct causes, $|S^*| = 12$.

More Direct Causes. We look at the case where the number of direct causes is set to be $|S^*| = 12$, which is half more the number of variables that ICP pre-selects with boosting (or lasso) in its default implementation. The purpose of this section is therefore to investigate the behavior of ICP when the number of direct causes is higher than the number of pre-selected variable it uses, compared to WVM for different levels $\alpha \in [0.1, 0.3, 0.5, 0.7]$ and with a number of pre-selected variables set at 18. Such a scenario can happen in practice since ICP cannot be used efficiently even on a moderate number of variables, and therefore has to restrict itself to a small subset of pre-selected variables, while WVM can easily handle dozens of variables.

The simulations used for this experiments are the same as in Sections 6 and G.1, with the only difference that we set S^* to be of size 12. As shown in Figures 11a and 11b ICP returns an empty set in many occurrences under this setting; WVM on the other hand significantly outperforms ICP by retrieving at least roughly half of the direct causes on average with almost no false positives; again, the difference between ICP and WVM is even more pronounced when α is higher than 0.1 for WVM.



(a) (b)

Figure 12: Number of false negatives (a) and number of false positives (b) for ICP and WVM with varying levels of α for a simulation with a smaller number of predictors, $p = 10$.

Smaller Number of Predictors. Finally, we consider the case where the number of predictors is 10, i.e. $p = 10$, so that there was no need to pre-select the variables with lasso for both ICP and WVM. Since $p = 10$, we also needed to change the number of direct causes to $|S^*| = 3$, the number of non-descendants to 6, number of non-ascendants to 4 and the average degree was set to $k = 3$. All the other parameters for the data generating process remained the same. This simulation shows that the higher power of WVM compared to ICP is not just a result of ICP requiring a more restrictive pre-selection step than WVM; with no pre-selection for both methods WVM still outperforms ICP (Figures 12a and 12b).

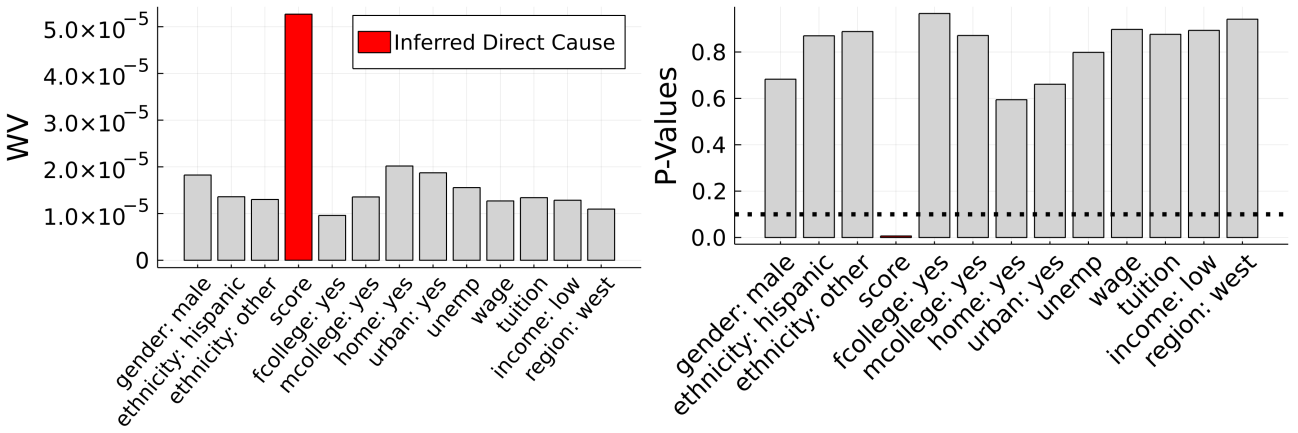
G.3 Real Data – Educational Attainment

We ran WVM on a real-world data set about the educational attainment of US teenagers (Rouse, 1995) – the data was accessed from Kleiber and Zeileis (2008). The dataset consists of 4739 students from 1100 US high schools; the purpose here is to study which factors are causal predictors of whether these students will obtain a Bachelor of Arts degree. Concretely, there are 13 features recorded, some of which include gender, ethnicity, composite score on an achievement test, whether the father or mother graduated college, etc. In this setting, the target variable is binary and indicates whether the student had greater than 16 years of education or not (the length of time required to obtain a Bachelor of Arts degree in the US). Even though WVM assumes a continuous target variable, we investigate its performance when the target is binary; we do so by applying WVM as is, but acknowledge that some modifications could improve its results.

To construct different environments from this observational data we consider an approach taken from the original ICP paper (c.f. section 3.3 Peters et al. (2016)) in which a variable U is chosen that is not the target and is known to be a non-descendent of the target in order to split the data by conditioning on U. A concrete example is when U precedes the target chronologically. We choose (as was done in ICP) the distance to the nearest 4-year college as the conditioning variable and split the data into two environments: students who lived within the median distance of 10 miles to a 4 year college, and students who lived farther away.

Figure 13 shows the output of WVM run on the educational attainment dataset. At $\alpha = 0.1$ WVM infers only one of the variables to be a direct cause of whether a student will obtain a Bachelor of Arts degree or not – the composite score on a standardized test (denoted as *score* in Figure 13). This result is similar to that of ICP's; they only infer one more variable as a cause: Whether the student's father received a college degree, the variable *fcollege*. We believe this discrepancy can be explained by model misspecification. Specifically, ICP can sometimes

return ancestors of the target that are not necessarily direct causes under model misspecifications such as hidden confounders, which there are likely to be in this real world dataset (c.f. Section 6.3 and Proposition 5 Peters et al. (2016)). Since, WVM is agnostic to hidden confounders (assuming that the confounding factor is independent of the environment, see Appendix B) this might explain why WVM does not consider *fcollege* as a direct cause (and why ICP does) as it is likely that *score* is a mediator between *fcollege* and the educational attainment of the student. We note however, that we are not trying to to make any causal claims here and include this example to showcase that our method could be applied to real data.



(a) (b)

Figure 13: Histograms of WV Values and P-values for the educational attainment dataset with two environments. The inferred direct cause from WVM is given in red.