Quantile-constrained Wasserstein projections for robust interpretability of numerical and machine learning models

Marouane Il Idrissi, Nicolas Bousquet, Fabrice Gamboa, Bertrand Iooss, Jean-Michel Loubes

Quantile-constrained Wasserstein projections for robust interpretability of numerical and machine learning models

Marouane Il Idrissi[a,b,c,e], Nicolas Bousquet[a,b,d], Fabrice Gamboa[c], Bertrand Iooss[a,b,c], Jean-Michel Loubes[c]

[a] *EDF Lab Chatou, 6 Quai Watier, 78401 Chatou, France*
[b] *SINCLAIR AI Lab., Saclay, France*
[c] *Institut de Mathématiques de Toulouse, 31062 Toulouse, France*
[d] *Sorbonne Université, LPSM, 4 place Jussieu, Paris, France*
[e] *Corresponding Author - Email: marouane.il-idrissi@edf.fr*

Abstract

Robustness studies of black-box models is recognized as a necessary task for numerical models based on structural equations and predictive models learned from data. These studies must assess the model's robustness to possible misspecification of regarding its inputs (e.g., covariate shift). The study of black-box models, through the prism of uncertainty quantification (UQ), is often based on sensitivity analysis involving a probabilistic structure imposed on the inputs, while ML models are solely constructed from observed data. Our work aim at unifying the UQ and ML interpretability approaches, by providing relevant and easy-to-use tools for both paradigms. To provide a generic and understandable framework for robustness studies, we define perturbations of input information relying on quantile constraints and projections with respect to the Wasserstein distance between probability measures, while preserving their dependence structure. We show that this perturbation problem can be analytically solved. Ensuring regularity constraints by means of isotonic polynomial approximations leads to smoother perturbations, which can be more suitable in practice. Numerical experiments on real case studies, from the UQ and ML fields, highlight the computational feasibility of such studies and provide local and global insights on the robustness of black-box models to input perturbations.

Keywords: interpretability, machine learning, sensitivity analysis, computer model, sensitivity analysis, robustness, epistemic uncertainty, domain uncertainty, quantiles, isotonic polynomials

1. Introduction

Multiple engineering fields require models for prediction and phenomenological understanding. Machine learning (ML) and uncertainty quantification (UQ) of numerical models are two essential approaches to developing and manipulating such models. Because they require, for their enlightened use, an adequate understanding of their characteristics, they share fundamental similarities. These two frameworks feed off each other through the duality of sensitivity analyses (SA), a fundamental methodological corpus in UQ, and ML interpretability methods, as explained by [90, 61]. In particular, recent advances in explainable ML leverage tools from SA to produce meaningful interpretations of black-box models [36, 13], and novel SA estimation schemes are heavily based on the construction of suitable ML models [18, 12]. Both SA and ML interpretability especially rely on the definition, estimation, and manipulation of diagnostics related to the characteristics of a model and how its behavior depends on its inputs [28, 74, 95]. Formally speaking, let a model f be defined as a mapping between *inputs* $X \in \mathcal{X}$ and *outputs* $Y \in \mathcal{Y}$ where $(\mathcal{X}, \mathcal{Y})$ are two metric spaces:

$$Y = f(X).$$

In a ML context, f is defined as a *predictive model* (e.g., penalized linear regression, neural network) linking a feature (or covariate) instance X to a prediction Y [54]. In the UQ framework, a so-called

computer model f represents the numerical implementation of a hypothetical-deductive link (e.g., by systems of ordinary differential equations, by finite element methods) between X and Y [100].

In both fields, the input $X \in \mathcal{X}$ is generally assumed to be random, inducing a general framework for handling uncertainties about the latter. Let P be the distribution of X. In the ML context, P is defined implicitly by an empirical measure: given a set of observations $x^{(1)}, \ldots, x^{(n)} \in \mathcal{X}$,

$$P = \frac{1}{n} \sum_{i=1}^{n} \delta_{x^{(i)}} \tag{1}$$

where δ denotes the Dirac measure. On the other hand, in the UQ setting, P is often explicitly chosen based on observations of X, expert assessment (domain knowledge), or stochastic inversion from observations of Y [105]. The diagnostics mentioned above correspond to estimations of key interpretable features of Y, or *quantities of interest* (QoI). In the SA literature, such a QoI is often referred as the *score* [94], while ML researchers rather speak about *predictive performance* [81]. Recall that SA aims to rank the dimensions of X according to their influence on this QoI [20]. For instance, in local SA, it is usually computed by way of a differential operator with respect to (w.r.t.) the dimensions of X [28]. In global SA, it can be typically chosen as the output's variance or its quantiles [36, 72]. More general objects characterizing the distribution of Y (e.g., a kernel embedding [9]) can also be of interest, possibly at the cost of a less immediate interpretability. In ML interpretability, local methods focus on a particular prediction instance, letting the QoI be the identity function [104] or a local linear decomposition of f [113]. Global ML interpretability often relies on QoI defined as performance metrics (e.g., accuracy, loss value) of f computed over a training dataset [49, 26, 60].

Accordingly, the use of these diagnostics to allow different levels of interpretability is subject to the same robustness problem: they must remain relevant when P suffers from *misspecification*. It would improve the confidence in both the usage and the insights that an ML model offers [10]. In this framework it is fundamentally connected to problems of domain adaptation and transfer learning [19, 70] (e.g., when the data used for the design of f suffer from selection bias w.r.t. operational data), including in particular the robustness to *covariate shift* [51, 109] (see Figures 1 and 2 for an illustration). More generally, whether in UQ or ML, this misspecification is due to the epistemic uncertainty affecting the knowledge of the properties of P (e.g., support, geometry, topology) due to the finiteness of the available information (e.g., data, expertise, boundary conditions) [58, 105].

This epistemic uncertainty about P thus impacts the generalization capability of f. It is essential

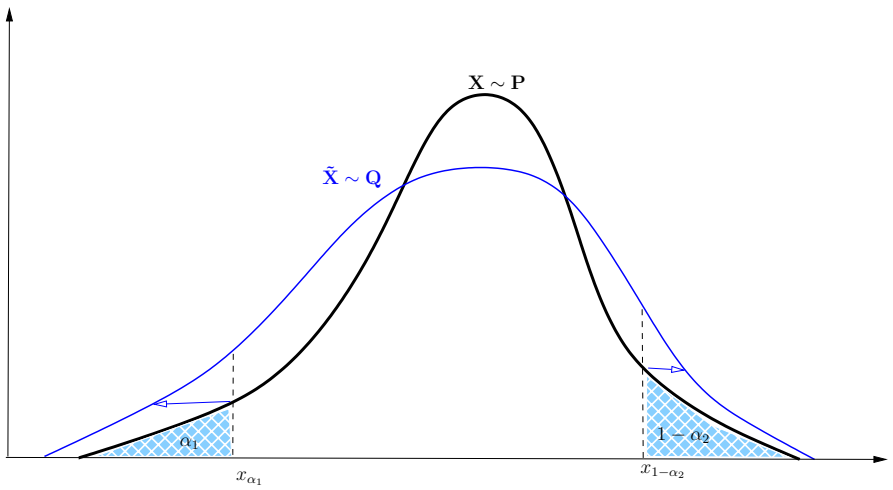

Figure 1: One-dimensional illustration of covariate shift in the UQ framework. A postulated density distribution for $X \sim P$ (solid black line) is dilated by modifying the tail order of low and high quantiles defining an application domain (blue arrows), resulting in a new distribution Q.

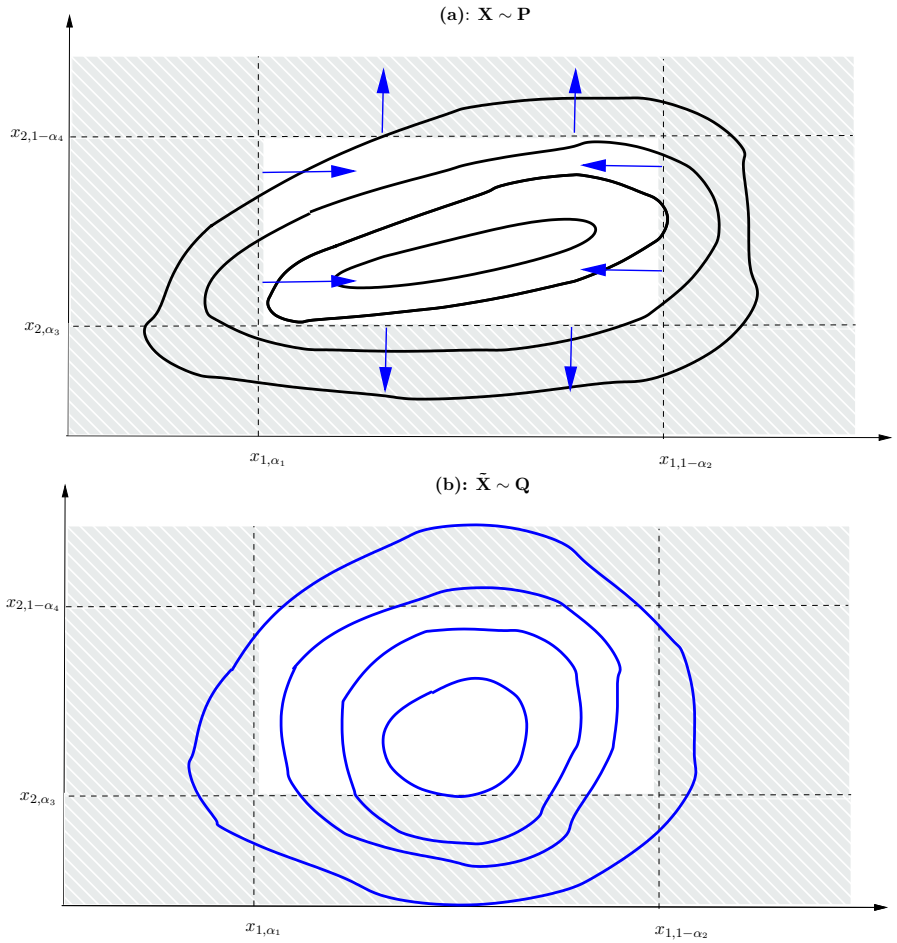

Figure 2: Two-dimensional illustration of covariate shift in the UQ framework. Graph **(a)** displays the isodensity curves of a joint distribution P on $X = (X_1, X_2)$. Graph **(b)** displays the result of a simultaneous quantile shrink for X_1 and quantile dilatation for X_2 while preserving the dependence structure.

to define a practical *application domain* [93], or usage domain of f, allowing to extrapolate beyond a validity domain based on selected situations (e.g., a domain including unobserved data points in ML, or unsure structural mechanisms in numerical models). Generalization bounds in ML usually depend on notions of capacities (e.g., the Vapnik–Chervonenkis dimension [111]). They rely on a sufficient number of observations to obtain generalization, i.e., linked to epistemic uncertainty reduction. Hence, generalization power and epistemic uncertainties are intimately linked [110]. Therefore, robustness studies of the application domain's uncertainty are particularly important for both fields. It motivated the search for adversarial situations in ML [4] and the use of imprecise probabilities in numerical model exploration [11].

The need to conduct robustness studies raises the question of how to reasonably vary P in a *perturbation class* $\mathcal{C} \subset \mathcal{P}(\mathcal{X})$, where $\mathcal{P}(\mathcal{X})$ is the set of probability measures supported on $\mathcal{X}$, by modifying the mass given to different parts of $\mathcal{X}$ in an interpretable way. Such variations allow checking the consistency of the information delivered by these diagnostics. Here, *reasonableness* means two things: a perturbed distribution Q must remain close to the original P in some formalized sense, and the computations required for the robustness analyses must have a moderate cost.

Several authors in both fields have conducted this type of study. In the ML setting, finding adversarial situations or detecting bias affecting a QoI can stand on moving the original data (e.g., by

optimal transport [89, 47]), resulting in an empirical shift of P. Researchers in UQ distinguish between *distributional sensitivity analysis* (DSA) [2, 76] which aims at providing a robustness diagnosis on a QoI used in SA (or connected similar studies, e.g., in financial risk analysis [25]), and other robustness approaches whose goal is to study the sensitivity of a given QoI to a gradual modification of P in $\mathcal{C}$. For instance [69], [43] and [63] proposed such indices for exceedance probabilities, quantiles or superquantiles of the distribution of Y under various perturbation schemes.

Many choices have been proposed for perturbation classes. They are often indexed by parameters θ and denoted $\mathcal{C}(\theta)$. For instance, in the SA setting, they can be defined as contamination classes for Bayesian SA [50, 92], as classes of generalized moments [69, 6] or as classes defined by Fisher-Rao derivatives [67, 43], among others. A unified view of these approaches is to define the perturbed distribution Q as the solution to the projection problem

$$\begin{aligned} Q = \underset{G \in \mathcal{P}(\mathcal{X})}{\operatorname{argmin}} \quad & \mathcal{D}(P, G) \\ \text{s.t.} \quad & G \in \mathcal{C}(\theta). \end{aligned} \tag{2}$$

where $\mathcal{D}$ is a discrepancy between probability measures and $\mathcal{C}(\theta)$ denotes a fixed perturbation class. The Kullback-Leibler (KL) divergence is often chosen as a suitable discrepancy (e.g. in [69, 6]), leading to entropic projections [27]. For instance in those developments, whenever $\mathcal{X} \subset \mathbb{R}^d$ ($d \geq 1$), $\mathcal{C}(\theta)$ can be defined as the distributions belonging to $\mathcal{P}(\mathcal{X})$ with a deviated mean η, different from $\mathbb{E}_P[X]$. Reusing the term proposed by [6], $\theta = \|\eta - \mathbb{E}[X]\|$ can be called a *perturbation intensity*, where θ belongs to an ordered set Θ. Hence, for a fixed perturbation intensity θ, the solution Q of (2) is the optimally perturbed distribution of P w.r.t. $\mathcal{C}(\theta)$.

Although rational, moment-based perturbations of P presents significant issues that limit the conclusions of interpretability analyses, and the subsequent unification of UQ and ML methodologies. First, the choices of θ, $\mathcal{C}(\theta)$, and Θ remain subjective and submitted to strong assumptions. For instance, in [69], some generalized moment of Q are required to exist, and the moment deviation intensity must subjectively chosen by the practitioner. When $\dim(X)$ is large, such assessments seem challenging to make. In [43], Θ comprises the values taken by the derivative of the Fisher metric, requiring regularity properties on P, which is often assumed to be in a restrictive parametric family, for computational reasons. In general, determining the boundaries of Θ remains challenging.

Second, choosing the KL divergence or the derivative of the Fréchet metric as in [53] as a discrepancy imposes strong restrictions on the space of reachable probability distributions. It result in a search on a restricted part of $\mathcal{P}(\mathcal{X})$ in (2). For instance, using the KL divergence implies that Q should be absolutely continuous w.r.t. P, which does not allow to consider continuous perturbed distributions Q given an empirical initial distribution P. Finally, the behavior of these discrepancies (purely related to information geometry) remains uneasy to explain to non-specialists.

This article, therefore, aims to address both of these issues, improving the connections between the interpretability analyses conducted in UQ and ML settings, and their robustness. More precisely, our contributions are twofold:

(a) We propose a meaningful and generic approach to perturb distributions through marginal quantile constraints, selecting the 2-Wasserstein distance for $\mathcal{D}$. These choices allow to solve the problems mentioned above. A key point of the proposed methodology is that no strong regularity assumptions must be assumed on f and it does not rely on f having a particular structure (e.g., belonging to a particular family of predictive models or based on specific physical equations). This effectively unifies UQ and ML approaches.

(b) We demonstrate the computational tractability of this methodology by implementing it on different types of QoI, on both numerical and predictive models, and studying some robustness indicators to perturbations in the particular case where f is considered to be a black box.

This article is organized as follows. In Section 2 we first introduce useful notations and definitions. Section 3 develops and motivates the choice of quantiles as an interpretable and generic basis for

defining meaningful perturbations and defines perturbation schemes relevant for different robust interpretability studies. Section 4 presents the general framework of probability measure projection using the 2-Wasserstein distance and proposes analytical solutions and numerical optimization schemes for solving the input perturbation problem through the help of isotonic polynomials. Section 5 showcases our method on two use-cases from both the ML and UQ fields, from which local and global robustness insights are highlighted. A discussion section ends this article, opening avenues for improvement. All proofs of technical results are postponed to a dedicated appendix.

2. Notations and main definitions

Let us introduce some useful notations and definitions. Let d and p be two positive integers. Let $\mathcal{X}$ be a subset of $\mathbb{R}^d$, and $\mathcal{P}_p(\mathcal{X})$ be the subset of $\mathcal{P}(\mathcal{X})$ of measures with finite p-th moment. The initial probability measure P is either defined through an explicit distribution, or empirically, as in (1). We will denote by $Q \in \mathcal{P}(\mathcal{X})$ the optimally perturbed distribution of P.

Furthermore, let us denote $\Omega_X \subset \mathcal{X}$ the application domain. It is the subset of $\mathcal{X}$ where f is intended to be used for predictions [93]. Both in ML and UQ, given a set $\mathbf{x_n} = \{x_1, \dots, x_n\}$ of training, validating or testing examples, the convex hull of $\mathbf{x_n}$ or a broader span of $\mathbf{x_n}$ are common candidates to define Ω_X [116]. More generally, Ω_X is an extrapolation domain where f is assumed to generalize well (e.g., a paving of a compact subspace of $\mathcal{X}$ selected by tree-based classification [56], confidence measures or cross-validation schemes [57, 78, 114, 66]). In ML specifically, including out-of-distribution data in Ω_X remains an open problem [55, 114, 99].

To echo the classical assumptions in ML and UQ, we assume that Ω_X is the union of compact subsets of $\mathcal{X}$. These subsets can be defined under some uncertainties, typically on their bounds. In a robustness analysis perspective, assuming that the dependence structure is maintained, the uncertainties on Ω_X can be interpreted as variations on the values (or thresholds) of the extreme quantiles of marginal distributions. Figure 3 illustrates a typical situation for a univariate marginal of X.

The upcoming developments deal with perturbations on the marginal distributions of P. The following definitions recall classical results on cumulative distribution functions (cdf) and generalized quantile functions (gqf) of univariate probability measures. Denote $\mathcal{F}$ the space of univariate

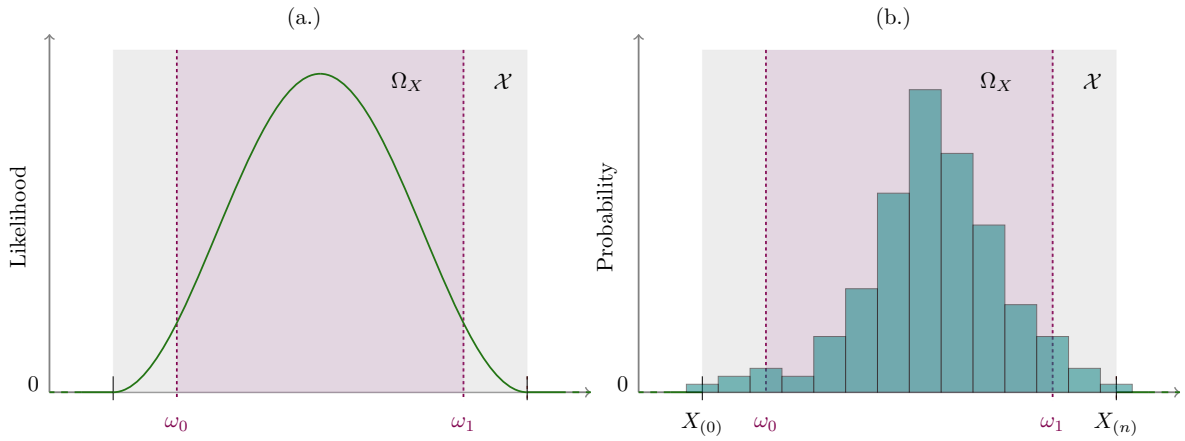

Figure 3: Application domain $\Omega_X \subset \mathcal{X}$ in the UQ (a.) and ML (b.) settings. In the UQ setting, $\mathcal{X}$ is the support of the explicitly chosen density (in green). In the ML setting, $\mathcal{X}$ is the interval between the minimum and maximum observed values. In both cases, Ω_X is included in $\mathcal{X}$, although it is not mandatory.

distribution functions:

$$\mathcal{F} = \Big\{ F : \mathbb{R} \to [0,1] \mid F \text{ is right-continuous, non-decreasing} \\ \text{such that } \lim_{x\to\infty} F(x) = 1 \text{ and } \lim_{x\to-\infty} F(x) = 0 \Big\}. \tag{3}$$

For any $P \in \mathcal{P}(\mathbb{R})$, $\mathcal{F}$ contains the cdf of P, defined as:

$$F_P(t) = \int_{(-\infty,t]} dP = P\big((-\infty, t]\big).$$

The gqf of any probability measure P can be formally defined as follows [91, 33, 64].

Definition 1 (Generalized quantile function)**.** *Let $P \in \mathcal{P}(\mathbb{R})$ with cdf F_P.*

(i) *The generalized quantile function (gqf) of P is the unique left-continuous, non-decreasing generalized inverse of F_P, defined, for every $a \in (0,1)$, as:*

$$\begin{aligned} F_P^{\leftarrow}(a) &= \sup \ \{t \in \mathbb{R} \mid F_P(t) < a\}, \\ &= \inf \ \{t \in \mathbb{R} \mid F_P(t) \geq a\}. \end{aligned} \tag{4}$$

(ii) *The unique right-continuous non-decreasing generalized inverse $F_P^{\rightarrow}$ of F_P, almost-everywhere equal to $F_P^{\leftarrow}$, is defined, for every $a \in (0,1)$, as:*

$$\begin{aligned} F_P^{\rightarrow}(a) &= \sup \ \{t \in \mathbb{R} \mid F_P(t) \leq a\}, \\ &= \inf \ \{t \in \mathbb{R} \mid F_P(t) > a\}, \\ &= F_P^{\leftarrow}\left(a^+\right) \end{aligned} \tag{5}$$

where $F_P^{\leftarrow}(a^+) = \lim_{x\to a^+} F_P^{\leftarrow}(x)$.

It is important to note that when F_P admits an inverse F_P^{-1} in the traditional sense (e.g., it is continuous, strictly increasing), then the following equality holds:

$$F_P^{-1} = F_P^{\leftarrow} = F_P^{\rightarrow},$$

which echoes and generalizes the traditional definition of a quantile function as the inverse of a cdf.

In the remainder of this work, any probability measure is assumed to have a finite 2-nd order moment; hence, their gqf is assumed to be square integrable. Accordingly, the perturbation θ considered in the projection problem (2) corresponds to quantile constraints placed on one-dimensional marginal variables. Therefore it is assumed that $\Theta \subset \mathbb{R}$.

3. Quantile perturbations

In this section, we define a particular set of perturbation classes. They are characterized as a set of constraints on the (generalized) quantiles of the marginal components of the inputs (or features) variables X of f. Before presenting a formal definition of such classes, we argue their relevance for practical studies.

3.1. Motivations

First, *generalized quantiles always exist*, unlike generalized moments considered by [69]. It is also what motivated [6], who proposed to define the perturbation range Θ in (2) as a compact set bounded by empirical quantiles. More precisely, recalling Definition 1, a gqf is the left-generalized inverse of a cdf. Moreover, they allow to uniquely characterize probability measures in $\mathcal{P}(\mathbb{R})$.

Remark 1. *The unique link between left-generalized inverses of functions in $\mathcal{F}$ and probability measures can be formalized as follows. Let $F^{\leftarrow}$ be a function in the set:*

$$\mathcal{F}^{\leftarrow} = \left\{ F^{\leftarrow} : (0,1) \to \mathbb{R} \mid F^{\leftarrow} \text{ is left-continuous and non-decreasing} \right\}. \tag{6}$$

Then there exists a unique probability measure in $\mathcal{P}(\mathbb{R})$ with cdf F, and such that $F^{\leftarrow}$ is its gqf. This classical result seems to be widely known in the literature. For completeness, a proof sketch is provided in the appendices.

Since generalized inverses of functions in $\mathcal{F}$ always exist, perturbing marginal quantiles in (2) do not require additional assumptions either on the initial probability measure P or on the shape of the target probability measure Q. Hence, it allows for generic, well-defined perturbations, which is key for merging both SA and ML interpretability.

Second, *constraints placed on (generalized) quantiles can be interpretable.* Considering $b_i \in \mathbb{R}$ as a representative magnitude of a real component X_i of X, then $\mathbb{P}(X_i < b_i) = \alpha_i \in (0,1)$ if and only if b_i minimizes the expected pinball cost function [23]:

$$\ell_i(b) \quad = \quad \mathbb{E}_{X_i}\left[|X_i - b| \left\{ \mathbb{1}_{\{X_i < b\}} + \frac{1-\alpha_i}{\alpha_i} \mathbb{1}_{\{X_i \geq b\}} \right\} \right].$$

This means that α_i can be interpreted as the ratio $c_1/(c_1+c_2)$ where $c_1|x_i - b|\mathbb{1}_{\{X_i<b\}}$ and $c_2|x_i - b|\mathbb{1}_{\{X_i \geq b\}}$ are relative first-order costs associated to estimating x_i by b. When X_i is an influential input, it seems relevant to produce rules to modify these costs (then the value of α_i), shifting the distribution of X_i. In particular, this information-theoretic interpretation of quantiles promotes their variation in sensitivity studies conducted in economic or financial contexts (see, e.g., [85]). Moreover, Bayesian statisticians have long recognized this representation as one of the most appropriate formal approaches to incorporating expert knowledge into statistical model inference [44, 73].

Third, in many applied problems, *(generalized) quantile specifications are often key to studying the influence of input variables on a decision-making output.* In both the UQ and ML framework, inputs X can themselves be partially derived from calculations from upstream learning or numerical models (e.g., for multi-physics problems). P is then often calibrated by quantile matching, which may introduce uncertainties on some of its marginal features [103]. Numerous applications dealing with economic stress tests or risk mitigation against natural hazards use quantiles as influential inputs of decision-helping models. For instance, in the drought risk studies in [35], the association between soil wetness, climatic, seismic, and socioeconomic variables (e.g., city-level description) is often carried out using marginal quantiles that play the role of features for cost predictive models. Input variations of daily value-at-risk percentiles, computed from legacy data, were recently required by the European Banking Authority for generating macroeconomic scenarios used for EU-wide stress tests [5]. Reverse SA studies for financial risk management, such as those conducted in [86], are primarily based on moving values-at-risk, which are quantiles.

Let us end this subsection with two motivating examples. They offer two additional concrete illustrations of using quantiles for influence analysis. They also illustrate two different quantile perturbation schemes: quantile shifting and application domain dilatation. These schemes are formally introduced in Section 3.3.

Example 1 (Economic stress test)**.** *Inspired by [16], assume that an economic shock happens in an abstract country. Prospective analyses announce a* \$200 *drop in the population median wage. Before*

the shock, the population wage distribution P is known (or observed), thanks to some annual census data. This distribution has a median wage of \$2000*. The new population wage distribution is unknown due to the lack of recent data. The economists would like to know if they can be confident in their predictive macro-economic model f w.r.t. this sudden change. A way to answer this problem would be assessing the behavior of the model f on a distribution Q, such that:*

$$F_Q^{\leftarrow}(0.5) = 1800.$$

Example 2 (River water level)**.** *This example is inspired from [62] and more deeply studied in Section 5.2. The safety of an industrial site located near a river is studied through the prediction of the water level $Y = f(X)$ where f is a numerical hydrodynamic model, and X gathers the physical features of the river. A key dimension of X is the Strickler roughness coefficient for the upstream water level [42], which is modeled as a truncated Gaussian distribution on $\Omega_X = [20, 50]$. However, this application domain is tainted with epistemic uncertainties on the actual nature of the riverbed (e.g., more or less subject to shrubby vegetation). The practical use of f would require assessing its predictive power under a wider interval $\Omega_X = [5, 65]$. A way to express this prospective study is to assess the model's behavior on a distribution Q, such that:*

$$F_Q^{\rightarrow}(0) = 5, \quad F_Q^{\leftarrow}(1) = 65.$$

3.2. A formal definition of quantile perturbation classes

Focusing on a univariate input $X \sim P \in \mathcal{P}(\mathbb{R})$, let us first recall the formal definition of a quantile. For $P \in \mathcal{P}(\mathbb{R})$ and $X \sim P$, for $\alpha \in [0, 1]$, an α-quantile of P is a number $p_\alpha \in \mathbb{R}$ such that:

$$P(\{X < p_\alpha\}) \leq \alpha \quad \text{and} \quad P(\{X \leq p_\alpha\}) \geq \alpha.$$

In certain cases, an α-quantile is not unique. For instance, assuming that P is purely atomic (e.g., an empirical measure) and that its cdf F_P takes the constant value α on an open interval (t_0, t_1) (i.e., it is the case if t_0 and t_1 are both atoms of an empirical probability measure), then any $t \in (t_0, t_1)$ is an α-quantile. By convention, the gqf of P defined by (4) is the smallest of the α-quantiles of P (in this case, $F_P^{\leftarrow}(\alpha) = t_0$).

As a result, given a chosen $b \in \mathcal{X}$, defining a perturbed version Q of P through the equality constraints $F_Q^{\leftarrow}(\alpha) = b$ seems somewhat arbitrary. It would implicitly imply constraining the smallest α-quantile value of Q. Arguably, the value of b should be constrained to be a part of the set of all possible α-quantiles:

$$b \in \{q_\alpha \in \mathcal{X} \,|Q(\{X < q_\alpha\}) \leq \alpha,\ Q(\{X \leq q_\alpha\}) \geq \alpha\},$$

or written equivalently

$$F_Q^{\leftarrow}(\alpha) \geq b \geq F_Q^{\leftarrow}(\alpha^+) = F_Q^{\rightarrow}(\alpha). \tag{7}$$

In the case where F_Q is invertible, it becomes a traditional equality constraint: any α-quantile is uniquely defined (i.e., $F_Q^{\leftarrow}(\alpha) = F_Q^{\rightarrow}(\alpha)$). Constraints of the form (7), which are referred to as *quantile constraints*. They are the basis to define quantile perturbation classes.

Definition 2 (Quantile perturbation class)**.** *Let $K \in \mathbb{N}^*$ be the cardinality of a collection of quantile constraints defined by $\alpha = (\alpha_1, \ldots, \alpha_K)^\top \in [0, 1]^K$ and $b = (b_1, \ldots, b_K)^\top \in \mathbb{R}^K$. The quantile perturbation class $\mathcal{Q} \subset \mathcal{P}(\mathbb{R})$ is the set of probability measures defined as:*

$$\mathcal{Q} = \left\{Q \in \mathcal{P}(\mathbb{R}) \mid F_Q^{\leftarrow}(\alpha_i) \leq b_i \leq F_Q^{\rightarrow}(\alpha_i), \quad i = 1, \ldots, K\right\}.$$

Under simple conditions on the values α and b, quantile perturbation classes are non-empty.

Lemma 1. *Let $\mathcal{Q}$ be a perturbation class defined over $\alpha \in [0,1]^K$ and $b \in \mathbb{R}^K$, which are assumed to be ordered, without loss of generality. If*

$$0 \leq \alpha_1 < \cdots < \alpha_K \leq 1, \quad \text{and} \quad b_1 < \cdots < b_K, \tag{8}$$

then $\mathcal{Q}$ is non-empty.

Since $F_Q^{\leftarrow}$ can oftentimes be discontinuous, smoothing restrictions can be envisioned. It entails restricting $\mathcal{Q}$ to probability measures whose quantile functions are *smooth* (e.g., continuous, derivable). Smooth quantile perturbation classes can be introduced as follows.

Definition 3 (Smooth quantile perturbation class)**.** *Let $K \in \mathbb{N}^*$ and let $\alpha = (\alpha_1, \ldots, \alpha_K)^\top \in [0,1]^K$, $b = (b_1, \ldots, b_K)^\top \in \mathbb{R}^K$. Additionally, let $\mathcal{V} \subseteq \mathcal{F}^{\leftarrow}$ be a given set of smooth non-decreasing functions. The smooth quantile perturbation class $\mathcal{Q}_{\mathcal{V}} \subset \mathcal{P}(\mathbb{R})$ is the set of probability measures defined as:*

$$\mathcal{Q}_{\mathcal{V}} = \left\{ Q \in \mathcal{P}(\mathbb{R}) \mid F_Q^{\leftarrow} \in \mathcal{V}, \quad F_Q^{\leftarrow}(\alpha_i) \leq b_i \leq F_Q^{\rightarrow}(\alpha_i), \quad i = 1, \ldots, K \right\}.$$

These classes are further discussed and illustrated in Section 4.2.2. Note that smooth perturbation classes generalize perturbation classes since $\mathcal{Q} = \mathcal{Q}_{\mathcal{F}^{\leftarrow}}$. Therefore, without loss of generality, perturbation classes are denoted $\mathcal{Q}_{\mathcal{V}}$ in the following. In the next few paragraphs, echoing the illustrative examples in Section 3.1, we formalize two types of quantile constraints: quantile shifts and application domain dilatations.

3.3. Two key quantile perturbations

3.3.1. Quantile shift

The *quantile shift perturbation* aims at increasing as well as decreasing some of the quantiles of the initial distribution. Formally, given a quantile level $\alpha \in [0,1]$, and an initial α-quantile $p_\alpha = F_P^{\leftarrow}(\alpha)$, the quantile shift entails defining values for b in (7) ranging over a compact interval $[\eta_0, \eta_1] \subseteq \Omega_X$ such that $\eta_0 < p_\alpha < \eta_1$. The next lemma formalizes an expression for b as a function of a perturbation intensity θ standardized on $\Theta = [-1,1]$.

Lemma 2. *Let $\Theta = [-1,1]$ and denote $\boldsymbol{\eta} = (\eta_0, \eta_1)$ with $\eta_0 < p_\alpha < \eta_1$. For $\theta \in \Theta$, let,*

$$b_\alpha(\boldsymbol{\eta}, \theta) = \begin{cases} p_\alpha(1+\theta) - \theta\eta_0 & \text{if } -1 \leq \theta < 0, \\ p_\alpha & \text{if } \theta = 0, \\ p_\alpha(1-\theta) + \theta\eta_1 & \text{if } 0 < \theta \leq 1. \end{cases}$$

Then, for $Q_\theta \in \mathcal{P}(\mathbb{R})$ such that

$$F_{Q_\theta}^{\leftarrow}(\alpha) \geq b_\alpha(\boldsymbol{\eta}, \theta) \geq F_{Q_\theta}^{\rightarrow}(\alpha),$$

one has that, $\forall \theta \in \Theta$:

$$\begin{aligned} -1 \leq \theta < 0 \quad &\Leftrightarrow \quad \eta_0 \leq F_{Q_\theta}^{\leftarrow}(\alpha) < p_\alpha, \\ \theta = 0 \quad &\Leftrightarrow \quad F_{Q_\theta}^{\leftarrow}(\alpha) = p_\alpha, \\ 0 < \theta \leq 1 \quad &\Leftrightarrow \quad p_\alpha < F_{Q_\theta}^{\leftarrow}(\alpha) \leq \eta_1. \end{aligned}$$

We refer to Figure 4 (a.) for an illustration of this perturbation scheme, and to Section 5.1.1 for a real-world application. The quantile shift perturbation class can, for a given initial quantile level $\alpha \in [0,1]$, and valid interval bounds $\boldsymbol{\eta} = (\eta_0, \eta_1)$, $\eta_0 < p_\alpha < \eta_1$, be formally defined as the collection of perturbation classes

$$\mathcal{T}(\boldsymbol{\eta}, \theta) = \{ Q \in \mathcal{P}(\mathbb{R}) \mid F_Q^{\leftarrow}(\alpha) \leq b_\alpha(\boldsymbol{\eta}, \theta) \leq F_Q^{\rightarrow}(\alpha) \} \tag{9}$$

indexed by the intensity $\theta \in [-1,1]$. Each choice of θ induces a perturbation class for which the projection problem in (2) must be solved.

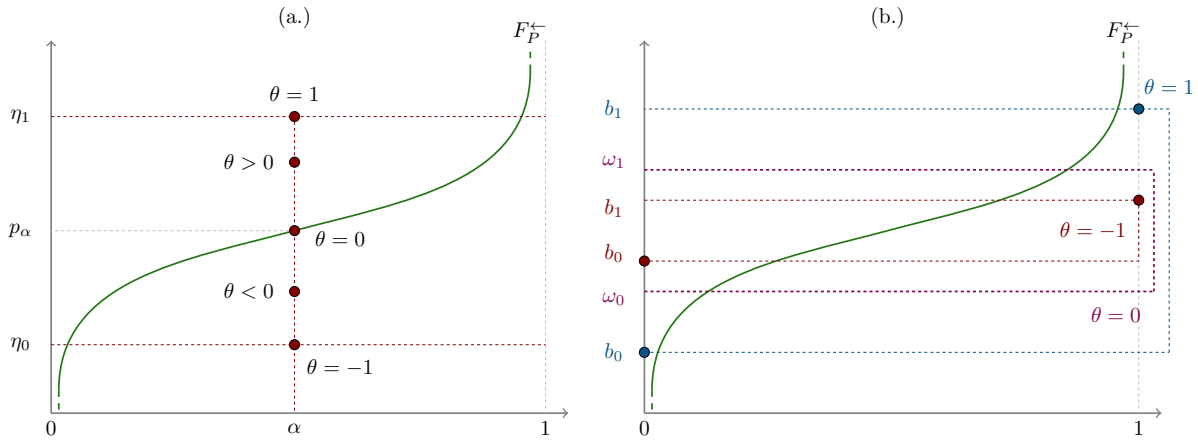

Figure 4: Quantile shift (a.) and application domain dilatation (b.) perturbation schemes. The initial quantile function is displayed in green. On the left, red points indicate different quantile shifting constraints between η_0 and η_1, leading to different intensity values θ. On the right, the application domain's width (in magenta) is up to doubled (blue points) or down to halved (red points), according to an intensity parameter θ evolving in a range $\Theta = [-1, 1]$.

3.3.2. Application domain dilatation

Application domain dilatation consists in perturbing the bounds of the application domain of an input. For a univariate $X \sim P$ with $\Omega_X = [\omega_0, \omega_1]$, the dilatation process amounts to widening or narrowing the width (or diameter $\text{diam}(\Omega_X)$) of Ω_X. It is similar to perturbing extreme quantile levels ($\alpha \in \{0, 1\}$) of P while preserving the midpoint of Ω_X. The dilatation is characterized by a parameter $\eta > 1$ controlling the rescaling magnitude of Ω_X while preserving its midpoint. In other words, one aims at finding a distribution Q with support $\text{Supp}(Q) = [b_0, b_1]$ for $b_0, b_1 \in \mathcal{X}$, $b_0 < b_1$, where the midpoint of $[b_0, b_1]$ is equal to the midpoint of Ω_X, but such that $\text{diam}(Q) := \text{diam}(\text{Supp}(Q))$ is rescaled compared to $\text{diam}(\Omega_X)$. Similarly to quantile shift, the next lemma formalizes expressions for these two bounds as a function of a perturbation intensity θ standardized on $\Theta = [-1, 1]$.

Lemma 3. *Let $\Theta = [-1, 1]$ and $\eta > 1$. For $\theta \in \Theta$, let $Q_\theta \in \mathcal{P}(\mathbb{R})$ such that*

$$F_{Q_\theta}^{\leftarrow}(0) \;=\; b_0(\eta, \theta) = \begin{cases} \frac{1}{2}\left(\omega_0\big(2 - \theta(\eta^{-1} - 1)\big) + \theta\omega_1(\eta^{-1} - 1)\right) & \text{if } -1 \le \theta < 0, \\ \omega_0 & \text{if } \theta = 0, \\ \frac{1}{2}\left(\omega_0\big(2 + \theta(\eta - 1)\big) - \theta\omega_1(\eta - 1)\right) & \text{if } 0 < \theta \le 1, \end{cases}$$

$$F_{Q_\theta}^{\leftarrow}(1) \;=\; b_1(\eta, \theta) = \begin{cases} \frac{1}{2}\left(\omega_1\big(2 - \theta(\eta^{-1} - 1)\big) + \theta\omega_0(\eta^{-1} - 1)\right) & \text{if } -1 \le \theta < 0, \\ \omega_1 & \text{if } \theta = 0, \\ \frac{1}{2}\left(\omega_1\big(2 + \theta(\eta - 1)\big) - \theta\omega_0(\eta - 1)\right) & \text{if } 0 < \theta \le 1. \end{cases}$$

Then, $\forall (\theta, \eta) \in \Theta \times [1, \infty)$,

$$b_0(\eta, \theta) + b_1(\eta, \theta) \;=\; \omega_0 + \omega_1 \quad \textit{(midpoints equality)}$$

and

$$\begin{aligned} -1 \le \theta < 0 \quad &\Leftrightarrow \quad \frac{1}{\eta}\,\textit{diam}(\Omega_X) \le \textit{diam}(Q_\theta) < \textit{diam}(\Omega_X), \\ \theta = 0 \quad &\Leftrightarrow \quad \textit{diam}(Q_\theta) = \textit{diam}(\Omega_X), \\ 0 < \theta \le 1 \quad &\Leftrightarrow \quad \textit{diam}(\Omega_X) < \textit{diam}(Q_\theta) < \eta\,\textit{diam}(\Omega_X). \end{aligned}$$

We refer to Figure 4 (b.) for an illustration of this perturbation scheme. The initial application domain is displayed in magenta and is subject to a dilatation of parameter $\eta = 2$. The red constraints halve its width, and the blue constraints double it. One can additionally check that in both cases, the midpoint of the original validity domain is preserved. Section 5.2.1 showcases the usage of application domain dilatation perturbation in practice.

Given a perturbation class $\mathcal{Q}_{\mathcal{V}}$ defined over some modeling constraints, a collection $\mathcal{C}_{\mathcal{V}}(\theta)$ of perturbation classes driven by θ can be easily defined:

$$\mathcal{C}_{\mathcal{V}}(\theta) \quad = \quad \mathcal{Q}_{\mathcal{V}} \cap \mathcal{T}(\eta, \theta)$$

where $\mathcal{T}(\eta, \theta) = \left\{Q \in \mathcal{P}(\mathbb{R}) \mid F_Q^{\leftarrow}(0) = b_0(\eta, \theta), F_Q^{\leftarrow}(1) = b_1(\eta, \theta)\right\}$ in the case of application domain dilatation perturbations, or $\mathcal{T}(\boldsymbol{\eta}, \theta)$ as defined in (9) in the case of a quantile shift.

Many perturbation settings can be defined by combining quantile shifts and domain dilatations. However, for the sake of simplicity, quantile shifts and domain dilatations are studied independently in Section 5.

3.4. Perturbing multiple inputs

In the previous sections, perturbation classes have been defined *marginally*, i.e., on uni-dimensional probability measures. Whenever multiple inputs (or features) are involved, an independence assumption is often assumed, allowing to perturb marginal distributions independently, as proposed by [69]. While this hypothesis facilitate the interpretation and use of models, it is questionable in practice. Therefore, one of the main challenges in ML interpretability and SA is to account for the potential dependence structure between the inputs (or features) [88]. Dependencies can often provide useful information which must be preserved through a simultaneous perturbation (e.g., to maintain the plausibility of perturbed information and avoid creating meaningless patterns [14]). In ML, perturbation problems dealing with data privacy are thus particularly concerned by the preservation of dependencies [71, 84]. In UQ and ML frameworks, marginal quantile perturbations on P must obey this requirement: the dependence structure between the initial and perturbed distributions must remain the same.

Dependencies between random variables can be modeled using copula-based representations [77]. Let us recall the definition of a copula (or *dependence function*) $C_P : [0,1]^d \to [0,1]$. Let the random d-dimensional vector of inputs $X = (X_1, \dots, X_d)^\top \sim P$, where $P \in \mathcal{P}(\mathcal{X})$ and $\mathcal{X} \subseteq \mathbb{R}^d$. Assume that the marginal cdfs F_{P_i}, $i = 1, \dots, d$ are continuous. Denote $U_1, \dots, U_d$ the random variables defined as:

$$U_i = F_{P_i}(X_i)$$

and denote $U = (U_1, \dots, U_d)^\top \sim U_P$. For any $\mathbf{u} = (u_1, \dots, u_d) \in [0,1]^d$, denote $H_{\mathbf{u}} = \bigtimes_{i=1}^d [0, u_i]$. The copula of X, denoted C_p is defined as:

$$\begin{aligned} C_P(u) &= Pr\left(U_1 \leq u_1, \dots, U_d \leq u_d\right) \\ &= \int_{H_{\mathbf{u}}} dU_P \end{aligned}$$

We refer to the proof of Remark 2 for a proper definition of copula for empirical measures. In the following remark, we show that by means of particular monotone transportation maps inspired by optimal transportation theory, the marginal inputs can be optimally perturbed while preserving their copula. Moreover, this result is suitable for both ML and UQ applications since it also holds whenever P is an empirical measure related to an observed set of data points.

Remark 2. *Let $\mathcal{X} \subseteq \mathbb{R}^d$, for d a positive integer, and $P \in \mathcal{P}(\mathcal{X})$. Let Q_i be the solution of the optimal projection problem (2) with $\mathcal{C} = \mathcal{C}_{\mathcal{V}}$, for every marginal distribution P_i of P, $i = 1, \dots, d$, and where $\mathcal{V} \subseteq \otimes_{j=1}^d \mathcal{F}_j^{\leftarrow}$. Let the random vectors*

$$X \sim P, \quad \widetilde{X} := T(X)$$

where

$$T : \begin{array}{ccc} \mathcal{X} & \to & \mathcal{X} \\ \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} & \mapsto & \begin{pmatrix} T_1(x_1) \\ \vdots \\ T_d(x_d) \end{pmatrix} \end{array} \tag{10}$$

where

$$T_j = \left(F_{Q_j}^{\leftarrow} \circ F_{P_j} \right), \quad j = 1, \dots, d.$$

(i) *If P is an empirical measure (i.e., X represents a dataset), then X and the perturbed dataset $\widetilde{X}$ have the same empirical copula. Moreover, the empirical measure of every perturbed marginal sample $\widetilde{X}_i$ converges towards Q_i, $i = 1, \dots, d$.*

(ii) *If P is atomless, and assuming additionally that $\mathcal{V}$ is such that every $F_{Q_i}^{\leftarrow}$, $i = 1, \dots, d$ is strictly increasing, then the random vectors X and $\widetilde{X}$ have the same copula. Moreover, each perturbed marginal $\widetilde{X}_i \sim Q_i$.*

In other words, applying the perturbation map (2) to the inputs allows for preserving their copula and hence their dependence structure. If only an initial dataset is observed, applying T to every observation results in a perturbed dataset with, for instance, the same Spearman correlation matrix. Moreover, these transportation maps achieve optimality for various univariate transportation costs (see the proof). Moreover, in the UQ framework, sampling from the perturbed inputs is as simple as applying these maps to simulated samples of P, which can naturally benefit from the large literature available on copula. However, it requires that the marginal gqfs $\{F_{Q_j}^{\leftarrow}\}_{1 \leq j \leq d}$ are accessible, which obviously depend on the projection problem (2). As shown in the next sections, and among many other practical and theoretical motivations, the particular choice of the 2-Wasserstein distance as a projection metric allows for characterizing the optimally perturbed marginal probability measures through their quantile functions and thus simplifies their accessibility.

4. Wasserstein projections

This section motivates the choice of the 2-Wasserstein distance as a projection metric for the perturbation problem (2). This distance is deeply rooted in optimal transportation theory [112] and has been used successfully in many ML and deep learning applications [40, 3]. It has also been extensively studied as a tool for guaranteeing distributional robustness to adversarial attacks in ML [32]. In SA, it has been used to produce novel sensitivity indices [38, 17].

Using the Wasserstein distance imposes to work on the subset $\mathcal{P}_p(\mathbb{R}^d) \subset \mathcal{P}(\mathbb{R}^d)$ of probability measures with finite p-th moment. The Wasserstein distance between multi-dimensional probability measures can be computationally expensive to evaluate [87]. However, as stated in Section 3.4, the fact that one wishes to preserve the copula between P and its optimally perturbed counterpart Q greatly simplifies the projection problem. Let $P, Q \in \mathcal{P}_p(\mathbb{R}^d)$ be two multi-dimensional probability measures, with marginals $P_1, \dots, P_d$ and $Q_1, \dots, Q_d$ respectively. Leveraging the work in [1], if P and Q share the same copula, one can rewrite their 2-Wasserstein distance as:

$$W_p^p(P, Q) = \sum_{i=1}^{d} W_p^p(P_i, Q_i). \tag{11}$$

Hence, finding Q that minimizes $W_p^p(P, Q)$ under constraints on the marginals $Q_1, \dots, Q_d$ is equivalent to solving d independent univariate projection problems with the p-Wasserstein distance as a projection metric. Furthermore, the transportation map defined in (10) is indeed optimal [1] (but not unique), in the sense that it minimizes (11).

In other words, finding Q that minimizes (11) such that $C_P = C_Q$ is equivalent to projecting each P_i under its relevant constraints, and applying the transportation map (10). As the problem reduces to perturbations of univariate marginal distributions, the $p-$Wasserstein distance may be easily computed, as recalled in the following definition.

Definition 4 (Wasserstein distance on the real line)**.** *Let $p \in \mathbb{N}^*$ and $P, Q \in \mathcal{P}_p(\mathbb{R})$ be two probability measures on $\mathbb{R}$ admitting F_P and F_Q as probability distribution functions, respectively. Then, the p-Wasserstein distance between P and Q is:*

$$W_p(P,Q) = \left(\int_0^1 \left| F_P^{\rightarrow}(x) - F_Q^{\rightarrow}(x) \right|^p dx \right)^{1/p}$$

where $F_P^{\rightarrow}$ (resp. $F_Q^{\rightarrow}$) has been defined in Definition 1.

The following subsections argue on the specific choice of the 2-Wasserstein distance. First, we highlight its attractive properties for conducting robust interpretability analyses. Then we investigate the solution of the perturbation problem (2), with and without regularity constraints, the latter enforced using isotonic, piece-wise polynomial approximations.

4.1. The 2-Wasserstein distance as a suitable perturbation discrepancy

The special choice of the 2-Wasserstein distance

$$W_2(P,Q) = \sqrt{\int_0^1 \left(F_P^{\rightarrow}(x) - F_Q^{\rightarrow}(x) \right)^2 dx}, \quad P, Q \in \mathcal{P}_2(\mathbb{R})$$

to instantiate the perturbation problem (2) is based on the following rationale.

First, W_2 metricizes weak convergence on $\mathcal{P}_2(\mathbb{R})$ ([112], Section 7.2). It means that W_2 is a measure of proximity on a broad set of probability measures. In other words, for any $P \in \mathcal{P}_2(\mathbb{R})$, and a sequence of probability measures $(Q_n)_{n\in\mathbb{N}^*} \in \mathcal{P}_2(\mathbb{R})$:

$$W_2(P,Q_n) \underset{n\to\infty}{\to} 0 \quad \Rightarrow \quad Q_n \xrightarrow{\text{d}} P$$

where $\xrightarrow{\text{d}}$ denotes the convergence in distribution (or weak convergence). That is, W_2 allows for assessing the point-wise proximity between two probability measures, as long as both admit finite second-order moments (a current assumption in both SA and ML fields). Contrary to the KL divergence, no additional conditions on the absolute continuity of Q_n w.r.t. P are needed. When it comes to the perturbation problem (2), two practical advantages in favor of the 2-Wasserstein distance can be drawn, compared to entropic projections (i.e., using the KL divergence): if P is an empirical measure (i.e., purely atomic), then Q is not restricted to be purely atomic; conversely, if P admits a density, then it does not restrict Q to admit a density.

These benefits are key in unifying the frameworks of SA and ML interpretability: the flexibility of W_2 allows for greater explicit control (e.g., through smoothing restriction) on the nature of Q, independently of that of P. Results of entropic projections entail a re-weighting of the atoms when P is empirical [6], or a perturbed density of Q proportional to the one of P when it is absolutely continuous w.r.t. the Lebesgue measure [69]. Using W_2 allows, for instance, to add additional atoms, or to allow Q to admit a density, independently of the regularity of P.

Second, W_2 facilitates the projection of P onto a quantile perturbation class $C_{\mathcal{V}}(\theta)$. Indeed, the next proposition shows that the optimization problem is equivalent to a projection in L^2 of its gqf onto $\mathcal{V}$ under interpolation constraints

Proposition 1. *Let $P \in \mathcal{P}_2(\mathbb{R})$, and $\mathcal{C}_{\mathcal{V}}(\theta)$ be a non-empty perturbation class defined by a subset $\mathcal{V} \subseteq \mathcal{F}^{\leftarrow}$. Consider for $\mathcal{V}$ the constraints system defined in § 3.2-3.3 associated with couples $(\alpha_i, b_i)_{1\leq i\leq K}$. In this frame, the solution Q of the perturbation problem (2), i.e.,*

$$Q = \underset{G\in\mathcal{P}_2(\mathbb{R})}{\text{argmin}}\, W_2(P,G) \quad s.t. \quad G \in \mathcal{C}_{\mathcal{V}}(\theta) \tag{12}$$

is characterized as the unique Lebesgue-Stieljes measure induced by the cdf F_Q with gqf $F_Q^{\leftarrow} \in \mathcal{F}^{\leftarrow}$:

$$
\begin{aligned}
F_Q^{\leftarrow} = \underset{L \in L^2([0,1])}{\operatorname{argmin}} & \left\{ \int_0^1 \left(L(x) - F_P^{\rightarrow}(x)\right)^2 \right\} \\
& s.t. \quad L(\alpha_i) \leq b_i \leq L\left(\alpha_i^+\right), \quad i = 1, \dots, K, \\
& \quad\; L \in \mathcal{V}.
\end{aligned}
\tag{13}
$$

The equivalent projection problem in (13) echoes with the result in Proposition 1. Projecting a measure w.r.t. the 2-Wasserstein distance is equivalent to projecting its gqf in L^2. One can then leverage the vast literature on function approximations in L^2, especially on monotonic approximations. Moreover, as eluded at the end of Section 3, the proposed perturbation scheme depends heavily on the knowledge of the gqf of Q. Solving (13) grants direct access to the gqf of Q, and thus allows applying perturbations fairly easily.

4.2. Quantile constrained Wasserstein projections

This subsection presents and discusses the main results of this paper. If no smoothing constraints on $F_Q^{\leftarrow} \in \mathcal{F}^{\leftarrow}$ is enforced, the perturbation problem (13) leads to a unique analytical solution. However, many studies require $F_Q^{\leftarrow}$ to be smooth (e.g., continuous). We propose enforce continuity by using isotonic interpolating piece-wise polynomials, which lead to a well-defined convex constrained quadratic program, easily solvable in practice.

4.2.1. Analytical solution without smoothing restrictions

The following proposition provides a convenient way to solve the perturbation problem (13) in the case when no smoothing constraint on $F_Q^{\leftarrow}$ is enforced.

Proposition 2. *Let P be a probability measure in $\mathcal{P}_2(\mathbb{R})$. Let $\mathcal{C}$ be a non-empty perturbation class characterized by a set of K quantile constraints. Assume, without loss of generality, for $i = 1, \dots, K$, that $\alpha_1 < \cdots < \alpha_K$ along with $b_1 < \cdots < b_K$. Let $\beta_i = F_P(b_i)$ for $i = 1, \dots, K$. Define the intervals $A_i = (c_i, d_i]$ for $i = 1, \dots, K$, such that:*

$$
\begin{aligned}
c_1 = \min(\beta_1, \alpha_1), \quad c_i = \min\Big[\max(\alpha_{i-1}, \beta_i), \alpha_i\Big], i = 2, \dots, K, \\
d_K = \max(\beta_K, \alpha_K), \quad d_j = \max\Big[\min(\beta_j, \alpha_{j+1}), \alpha_j\Big], j = 1, \dots, K-1.
\end{aligned}
$$

Let $A = \bigcup_{i=1}^K A_i$ and $\overline{A} = [0,1] \setminus A$. Then the problem (13) has a unique solution which can be written as, for any $y \in [0,1]$:

$$
F_Q^{\leftarrow}(y) = \begin{cases} F_P^{\rightarrow}(y) & \text{if } y \in \overline{A}, \\ b_i & \text{if } y \in A_i, \quad i = 1, \dots, K. \end{cases}
\tag{14}
$$

In order to interpret this result, illustrated in Figure 5, let us recall that when a quantile function is constant on an interval, it implies that its related probability measure admits an atom at the constant value taken by the gqf. Moreover, the mass allocated to this atom is equal to the length of the interval. Additionally, each jump of the quantile function induces an interval with no mass. The solution displayed in (14) shows that on $\overline{A}$, both initial and perturbed quantile functions are equal. However, they differ on every interval A_i in the following fashion:

- Q have atoms at each constraint point b_i, $i = 1, \dots, K$;
- Each of these atoms have mass $Q(\{b_i\}) = d_i - c_i$, for $i = 1, \dots, K$;

- Each open interval $I_i \subset \mathbb{R}$ defined as

$$I_i = \begin{cases} \Big(\max(F_P^{\leftarrow}(\alpha_i), b_{i-1}), b_i\Big), & \text{when } b_i > F_P^{\leftarrow}(\alpha_i), \\ \Big(b_i, \min\left(b_{i+1}, F_P^{\leftarrow}(\alpha_i)\right)\Big), & \text{when } b_i < F_P^{\leftarrow}(\alpha_i) \end{cases} \tag{15}$$

  with, by convention, $b_0 = -\infty$ and $b_{K+1} = \infty$, has no mass. To put it briefly, $Q(I_i) = 0$ for every $i = 1, \ldots, K$.

In other words, whenever an α-quantile p_α is shifted up to a value b, the perturbation entails sending every possible values in the range (p_α, b) to b. Hence, every value in (p_α, b) cannot be sampled according to Q. Moreover, the singleton $\{b\}$ now admits a probability of being observed equal to the initial probability of this interval, i.e., $Q(\{b\}) = P\big((p_\alpha, b)\big)$. When an α-quantile is shifted down to b, the interval becomes (b, p_α), and the same reasoning can be done.

The statement of Proposition 2 is rather intuitive. Indeed, the Wasserstein distance quantifies the amount of *work* needed to transform a probability measure into another one [96]. When using W_2, the amount of work is quantified using the Euclidian distance, i.e., transporting a point x_0 to x_1 requires $(x_0 - x_1)^2$ units of work. This intrinsic *point-wise* way of quantifying similarities can be sensed in the previous result: perturbing an α-quantile entails giving the initial mass of an interval adjacent to b to the singleton $\{b\}$ in order to satisfy the constraint.

4.2.2. Projection solution under smoothness restrictions

The analytical solution provided in Proposition 2 presents a significant drawback: part of the application domain Ω_X of the perturbed input receives no mass. From a practical standpoint, it cannot be explored in a robustness study. It is because $\mathcal{F}^{\leftarrow}$ contains discontinuous functions. Ensuring a smoother solution relies on specifying a smooth perturbation class $\mathcal{Q}_\mathcal{V}$ where $\mathcal{V}$ is a set of continuous, non-decreasing functions. To solve the perturbation problem, we can take advantage of its $L^2([0,1])$ equivalent formulation 13) to project $F_P^{\leftarrow}$ onto $\mathcal{V}$. However, the main challenge is to ensure that $\mathcal{V}$ only contains monotonic functions.

We propose to projection $F_P^{\leftarrow}$ onto a space of piece-wise continuous polynomials. It implies that the support of Q must be bounded. These bounds are made explicit using extremal quantile constraints (i.e., $F_Q^{\leftarrow}(0)$ and $F_Q^{\leftarrow}(1)$ are constrained to take finite values).

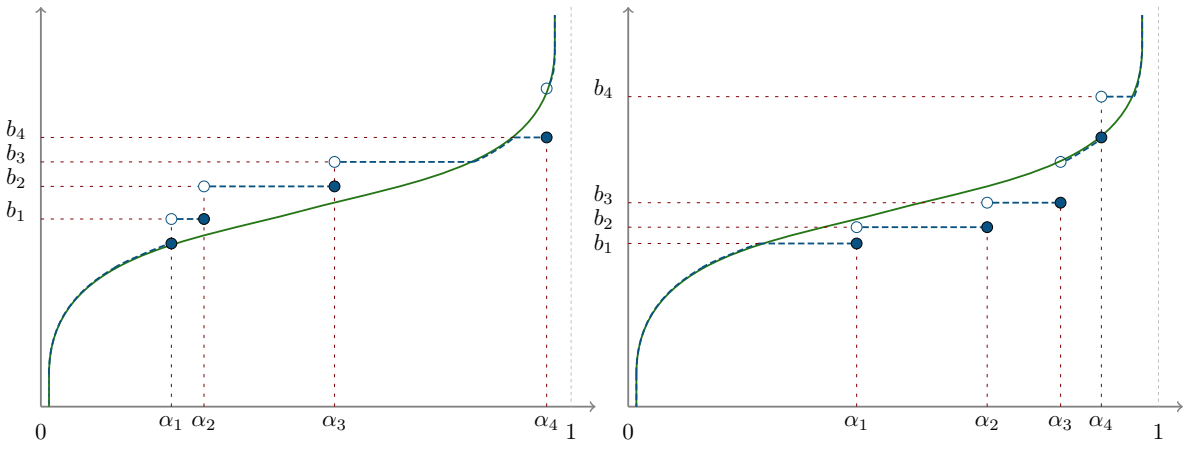

Figure 5: Characterizing quantile function of the solution of the perturbation problem (dashed blue). The initial quantile function (i.e., $F_P^{\leftarrow}$) is displayed in green, and dashed red lines identify the quantile constraints. (a.) and (b.) illustrate different possible perturbation configurations, increasing or decreasing several initial quantile values.

Formally, we aim to find a piece-wise polynomial of the form

$$G(x) = \begin{cases} G_0(x) & \text{if } \alpha_0 := 0 \leq x < \alpha_1, \\ \vdots & \\ G_i(x) & \text{if } \alpha_i \leq x < \alpha_{i+1}, \\ \vdots & \\ G_K(x) & \text{if } \alpha_K \leq x \leq 1 =: \alpha_{K+1}. \end{cases} \tag{16}$$

under the continuity constraints at each knot of the grid $\alpha_1 < \cdots < \alpha_K$

$$G_i(\alpha_{i+1}) = G_{i+1}(\alpha_{i+1}), \quad i = 0, \ldots, K-1.$$

Here, each $G_j \in \mathbb{R}[x]_{\leq p}$, for $j = 0, \ldots, K$ where we recall that $\mathbb{R}[x]_{\leq p}$ denotes the set of all real polynomials of degree at most equal to p. Let $\mathcal{S}_p$ denote the space of functions defined by (16). Restricting the solution of the perturbation problem (13) leads to the following optimization problem

$$\begin{aligned} F_Q^{\leftarrow} = \underset{L \in L^2([0,1])}{\operatorname{argmin}} & \left\{ \int_0^1 \left(L(x) - F_P^{\rightarrow}(x)\right)^2 dx \right\} \\ \text{s.t.} \quad & L(\alpha_i) = b_i, \quad i = 1, \ldots, K, \\ & L \in \mathcal{F}^{\leftarrow} \cap \mathcal{S}_p. \end{aligned} \tag{17}$$

Hence, the smooth perturbation class is defined by $\mathcal{V} = \mathcal{F}^{\leftarrow} \cap \mathcal{S}_p$. Since the design enforced the polynomials in $\mathcal{S}_p$ to be defined on the grid $\alpha_0 < \alpha_1 < \cdots < \alpha_K < \alpha_{K+1} = 1$, solving (17) reduces to solve several sub-problems on each sub-interval $[\alpha_i, \alpha_{i+1}], i = 0, \ldots, K$ of $[0, 1]$. The optimization problem is indeed separable into $K+1$ independent optimization sub-problems. Each of them defines an optimal component G_i of the piece-wise polynomial G as defined in (16).

Any of these problems can be formulated generically as follows. Let $[t_0, t_1] \subset [0, 1]$, and $z_0, z_1 \in \mathbb{R}$ be interpolation values at t_0 and t_1 respectively. We aim to find the solution to the optimization sub-problem

$$\begin{aligned} S = \underset{L \in \mathbb{R}[x]_{\leq p}}{\operatorname{argmin}} & \left\{ \int_{t_0}^{t_1} (F_P^{\leftarrow}(x) - L(x))^2 dx \right\} \\ \text{s.t.} \; & L(t_0) = z_0, L(t_1) = z_1, \\ & L'(x) \geq 0, \quad \forall x \in [t_0, t_1]. \end{aligned} \tag{18}$$

This optimization sub-problem is nothing more than the L^2 isotonic (i.e., monotonic, in this case non-decreasing) polynomial approximation on a compact interval [75], with interpolation constraints at the boundaries. The interpolating polynomials have been extensively studied in the literature [39], as well as isotonic polynomial regression and approximation [97, 115]. However, to our knowledge, this specific optimization problem does not seem to have been particularly studied.

We propose to solve (18) using the *sum-of-squares* (SOS) polynomials [68] representation of non negative polynomials. Furthermore, we leverage in particular the representation of SOS polynomials using semi-definite positive (SDP) matrices [82, 83, 101]. A similar characterization of isotonic polynomials has been proposed in [101]. The following result, the second main outcome of this article, shows that the problem to solve falls into the category of strictly convex programs: the solution in (21) is unique [15].

Theorem 1. *Let $[t_0, t_1] \subset [0, 1]$. Let M be the symmetric positive definite $((d+1) \times (d+1))$ moment matrix of the Lebesgue measure on $[t_0, t_1]$, i.e. for $i, j = 1, \ldots, d+1$,*

$$M_{ij} = \int_{t_0}^{t_1} x^{i+j-2} dx = \frac{(t_1)^{i+j-1} - (t_0)^{i+j-1}}{i+j-1}, \tag{19}$$

and denote $r \in \mathbb{R}^{d+1}$ *the moment vector of* $F_P^{\rightarrow}(x)$*, i.e., for* $i = 0, \dots, d$

$$r_i = \int_{t_0}^{t_1} x^i F_P^{\rightarrow}(x) dx. \tag{20}$$

Then, the vector $s^* = (s_0, \dots, s_d)^\top \in \mathbb{R}^{d+1}$ *of coefficients characterizing the polynomial* S *in (18) is the solution of the following convex constrained quadratic program*

$$\begin{aligned} s^* = \underset{s \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \; & s^\top M s - 2 s^\top r \\ s.t. \; & s \in \mathcal{K}, \end{aligned} \tag{21}$$

where $\mathcal{K}$ *is an identifiable closed convex subset of* $\mathbb{R}^{p+1}$ *(for the sake of conciseness,* $\mathcal{K}$ *is characterized within the proof).*

As the computation of s^* is a convex constrained quadratic program, it can be addressed efficiently using devoted solvers. The problem (17) can be addressed by solving $K+1$ optimization problems of the form (21). Furthermore, computations can be done in parallel, leading to fast computational times. Notice that (21) can be formulated and solved using `CVXR`. This is an `R` package for disciplined convex programming [41]. The pretty generic low-level logic behind the optimization scheme can be found in Algorithm 1.

Algorithm 1 Isotonic interpolating piece-wise continuous polynomial optimization strategy

Require: α, b, $F_P^{\rightarrow}$, p
1: **for** $i = 0, \dots, K$ **do** (in parallel)
2: Compute M on $[\alpha_i, \alpha_{i+1}]$ (19).
3: Compute r on $[\alpha_i, \alpha_{i+1}]$ (20).
4: Setup `CVXR` constraints.
5: $s^{(i)} \leftarrow$ Solve (21).
6: $G_i(x) \leftarrow \sum_{j=0}^{p} s_j^{(i)} x^j$
7: **end for**
8: **return** $G(x) \leftarrow \sum_{i=0}^{K} G_i(x) \mathbb{1}_{[\alpha_i, \alpha_{i+1}]}(x)$

While computing the Lebesgue moment matrix M on each sub-interval of $[0,1]$ is straightforward, computing strategies for r, the moment vector of $F_P^{\leftarrow}$, can vary depending on the nature of P. Additional computational details are given in Appendix Appendix B. The set-up of the `CVXR` constraints is detailed in the accompanying GitLab repository[1].

To provide a frame of reference for the practical usage of our method, the empirical computational time of solving one element of G, w.r.t. the polynomial degree is studied as follows. Values $t_0, t_1 \in [0,1]$, and $z_0, z_1 \in \Omega_X$ are randomly selected, and an isotonic interpolating piece-wise continuous polynomial is fitted (i.e., solving (21)). Polynomials of degrees ranging from 2 to 50 are fitted for each experiment, repeated 150 times. The execution time has been recorded and is displayed in Figure 6. One can notice that the mean computational time seems to be linear w.r.t. the polynomial degree. However, the higher the degree, the wider the 90% time coverage seems to be, which may be caused by the complexity of the underlying optimization problem. In our limited testing, further numerical experiments showed that small polynomial degrees (≤ 7) often appear sufficient to obtain good approximations and that the approximation error tends to stabilize, w.r.t. the polynomial degree, rather rapidly.

Remark 3. *The numerical solver used is* `SCS V3.2.1` *[79]. To improve numerical stability, the quantile functions have been mapped to take values between* $[-1, 1]$*. All the figures and all obtained optimal perturbations have been computed by performing this pre-processing step first.*

[1] https://gitlab.com/milidris/qcWasserteinProj

5. Robustness diagnostics to distributional perturbations

We illustrate the previous method on two use cases. First, in an ML context, we assess the robustness to feature perturbations of a classification model (i.e., a one-layer neural network) trained on an acoustic fire extinguisher dataset. Then, in the UQ framework, we extend Example 2, studying the impact of input perturbation on the output of a numerical hydrological model predicting a river water level.

Remark 4. *In the following applications, isotonic polynomial smoothing is applied with an arbitrarily high degree. We chose the degree based on an empirical inspection of the solutions and by ensuring that the approximation error remains relatively the same w.r.t. higher degrees.*

Our method is in line with the general SIPA (Sampling, Intervention, Prediction, Aggregation) framework for model-agnostic interpretation proposed in [98]. More precisely, the four step can be broken down as follows:

1. **Sampling**: In Section 5.1, we have access to an i.i.d. sample of the inputs. In Section 5.2, a sampling strategy is used to simulate data.
2. **Intervention**: In both use-cases, we define meaningful quantile perturbations, and solve the subsequent perturbation problem. Then we apply the transformation in (10), resulting in perturbed samples.
3. **Prediction**: We predict using the available black-box model (a neural network in 5.1, and a numerical model in Section 5.2), resulting in perturbed outputs.
4. **Aggregation**: The resulting perturbed outputs are aggregated w.r.t. the perturbation intensity, leading to global robustness metrics, or are simply plotted against the initial and perturbed of the perturbed input values, allowing for local robustness assessments.

5.1. ML application: Acoustic fire extinguisher dataset

The acoustic fire extinguisher dataset is composed of 15390 experiments of fire extinguishing tests of three different liquid fire fuels. Amplified subwoofers are placed in a collimator with an opening.

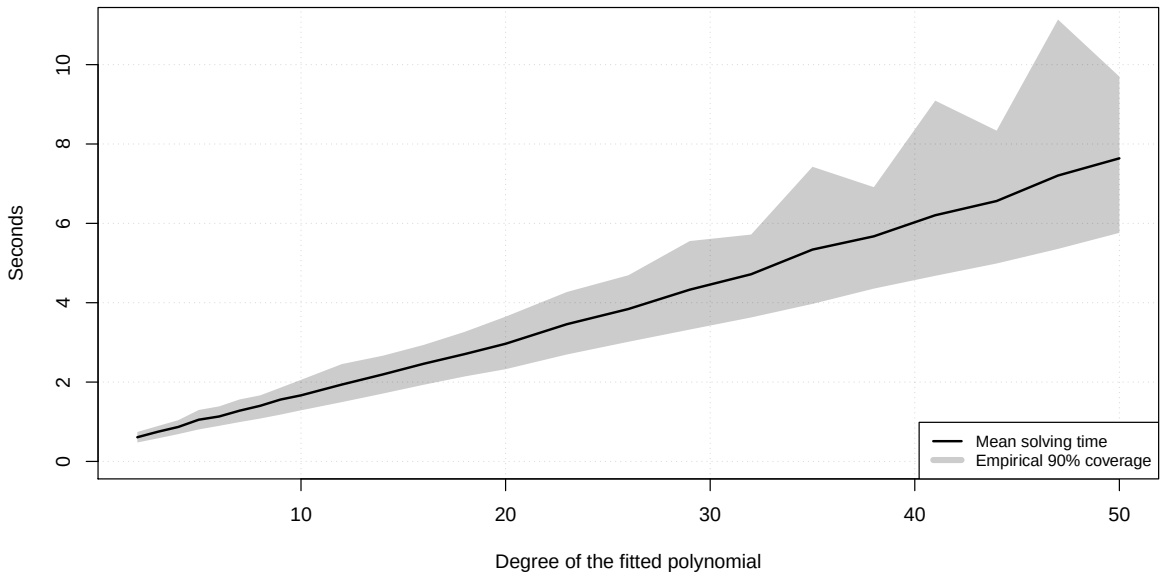

Figure 6: Computational solving time in seconds of the optimization problem (21) using `CVXR`, w.r.t. the chosen degree of the polynomial.

When activated at different frequencies, the acoustic waves produce an escape of air through the opening, which is used to extinguish fires. Three features are set using a design of experiment (DoE), and two are measured using appropriate equipment. For more details on the experiments settings, one can refer to the in-depth descriptions in [65, 107]. Table 1 gives additional details on the nature of the features.

Feature	Unit	Mode of measure	Description
TankSize	cm	DoE	Discrete feature (5 levels) describing the size of the tank containing the fuel.
Fuel		DoE	Type of fuel used (3 levels: Gasoline, Kerosene, Thinner).
Distance	cm	DoE	Distance of the flame to the collimator opening.
Frequency	Hz	DoE	Sound frequency range.
Decibel	dB	Measured	Sound pressure level.
Airflow	m/s	Measured	Airflow created by the sound waves.

Table 1: Description of the features of the acoustic fire extinguisher dataset.

For each experiment, a binary output variable Y is measured, representing the result of the experiment, i.e., whether the fire has been put out ($Y = 1$) or not ($Y = 0$). The two output classes are relatively balanced (i.e., 48.97% of the observations describe effectively put out fires). The distribution, correlation structure, and relationship of the continuous features with the output are represented in Figure 7. Some variables seem fairly correlated (in Spearman's sense, i.e., the linear correlation of the rank-transformed data), such as Frequency and Decibel, as well as Distance and Airflow.

The classification black-box model is a one-layer neural network (composed of 100 neurons), trained on 500 epochs, with a learning rate of 10^{-4}, similar to the study conducted in [106]. 5% of the data has been randomly selected to serve as validation data. The model resulted in a good prediction accuracy: 95.15% of the training data and 94.26% of the validation data are correctly classified. Figure 8 depicts the ROC curve and confusion matrix of the trained black-box model. The model's predictive performance can be validated globally with an AUC of 0.992 and less than 3% of type 1 and 2 prediction errors.

However, global predictive performance only focuses on effectively observed data points. It is mandatory to study the model's behavior on predictions outside of these points to improve confidence in its usage. Hence, one can be interested in the robustness of the model w.r.t. perturbations on its inputs. Note that ground truths cannot be observed for perturbed data. However, the impact, either globally or locally, of these perturbations on the predictive behavior of the model can still be assessed using predictions on the perturbed data. In the following, the feature perturbation scheme is detailed and motivated, and then the model's behavior is studied under these perturbations.

5.1.1. Perturbation strategy

We propose a straightforward perturbation strategy. Only the Airflow feature is perturbed. The perturbation is composed of the $K = 14$ constraints:

- The application domain of the feature is preserved by setting both the 0 and 1-quantiles to the minimum and the maximum observed value of the dataset.

- The left tail of the distribution is preserved by constraining every quantile of level 0.1 to 0.6 with a step of 0.05 to interpolate the empirical quantile function of the feature.
- A quantile shift perturbation is put on the 0.8-quantile of the feature, with an initial value of $F_P^{\leftarrow}(0.8) = 12$, being shifted between 9.5 ($\theta = -1$) and 14.5 ($\theta = 1$).

Additionally to these perturbations, smoothness is enforced using piece-wise continuous isotonic polynomials, as described in Section 4.2.2. The degree of each monotone polynomial has been arbitrarily chosen to be up to 9. The constraints and the resulting quantile-constrained Wasserstein projections are illustrated in Figure 9 for intensity values -1, 0, and 1.

The perturbed quantile level has been chosen in relation to the model's decision boundary: no observation in the initial dataset with an Airflow value exceeding 12.3m/s is classified by the model as not extinguishing the fire, regardless of the values taken by the other features. Perturbing the 0.8-quantile of the Airflow variable allows for exploring the model's behavior in regions close to this decision boundary. More importantly, it allows assessing the predictive robustness of the neural network in this region under perturbations of varying magnitude. Generally, this quantile shift regime can be understood as a perturbation on the right tail of the initial distribution, i.e., on values higher than the 0.6-quantile.

5.1.2. Model robustness assessment

First, we are interested in assessing the robustness of the neural network model in a global fashion. The left plot of Figure 10 presents the proportion of perturbed observations with predictions of 1

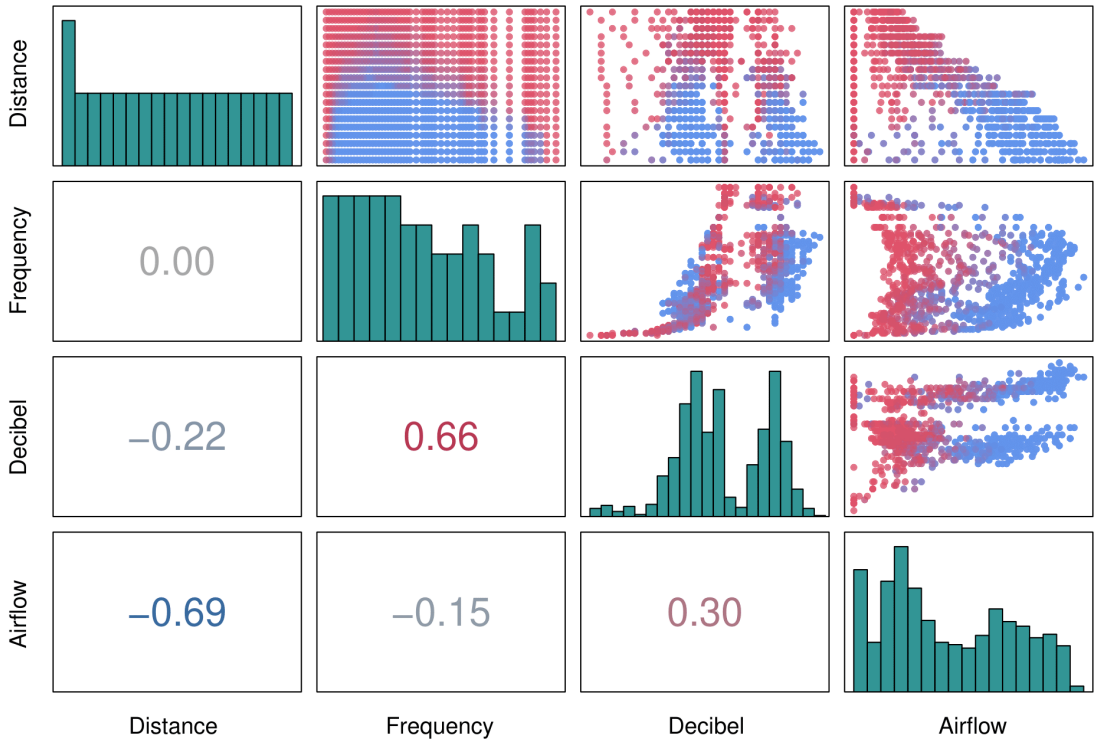

Figure 7: Histogram, cross-scatterplot, and Spearman's correlation coefficient of the input features. Red dots represent observations resulting in $Y = 0$, and blue dots are observations resulting in $Y = 1$.

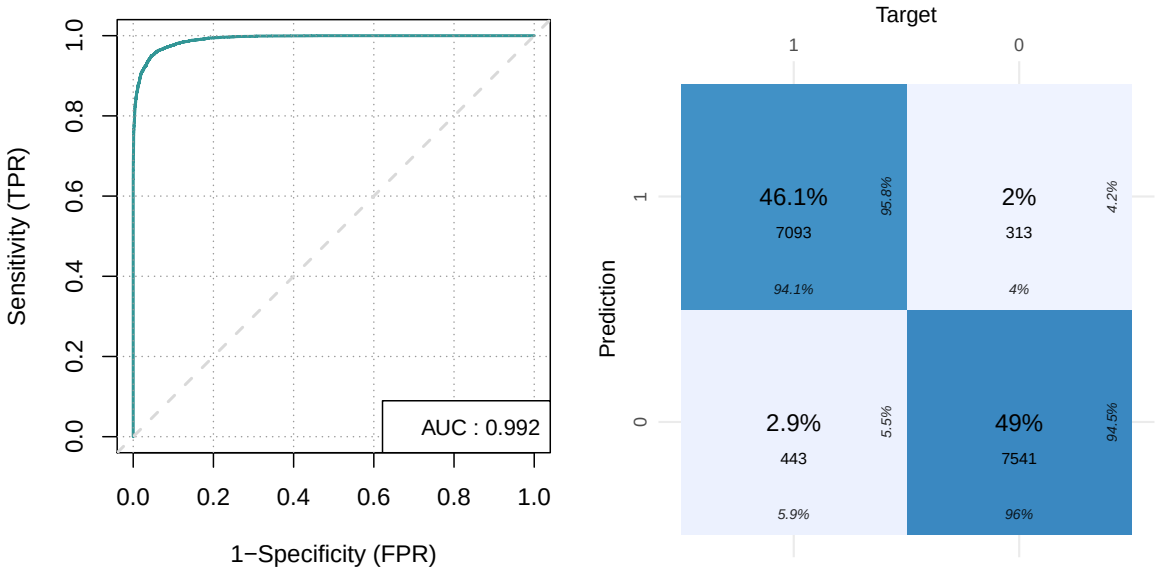

Figure 8: ROC curve (left) and confusion matrix (right) of the neural network model trained on the acoustic fire extinguisher dataset.

w.r.t. to the intensity of the perturbation. Notice that the proportion is increasing, along with θ. Hence, decreasing the value of the initial 0.8-quantile tend to result in a lower number of predicted put-out fires, and increasing its value results in an increasing number of predicted put-out fires. This interpretation is rather intuitive: all other things being equal, a higher Airflow value entails a higher chance of predicting $Y = 1$. The right plot of Figure 10 presents the proportion of prediction shift with respect to θ. Notice that the higher the magnitude of the perturbation (either positively or negatively), the more predictions tend to change, and the closest θ is to 0, the fewer predictions shift. This observation informs on the predictive stability in the vicinity of the decision boundary of the model: small perturbations tend to result in less prediction shift than bigger perturbations.

Second, we study the robustness of global SA results. Figure 11 presents the target Shapley effects [59], a global SA input importance measure for binary black-box model outputs with dependent inputs, w.r.t. the perturbation intensity parameter θ. These indices have been computed using the nearest-neighbor (KNN) approach proposed in [18] (with an arbitrarily chosen number of neighbors equal to 6). Recall that our perturbation method allows preserving the empirical copula between the features, justifying the use of the KNN approach. Studying the behavior of importance measures allows to verify if the feature importance order shifts due to the perturbations, i.e., if the importance hierarchy between the inputs changes due to perturbations around the model's decision boundary. The left barplot presents the initial target Shapley effects, computed on the model's prediction on the observed data, and the right plot presents their behavior under the airflow perturbation. One can notice that the importance indices remain stable w.r.t. θ. This result indicates that the global SA of the neural network is robust to the distributional perturbations driven by θ. Hence, we can be confident in the importance measures under uncertainties in the region near the model's decision boundary.

Finally, the robustness of the neural network can also be assessed locally. Figure 12 allow visualizing whether a prediction has shifted w.r.t. to the effective magnitude of the perturbation. The black line indicates no perturbation change: the airflow value of an observation has been mapped to itself. For a fixed initial airflow datapoint, its vertical distance to the black line indicates the (signed) magnitude of the applied perturbation. Red points indicate that the prediction has shifted w.r.t. the initial dataset, and blue points indicate no predictive change. One can note the presence of red dots close to the black line around the prediction boundary of the model. Small perturbations for observations with airflow

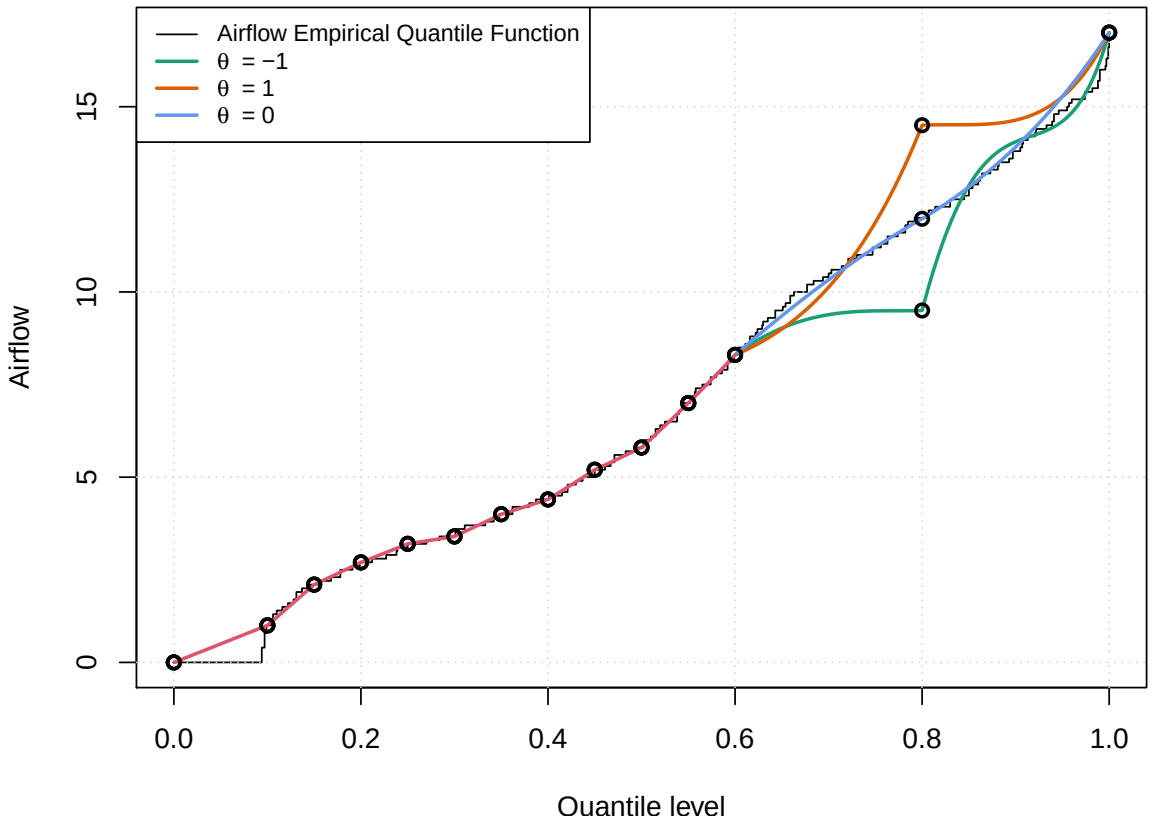

Figure 9: Quantile functions of the optimally perturbed Airflow feature, with a chosen polynomial degree equal to 9. The red line represents the preserved tail; meanwhile, the green, blue and yellow lines represent various quantile shift intensity levels ($\theta = -1$, $\theta = 0$, and $\theta = 1$, respectively).

values around 12, all other features being equal, can lead to a prediction change. Hence, the confidence in predictions on observations in this region can be questioned. However, notice the lack of red dots near the black line for airflow values on the interval $[13, 17]$ and on the interval $[7, 10]$. Hence, we can be confident in the model's predictions for Airflow values on these intervals, which seem to be robust w.r.t. the quantile shift.

One may notice the presence of small perturbation resulting in prediction changes for small airflow values. However, since the perturbation scheme focuses on exploring the model's behavior around its Airflow decision boundary, their interpretation is voluntarily omitted: a different perturbation scheme involving perturbing the left tail of the airflow distribution would be advised.

In summary, besides the model's good prediction accuracy, it also seems globally robust to distributional perturbation focused around its decision boundary. Moreover, we can be confident in the feature importance indices since they remain relatively similar under perturbation. Locally, the model prediction seems stable w.r.t. small perturbations, except on a small interval around its decision boundary (a behavior generally expected in ML applications). In conclusion, this robust interpretability analysis further assesses the model's behavior beyond classical accuracy metrics and provides additional arguments for its validation.

5.2. SA application: Simplified hydrological model

This use-case focuses on a simplified model of the water level of a river. This model has been extensively used in the safety and reliability of industrial sites, where the occurrence of a flood can lead to dramatic human and ecological consequences. It consists of a substantial simplification of the one-dimensional Saint-Venant equation, with a uniform and constant flow rate, inspired from [62, 42]. The maximal annual water level from sea level is modeled as:

$$Y = Z_v + \left(\frac{Q}{BK_s \sqrt{\frac{Z_m - Z_v}{L}}} \right)^{3/5}$$

where the description of each input variable and their explicit marginal probabilistic structure is detailed in Table 2.

Additionally, similarly to [21], a dependence structure is modeled using a Gaussian copula, with the covariance matrix

$$R_P = \begin{pmatrix} 1 & 0.5 & 0 & 0 & 0 & 0 \\ 0.5 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.3 & 0 & 0 \\ 0 & 0 & 0.3 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0.3 \\ 0 & 0 & 0 & 0 & 0.3 & 1 \end{pmatrix}, \quad \text{where} \quad \begin{pmatrix} Q \\ K_s \\ Z_v \\ Z_m \\ L \\ B \end{pmatrix} \sim P.$$

Echoing Example 2, we are interested in uncertainties on the application domain of the K_s input, i.e., the Strickler riverbed roughness coefficient (which is the inverse of the Manning coefficient). Its value can range from around 3 (proliferating algae) to around 90 (smooth concrete). We refer the interested reader to the in-depth study in [42] for more details on the determination and inference of the Strickler coefficient for realistic rivers. In this use-case, initially, the application domain Ω_X of the Strickler coefficient is set between the values of 20 and 50, corresponding to situations from

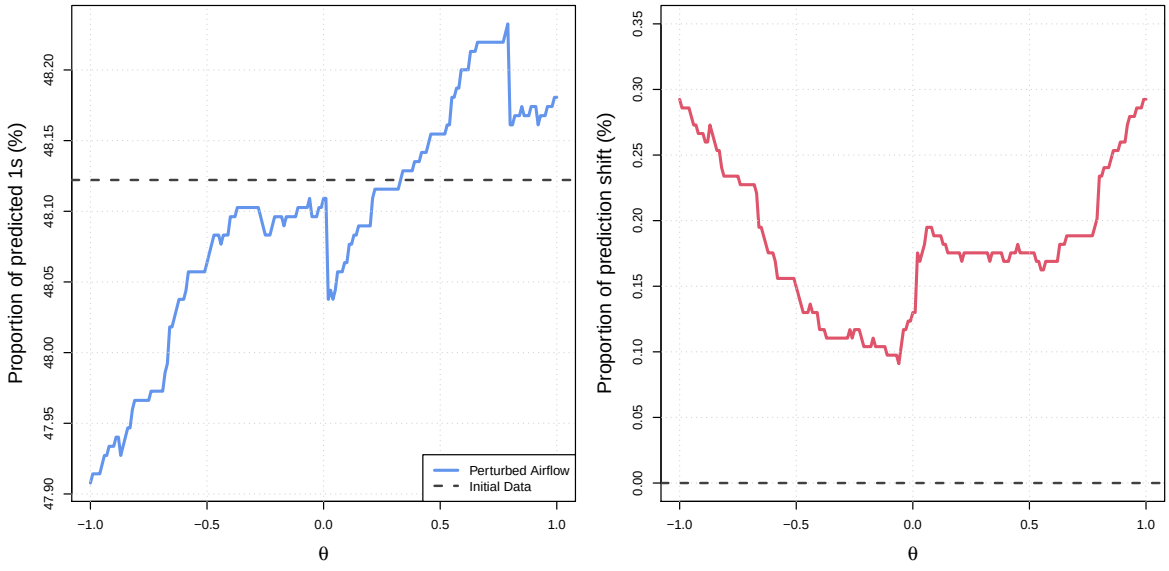

Figure 10: Proportion of predictions $Y = 1$ (left) and proportion of classification prediction shift (right) compared to the initial data, w.r.t. the perturbation intensity parameter θ.

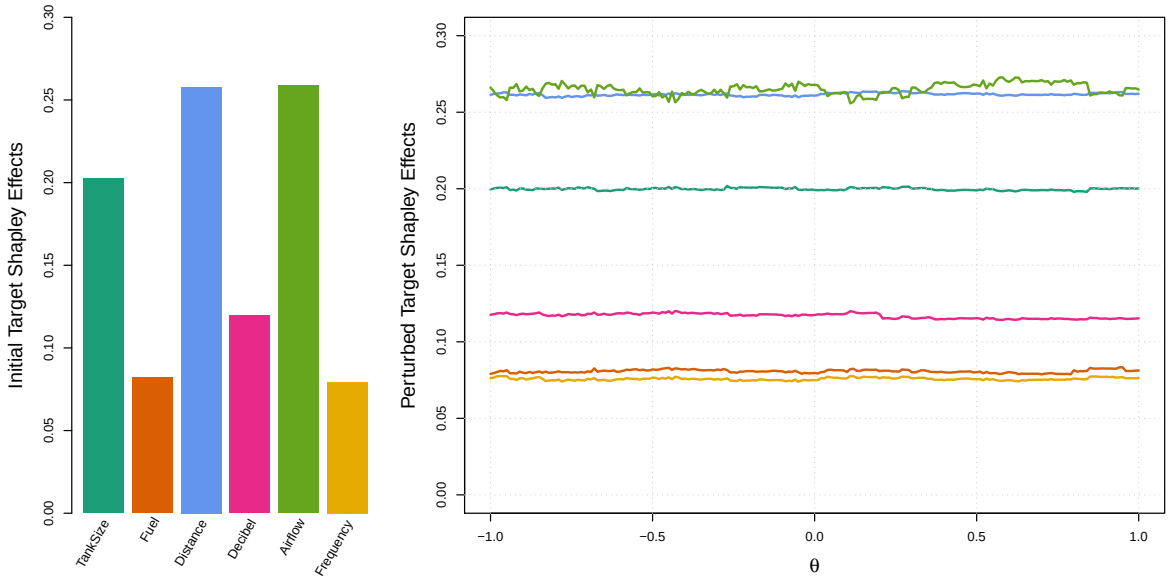

Figure 11: Initial (left) and perturbed (right) target Shapley effects, w.r.t. the intensity parameter θ, using the same color panel.

Input	Unit	Distribution	Application Domain	Description
Q	m^3/sec	$\mathcal{G}(1013, 558)$ trunc.	$[500, 3000]$	River maximum annual water flow rate.
K_s		$\mathcal{N}(35, 5)$ trunc.	$[20, 50]$	Strickler riverbed roughness coefficient.
Z_v	m	$\mathcal{T}(49, 50, 51)$	$[49, 51]$	Downstream river level.
Z_m	m	$\mathcal{T}(54, 55, 56)$	$[54, 56]$	Upstream river level.
L	m	$\mathcal{T}(4990, 5000, 5010)$	$[4990, 5010]$	River length.
B	m	$\mathcal{T}(295, 300, 305)$	$[295, 305]$	River width.

Table 2: Inputs of the simplified river water level model and their explicit marginal distributions. $\mathcal{G}, \mathcal{N}, \mathcal{T}$ denote Gumbel, Normal and Triangular distributions, respectively (trunc means truncated).

very cluttered riverbeds to earthen channels. However, to illustrate our robustness method, epistemic uncertainties are assumed to affect this application domain.

5.2.1. Perturbation strategy

In this use case, the three following inputs are perturbed. The river maximum annual water flow rate Q, the river length L, and the upstream river level Z_m are subject to the following quantile constraints:

- Quantile perturbations on Q:

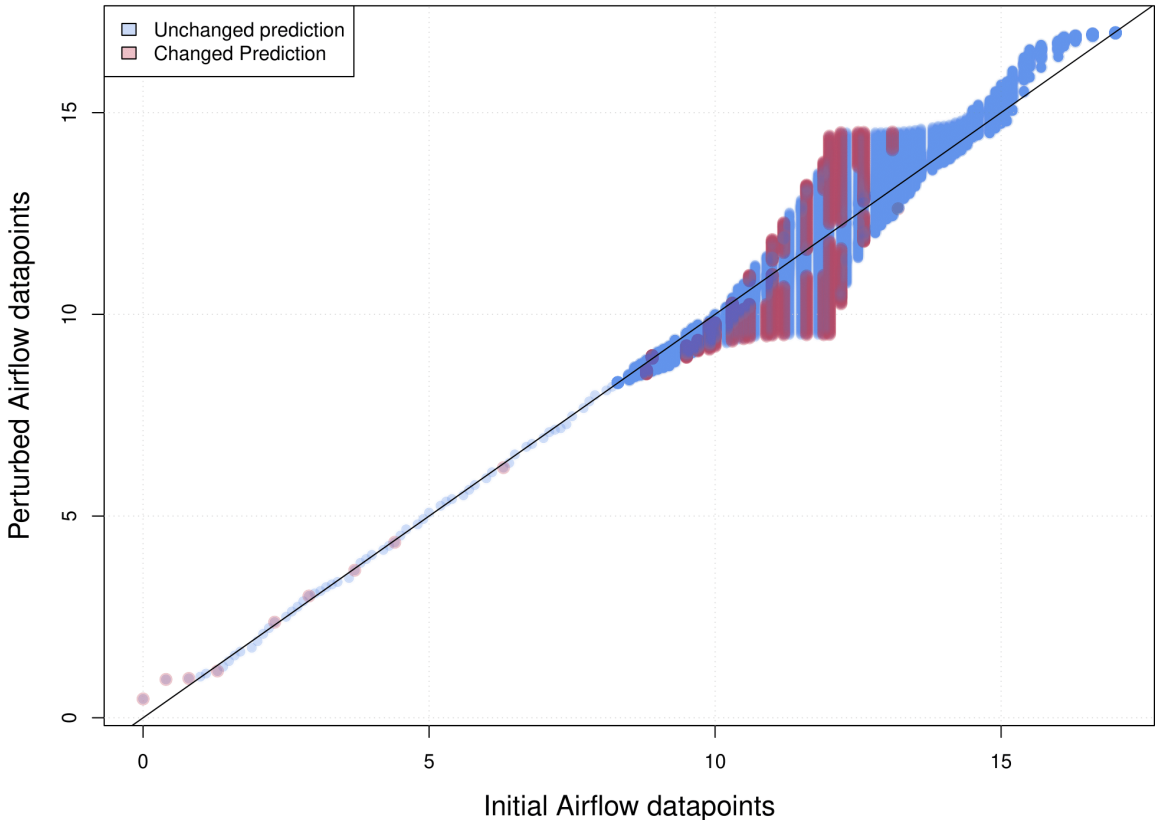

Figure 12: Perturbed datapoints w.r.t. their initial values. The black line represents no perturbation. The red and blue dots represent either a classification shift due to the perturbation or no classification shift, respectively.

  - Shift of the application domain from $[500, 3000]$ to $[500, 3200]$;
  - Preserve the median of the distribution;
  - Increase the initial 0.15-quantile by 75;
  - Decrease the initial 0.75-quantile by 125;

- Quantile perturbations on L:
  - Shift the application domain from $[4990, 5010]$ to $[4988, 5012]$;
  - Preserve the median of the distribution;

- Quantile perturbations on Z_m:
  - Preserve the application domain and the median of the initial distribution;
  - Increase the 0.8 and 0.9-quantiles by 0.1;
  - Decrease the 0.25-quantile by 0.05.

The initial input distributions, their application domain, and the optimally perturbed results are illustrated in Figure 13. These constraints are mainly enforced to illustrate that multiple inputs can be perturbed simultaneously while preserving their dependence structure.

In addition to these constraints, the Strickler coefficient K_s is subject to an application domain dilatation perturbation, with a scaling parameter $\eta = 2$. Each perturbation intensity represents a degree

of uncertainty on the type of riverbed roughness. When $\theta = -1$, the width of the initial application domain is halved, i.e., from $[20, 50]$ to $[27.5, 42.5]$, which can be interpreted in a situation where the epistemic uncertainty on the riverbed roughness is narrower, between a slow winding natural river, up to a plain river without shrub vegetation. When $\theta = 1$, the epistemic uncertainty on the riverbed is much wider, with an application domain equal to $[5, 65]$, which depicts a range of riverbed roughness from proliferating algae up to smooth concrete. Figure 14 illustrates the initial K_s distribution, along with the optimally perturbed quantile functions for θ being equal to -1 and 1.

Additionally, the perturbations' smoothness is enforced using piece-wise continuous isotonic polynomials of degree up to 12, chosen arbitrarily.

5.2.2. Robustness of the sensitivity analysis

From a global standpoint, one can be interested in the impact of the distributional perturbations on key statistics of the random output of the river water level model. Figure 15 presents estimated values for the mean, standard deviation, 0.025 and 0.975-quantiles (shown by the 95% coverage), and minimum and maximum values of the random output, computed on 10^5 Monte Carlo samples, w.r.t. the dilatation intensity θ. These values are compared to the reference ones according to the initial distribution of the inputs, estimated on a 2×10^5 Monte Carlo sample.

Notice that the expectation, standard deviation, 95% coverage quantiles, and minimum value of the model output remain stable under the distributional perturbations on the application domain of the Strickler coefficient. However, the estimated upper bound of the output support increases exponentially for positive values of θ. Widening the uncertainty on the type of riverbed allows for relatively rare events of high river water levels since the 0.975-quantile does not seem to be dramatically affected by the distributional perturbations.

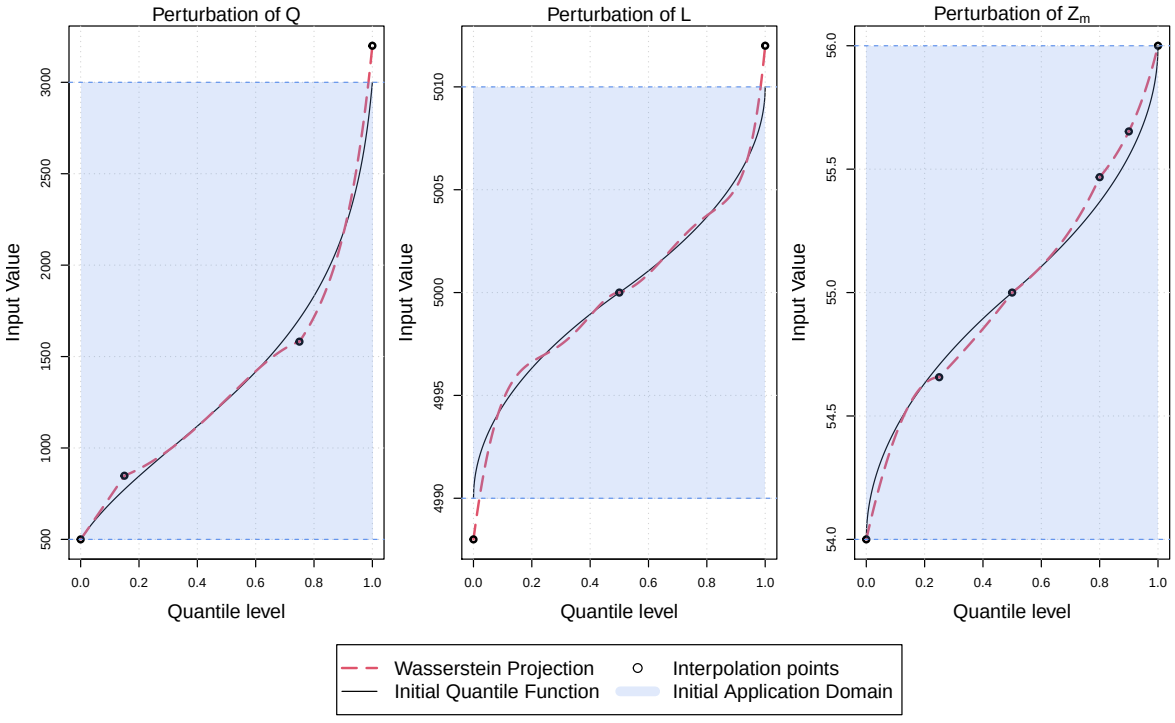

Figure 13: Initial quantile functions, application domains, and corresponding optimally perturbed quantile functions of the Q, L, and Z_m inputs.

Figure 16 presents the Shapley effects [80], which are global SA importance measure for real-valued model outputs with dependent inputs. These indices have been computed using a double Monte Carlo scheme as depicted in [102], with fixed simulated sample sizes, for each perturbed distribution Q driven by a value of θ, $N_v = 10^4$ for estimating $\mathrm{Var}_Q(Y)$, as well as $N_o = 10^3$ and $N_i = 100$ to estimate $\mathbb{E}_Q\left[\mathrm{Var}_Q\left(Y \mid X_A\right)\right]$ for every subset $X_A, A \subseteq \{1, \ldots, d\}$ of variables. Additionally, the reference Shapley effects have been computed under the initial distribution with sample sizes $N_v = 10^5$, $N_o = 3 \times 10^3$ and $N_i = 300$.

Note that the distributional perturbations have an impact on the importance measures. More precisely, increasing the range of the uncertainty of the riverbed roughness increases its importance for positive values of θ. Conversely, the importance of both Q and Z_v decreases accordingly. However, the variable importance hierarchy induced by the Shapley effects is preserved. It is also essential to notice that both Q and Z_v tend to be considered equally important as θ gets large. Hence, this SA does not seem robust to distributional perturbations and, more precisely, to a widening of the support of the Strickler coefficient in combination with the quantile perturbations put on Q, L, and Z_m.

5.3. Conclusions

These two use cases illustrate the usefulness of our method in both UQ and ML studies. On the ML side, for classification tasks, it allows assessing the global behavior of black-box models under input perturbations. This assessment is quantified either through studying the prediction shifts due to the perturbation, or through the behavior of feature importance metrics. Locally, it allows the detection of low-stability regions of interest (regions where small perturbations induce a classification change). Overall, in addition to classical accuracy metrics, our method can be used to assess confidence in a predictive model. On the UQ side, it allows for studying the impact of distributional perturbations (whose

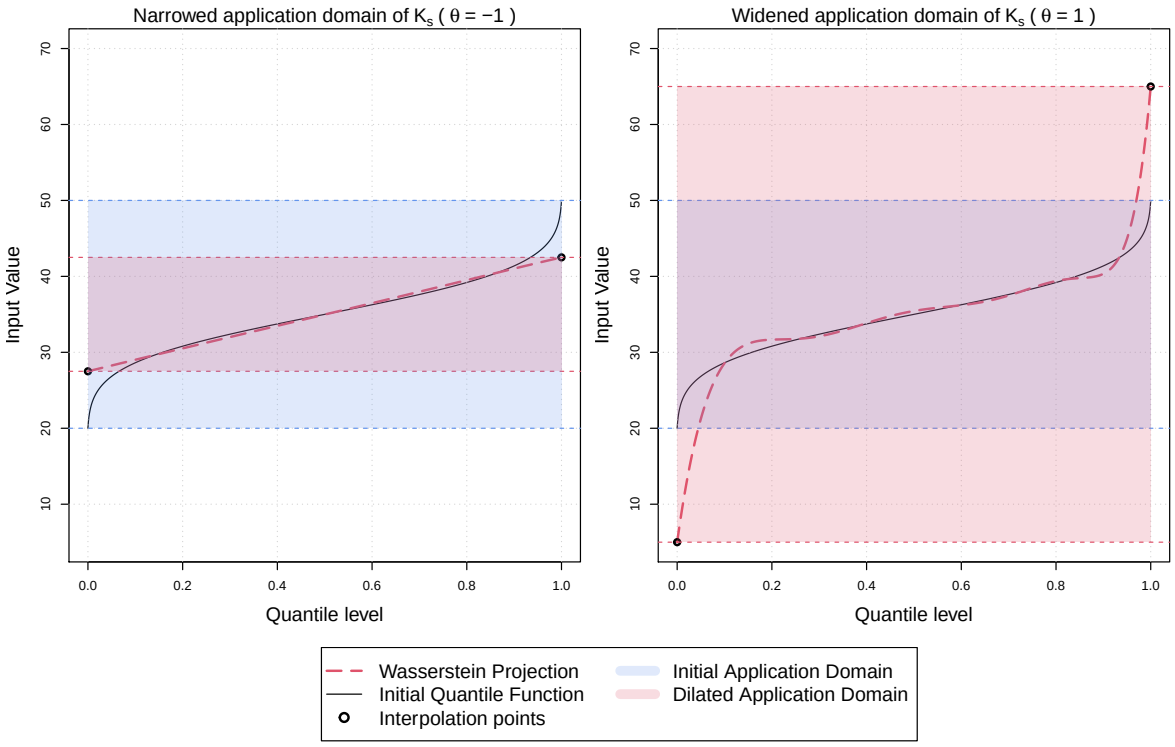

Figure 14: Initial quantile function, application domain and corresponding optimally perturbed quantile functions for K_s, for θ being equal to -1 (left) and 1 (right), for a scaling parameter $\eta = 2$.

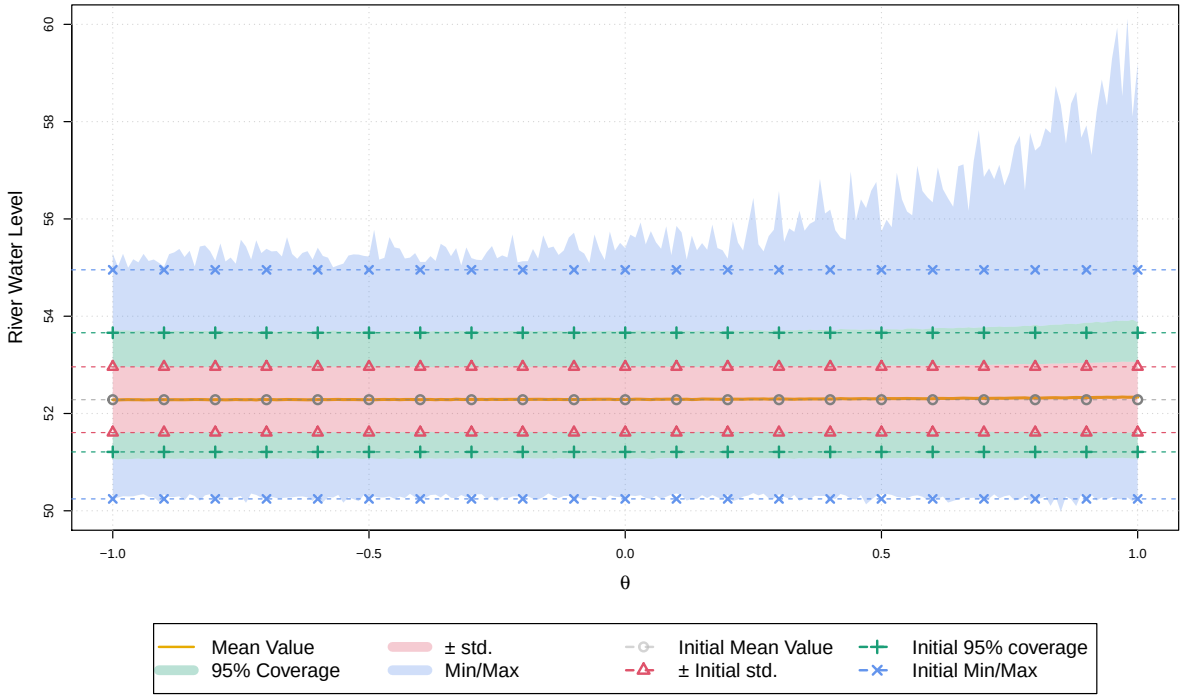

Figure 15: Expectation, standard deviation, 95% coverage, minimum and maximum estimators of the river water level, w.r.t. the application domain dilatation intensity θ.

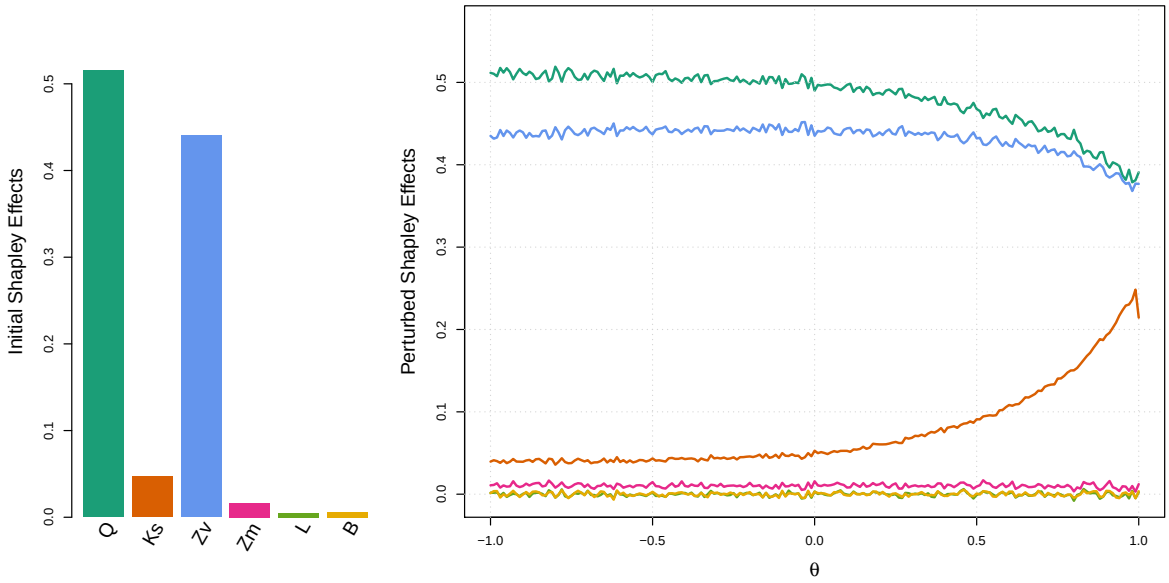

Figure 16: Reference Shapley effects (left) and Shapley effects of the river water level model under optimally dilated application domain w.r.t. θ (right), using the same color panel.

intensity can be tuned to represent epistemic uncertainties) on the model output, even in situations where inputs are correlated. Furthermore, in a SA context, the behavior of classical sensitivity indices under those perturbations can also be studied, and their robustness (for instance, the preservation of the input importance hierarchy) w.r.t. the probabilistic modeling on the inputs can be assessed.

6. Discussion and perspective

Obtaining a robustness diagnosis on the influence of input variables and the behavior of a model considered a black box is essential for its acceptance and use. Such models can either implement solutions of mechanistic equations or be learned from data. In both cases, the sensitivity of key indicators to a misspecification of the probabilistic input model must be evaluated. It is essential to have specific intelligible rules and well-defined computational tools to achieve this goal. This paper provides an answer to this question by proposing to modify the distributions of the input variables, seen as features. These perturbations modify the quantile of marginal distributions while preserving the dependence structure. We project the initial distribution under a Wasserstein cost to design optimal perturbations. The proposed approach allows the inclusion of regularity conditions through piece-wise isotonic polynomials. The robustness analyses conducted on real case studies illustrate its potential flexibility and speed, which are essential for scaling up (increasing dimension and size of data). These methods and examples have been implemented within a dedicated openly accessible GitLab repository.

We will continue this work by examining the following points, which offer specific avenues for further research. The first technical problem is determining rules for the degree of the nonnegative polynomials in smooth approximations. It can be understood as the injection of prior information on the order of differentiability of the sought-after perturbed gqf. If $F_P^{\leftarrow}$ is supposedly regular (on all or some parts of $[0,1]$) and its differentiability can be estimated, a rule-of-thumb heuristic could be to impose the same differentiability on the resulting perturbed gqf $F_Q^{\leftarrow}$. In an ML framework, nonparametric approaches to isotonic regression of the marginal gqfs of P can provide answers through statistical testing [34, 29] or criteria enforcing a trade-off between approximation error and sparsity (e.g., inspired from AIC or BIC).

The second problem to be addressed is providing rules for modifying the dependence structure between features. In an UQ context, classes of parametric high-dimensional (vine) copulas have been recently explored by [108, 14, 88, 8] for several tasks, including SA. In the light of the dedicated literature, association and concordance measures appear as the most interpretable tools for these multivariate distributions (and therefore frequently used to incorporate expert opinion) [24, 117, 14]. Relevant classes of parametric vine copulas could be defined objectively by the Fréchet-Hoeffding bounds on the conditional bivariate dependence structures of such copulas or by computing extreme dependence structures relevant for the phenomenon of interest, as proposed by [14]. The latter authors use a sparsity hypothesis to avoid prohibitive computational costs related to the curse of dimensionality. It echoes the strong need for sparsity assumptions about feature dependence in an ML framework, to which preliminary dimensionality reduction steps usually respond. In this context, many ML methods pay particular attention to the search for decorrelated, if not independent, features or small groups of features. It ensures the good behavior of the learning models and the relevance of the indicators allowing their interpretability [46]. For instance, it is well known that permutation-based indicators for random forests provide relevant information on the importance of input features solely when they are independent [48]. New solutions have recently been developed to overcome this limitation [12]. In addition to computational limitations, it seems difficult, if not impossible, to produce understandable indicators when dealing with a large number of features.

An alternative approach to account for feature dependence could be based on multivariate quantile extensions. Among the many approaches to defining such a notion, the most theoretically accomplished today is the one resulting from the concept of *center-outward distribution function*, based on optimal transportation ideas. It was first proposed and studied by [22] and [52]. The resulting *center-outward quantile function* offers desirable continuity and invertibility properties, which ensures the existence of closed and nested quantile contours [37]. Suggesting that characteristics of this multivariate quantile

function can be used to define variation classes seems natural to generalize our approach to multidimensional perturbations instead of focusing on marginal perturbations. Therefore we suggest that studying the interpretability of these features and the computational work required to handle quantile contours can be useful to improve our approach.

On the subject of isotonic polynomials, one can notice that the solution of the proposed approximation scheme can not be differentiable at the points of the grid. To circumvent this drawback and enforce the derivability of the smoothed gqf on $[0, 1]$, one could enforce additional constraints on the derivate of the polynomials. It would result in smoother perturbed gqf. Doing so would bear resemblance with splines [54], and in particular isotonic splines, which have been extensively studied in the literature [97, 39, 115].

As stated in the proof of Theorem 1, we chose to use the SOS representation of nonnegative polynomials and their subsequent characterization using semi-definite positive matrices. Works in [31, 75] could allow for less convoluted representations, leading to faster solving times. However, our developments allow for more apparent extensions to the definition of multivariate nonnegative polynomials. This decision has been made to account for using multivariate polynomials for smooth perturbation classes of quantile contours, which generalizes our approach to multidimensional perturbations.

Other spaces of functions can also be used for smoothing purposes. Following the work of [7], abstract reproducing kernel Hilbert space of nonnegative functions can be reached through particular kernels. Hence, it would allow accessing different sets of nonnegative functions, whose regularities can be assessed through a thorough study of these kernels.

Finally, one of the primary motivations for using the 2-Wasserstein distance as a projection metric is that it metricizes weak convergence on a broad set of probability measures. Other distances between probability measures are endowed with similar properties, such as the Prokhorov-Levy distance. An exciting improvement of our method would be to leverage the different relationships between such distances (see [45]) to assess their use's relevancy for robustness studies.

Acknowledgements

Support from the ANR-3IA Artificial and Natural Intelligence Toulouse Institute is gratefully acknowledged.

The authors warmly thank Jean-Bernard Lasserre (Institut de Mathématique de Toulouse) and Guillaume Dalle (CERMICS) for their help in solving the optimization problem at the heart of this work, as well as Clément Bénesse (Institut de Mathématiques de Toulouse) and Antoine Paolini (UVSQ Université Paris Saclay) for their support on some mathematical aspects of this study.

References

[1] A. Alfonsi and B. Jourdain. A remark on the optimal transport between two probability measures sharing the same copula. *Statistics & Probability Letters*, 84:131–134, January 2014.

[2] D.L. Allaire and K. E. Willcox. Distributional sensitivity analysis. *Procedia - Social and Behavioral Sciences*, 2:7595–7596, 2010.

[3] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 214–223, 2017.

[4] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. Synthesizing robust adversarial examples. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning (ICML), 10-15, 2018*, volume 80, pages 284–293, 2018.

[5] European Banking Authority. *2021 EU-Wide Stress Test*. European Banking Authority, 2020.

[6] F. Bachoc, F. Gamboa, M. Halford, J-M. Loubes, and L. Risser. Explaining Machine Learning Models using Entropic Variable Projection. *arXiv:1810.07924 (submitted)*, December 2020.

[7] J. A. Bagnell and A-M Farahmand. Learning positive functions in a hilbert space. *Preprint*, 2015.

[8] Z. Bai, H. Wei, Y. Xiao, S. Song, and S. Kucherenko. A Vine Copula-Based Global Sensitivity Analysis Method for Structures with Multidimensional Dependent Variables. *Mathematics*, 9:2489, 2021.

[9] J. Barr and H. Rabitz. A generalized kernel method for global sensitivity analysis. *SIAM/ASA Journal on Uncertainty Quantification*, 10(1):27–54, 2022.

[10] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, June 2020.

[11] M. Beer, S. Ferson, and V. Kreinovich. Imprecise probabilities in engineering analyses. *Mechanical Systems and Signal Processing*, 37(1):4–29, 2013.

[12] C. Bénard, G. Biau, S. Da Veiga, and E. Scornet. SHAFF: Fast and consistent SHApley eFfect estimates via random Forests. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 5563–5582, 2022.

[13] C. Bénard, S. Da Veiga, and E. Scornet. Mean decrease accuracy for random forests: inconsistency, and a practical solution via the Sobol-MDA. *Biometrika*, asac017, https://doi.org/10.1093/biomet/asac017, 2022.

[14] N. Benoumechiara, N. Bousquet, B. Michel, and P. Saint-Pierre. Detecting and modeling critical dependence structures between random inputs of computer models. *Dependence Modeling*, 8(1):263–297, 2020.

[15] D. P. Bertsekas. *Nonlinear programming*. Athena scientific, Belmont, Mass, 3rd ed edition, 2016.

[16] N. Bloom. The impact of uncertainty shocks. *Econometrica*, 77(3):623–685, 2009.

[17] E. Borgonovo, A. Figalli, E. Plischke, and G. Savare. Probabilistic Sensitivity with Optimal Transport. *Preprint*, 2022.

[18] B. Broto, F. Bachoc, and M. Depecker. Variance Reduction for Estimation of Shapley Effects and Adaptation to Unknown Input Distribution. *SIAM/ASA Journal on Uncertainty Quantification*, 8(2):693–716, 2020.

[19] L. Bruzzone and M. Marconcini. Domain Adaptation Problems: A DASVM Classification Technique and a Circular Validation Strategy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):770–787, 2010.

[20] D. G. Cacuci. *Sensitivity and uncertainty analysis - Theory*. Chapman & Hall/CRC, 2003.

[21] G. Chastaing, F. Gamboa, and C. Prieur. Generalized Hoeffding-Sobol decomposition for dependent variables - Application to sensitivity analysis. *Electronic Journal of Statistics*, 6:2420–2448, 2012.

[22] V. Chernozhukov, A. Galichon, M. Hallin, and M. Henry. Monge-Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223 – 256, 2017.

[23] Y. Chung, W. Neiswanger, I. Char, and J. Schneider. Beyond pinball loss: Quantile methods for calibrated uncertainty quantification. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10971–10984, 2021.

[24] R. T. Clemen and T. Reilly. Correlations and copulas for decision and risk analysis. *Management Science*, 45(2):208–224, 1999.

[25] R. Cont, R. Deguest, and G. Scandolo. Robustness and sensitivity analysis of risk measurement procedures. *Quantitative Finance*, 10:593–606, 2008.

[26] I. Covert, S. Lundberg, and S.-I. Lee. Understanding Global Feature Contributions With Additive Importance Measures. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17212–17223, 2020.

[27] I. Csiszár. I-Divergence Geometry of Probability Distributions and Minization problems. *The Annals of Probability*, 3(1):146–158, 1975.

[28] S. Da Veiga, F. Gamboa, B. Iooss, and C. Prieur. *Basics and Trends in Sensitivity Analysis. Theory and Practice in R.* SIAM. Computational Science and Engineering, 2021.

[29] S. Da Veiga and A. Marrel. Gaussian process modeling with inequality constraints. *Annales de la Faculté des Sciences de Toulouse*, 3:529–555, 2012.

[30] A. de la Fortelle. A study on generalized inverses and increasing functions Part I: generalized inverses. *INRIA Report hal-01255512*, 2015.

[31] H. Dette and W. J. Studden. *The theory of canonical moments with applications in statistics, probability, and analysis.* Wiley series in probability and statistics. Wiley, New York, 1997.

[32] J. C. Duchi, P. W. Glynn, and H. Namkoong. Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach. *Mathematics of Operations Research*, 43:835–1234, 2021.

[33] J.-M. Dufour. Distribution and quantile functions. *McGill University Report*, 1995.

[34] C. Durot and A.-S. Tocquet. Goodness of fit test for isotonic regression. *ESAIM:P&S*, 5:119–140, 2001.

[35] G. Ecoto, A. Bibault, and A. Chambaz. One-step ahead Super Learning from short time series of many slightly dependent data, and anticipating the cost of natural disasters. *arXiv:2107:13291*, 2021.

[36] T. Fel, R. Cadene, M. Chalvidal, M. Cord, D. Vigouroux, and T. Serre. Look at the Variance! Efficient Black-box Explanations with Sobol-based Sensitivity Analysis. In *Advances in Neural Information Processing Systems*, volume 34, pages 26005–26014, 2021.

[37] A. Figalli. On the continuity of center-outward distribution and quantile functions. *Nonlinear Analysis*, 177:413–421, 2018. Nonlinear PDEs and Geometric Function Theory, in honor of Carlo Sbordone on his 70th birthday.

[38] J-C Fort, T. Klein, and A. Lagnoux. Global Sensitivity Analysis and Wasserstein Spaces. *SIAM/ASA Journal on Uncertainty Quantification*, 9(2):880–921, 2021.

[39] S. Fredenhagen, H. J. Oberle, and G. Opfer. On the Construction of Optimal Monotone Cubic Spline Interpolations. *Journal of Approximation Theory*, 96(2):182–201, 1999.

[40] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T.A. Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[41] A. Fu, B. Narasimhan, and S. Boyd. CVXR: An R package for disciplined convex optimization. *Journal of Statistical Software*, 94(14):1–34, 2020.

[42] S. Fu, M. Couplet, and N. Bousquet. An adaptive kriging method for solving nonlinear inverse statistical problems. *Environmetrics*, 28(4):e2439, 2017.

[43] C. Gauchy, J. Stenger, R. Sueur, and B. Iooss. An information geometry approach to robustness analysis for the uncertainty quantification of computer codes. *Technometrics*, 64:80–91, 2022.

[44] A.E. Gelfand, B.K. Mallick, and D.K. Dey. Modeling expert opinion arising as a partial probabilistic specification. *Journal of the American Statistical Association*, 90(430):598–604, 1995.

[45] A.L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review / Revue Internationale de Statistique*, 70(3):419–435, 2002.

[46] L.H. Gilpin, D. Bau, B.Z. Yuan, A. Bajwa, M.A. Specter, and L. Kagal. Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89, 2018.

[47] P. Gordaliza, E. Del Barrio, F. Gamboa, and J.-M. Loubes. Obtaining Fairness using Optimal Transport Theory. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 2357–2365, 2019.

[48] B. Gregorutti, B. Michel, and P. Saint-Pierre. Correlation and variable importance in random forests. *Statistics and Computing*, 27(3):659–678, 2017.

[49] U. Grömping. Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7:137–152, 2015.

[50] P. Gustafson and L. Wasserman. Local sensitivity diagnostics for Bayesian inference. *The Annals of Statistics*, 23(6):2153–2167, 1995.

[51] Shimodaira; H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[52] M. Hallin, E. del Barrio, J. Cuesta-Albertos, and C. Matrán. Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*, 49(2):1139–1165, April 2021. Publisher: Institute of Mathematical Statistics.

[53] J. Hart and P.A. Gremaud. Robustness of the sobol' indices to distributional uncertainty. *International Journal for Uncertainty Quantification*, 9(5):453–469, 2019.

[54] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer: New York, 2009.

[55] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8320–8329, 2021.

[56] G. Hooker. Diagnosing extrapolation: tree-based density estimation. In W. Kim and R. Kohavi, editors, *Proceedings of the tenth ACM SIGKDD International conference on Knowledge Discovery and Data Mining (KDD '04)*, pages 569–574, 2004.

[57] G. Hooker and S. Rosset. Prediction-based regularization using data augmented regression. *Statistics and Computing*, 22:237–249, 2012.

[58] E. Hüllermeier and W. Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

[59] M. Il Idrissi, V. Chabridon, and B. Iooss. Developments and applications of Shapley effects to reliability-oriented sensitivity analysis with correlated inputs. *Environmental Modelling and Software*, 143:105115, 2021.

[60] B. Iooss, V. Chabridon, and V. Thouvenot. Variance-based importance measures for machine learning model interpretability. In *Congrès Lambda-Mu 23*, volume 23, Saclay, France, 2022.

[61] B. Iooss, R. Kennet, and P. Secchi. Different views of interpretability. In A. Lepore, B. Palumbo, and J-M. Poggi, editors, *Interpretability for Industry 4.0: Statistical and Machine Learning Approaches*. Springer, 2022.

[62] B. Iooss and P. Lemaître. A review on global sensitivity analysis methods. In G. Dellino and C. Meloni, editors, *Uncertainty Management in Simulation-Optimization of Complex Systems: Algorithms and Applications*, pages 101–122. Springer US, 2015.

[63] B. Iooss, V. Vergès, and V. Larget. BEPU robustness analysis via perturbed-law based sensitivity indices. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 236:655–665, 2022.

[64] O. Kallenberg. *Foundations of modern probability*. Probability theory and stochastic modelling. Springer, Cham, Switzerland, 2021.

[65] M. Koklu and Y. S. Taspinar. Determining the Extinguishing Status of Fuel Flames With Sound Wave by Machine Learning Methods. *IEEE Access*, 9:86207–86216, 2021.

[66] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville. Out-of-distribution generalization via risk extrapolation (rex). In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 5815–5826, 2021.

[67] S. Kurtek and K. Bharath. Bayesian sensitivity analysis with the Fisher–Rao metric. *Biometrika*, 102(3):601–616, 2015.

[68] J-B. Lasserre. *An Introduction to Polynomial and Semi-Algebraic Optimization*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 2015.

[69] P. Lemaître, E. Sergienko, A. Arnaud, N. Bousquet, F. Gamboa, and B. Iooss. Density modification-based reliability sensitivity analysis. *Journal of Statistical Computation and Simulation*, 85(6):1200–1223, 2015.

[70] P. Lemberger and I. Panico. A Primer on Domain Adaptation. Theory and Applications. *arXiv:2001.09994*, 2020.

[71] K. Liu, H. Kargupta, and J. Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on Knowledge and Data Engineering*, 18:92–106, 2006.

[72] V. Maume-Deschamps and I. Niang. Estimation of quantile oriented sensitivity indices. *Statistics and Probability Letters*, 134:122–127, 2018.

[73] P. Mikkola, O. Martin, S. Chandramouli, M. Hartmann, O. Pla, O. Thomas, H. Pesonen, J. Corander, A. Vehtari, S. Kaski, P-C Bürkner, and A. Klami. Prior knowledge elicitation: The past, present, and future, 2021. arXiv:2112.01380.

[74] C. Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. leanpub.com, 1 edition, 2021.

[75] K. Murray, S. Müller, and B. A. Turlach. Fast and flexible methods for monotone polynomial fitting. *Journal of Statistical Computation and Simulation*, 86(15):2946–2966, 2016.

[76] A. Narayan and D. Xiu. Distributional sensitivity for uncertainty quantification. *Communications in Computational Physics*, 10(1):140–160, 2011.

[77] R. B. Nelsen. *An introduction to copulas*. Springer series in statistics (2nd edition). Springer, New York, 2006.

[78] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro. Exploring generalization in deep learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5949–5958, 2017.

[79] B. O'Donoghue, E. Chu, N. Parikh, and S. Boyd. Conic Optimization via Operator Splitting and Homogeneous Self-Dual Embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068, 2016.

[80] A. B. Owen. Sobol' Indices and Shapley Value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):245–251, 2014.

[81] T. Paananen, J. Piironen, M. Riis Andersen, and A. Vehtari. Variable selection for Gaussian processes via sensitivity analysis of the posterior predictive distribution. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pages 1743–1752, 2019.

[82] P. A. Parrilo. Algebraic Optimization and Semidefinite Optimization. *MIT Lectures Notes (EIDMA Minicourse)*, 2010.

[83] P. A. Parrilo. Polynomial optimization, sums of squares, and applications. In *Semidefinite Optimization and Convex Algebraic Geometry*, pages 47–157. SIAM, 2012.

[84] M.K. Paul, M.R. Islam, and Sarowar Sattar A.H.M. An efficient perturbation approach for multivariate data in sensitive and reliable data mining. *Journal of Information Security and Applications*, 62:102954, 2021.

[85] S. M. Pesenti, A. Bettini, P. Millossovich, and A. Tsanakas. Scenario Weights for Importance Measurement (SWIM) – an R package for sensitivity analysis. *Annals of Actuarial Science*, 15(2):458–483, 2021.

[86] S.M. Pesenti. Reverse Sensitivity Analysis for Risk Modelling. *Risks*, 10:141, 2022.

[87] G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[88] E. Plischke and E. Borgonovo. Copula theory and probabilistic sensitivity analysis: Is there a connection? *European Journal of Operational Research*, 277(3):1046–1059, 2019.

[89] M.S. Pydi and V. Jog. Adversarial Risk via Optimal Transport and Optimal Couplings. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 37th International Conference on Machine Learning*, pages 2357–2365, 2020.

[90] S. Razavi, A. Jakeman, A. Saltelli, C. Prieur, B. Iooss, E. Borgonovo, E. Plischke, S. Lo Piano, T. Iwanaga, W. Becker, S. Tarantola, J.H.A. Guillaume, J. Jakeman, H. Gupta, N. Melillo, G. Rabitti, V. Chabridon, Q. Duan, X. Sun, S. Smith, R. Sheikholeslami, N. Hosseini, M. Asadzadeh, A. Puy, S. Kucherenko, and H.R. Maier. The Future of Sensitivity Analysis: An essential discipline for systems modeling and policy support. *Environmental Modelling and Software*, 137:104954, 2021.

[91] S. I. Resnick. Preliminaries. In S. I. Resnick, editor, *Extreme Values, Regular Variation and Point Processes*, Springer Series in Operations Research and Financial Engineering, pages 1–37. Springer, New York, NY, 1987.

[92] M. Roos, T.G. Martins, L. Held, and H. Rue. Sensitivity Analysis for Bayesian Hierarchical Models. *Bayesian Analysis*, 10:321–349, 2015.

[93] C.J. Roy and W.L. Oberkampf. A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing. *Computer Methods in Applied Mechanics and Engineering*, 200(25):2131–2144, 2011.

[94] R.Y. Rubinstein. Sensitivity analysis and performance extrapolation for computer simulation models. *Operation Research*, 37(1):72–81, 1989.

[95] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K-R. Müller, editors. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*. Springer International Publishing, Cham, 2019.

[96] F. Santambrogio. *Optimal Transport for Applied Mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Springer International Publishing, Cham, 2015.

[97] J. W. Schmidt and W. Heß. Positivity of cubic polynomials on intervals and positive spline interpolation. *BIT Numerical Mathematics*, 28(2):340–352, 1988.

[98] Christian A. Scholbeck, Christoph Molnar, Christian Heumann, Bernd Bischl, and Giuseppe Casalicchio. Sampling, Intervention, Prediction, Aggregation: A Generalized Framework for Model-Agnostic Interpretations. In Peggy Cellier and Kurt Driessens, editors, *Machine Learning and Knowledge Discovery in Databases*, Communications in Computer and Information Science, pages 205–216, Cham, 2020. Springer International Publishing.

[99] Z. Shen, J. Liu, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui. Towards Out-Of-Distribution Generalization: A Survey. *arXiv:2108.13624*, 2021.

[100] R. C. Smith. *Uncertainty Quantification: Theory, Implementation, and Applications*. Computational Science & Engineering. SIAM, 2014.

[101] O. Sobrie, N. Gillis, V. Mousseau, and M. Pirlot. UTA-poly and UTA-splines: Additive value functions with polynomial marginals. *European Journal of Operational Research*, 264(2):405–418, 2018.

[102] E. Song, B. L. Nelson, and J. Staum. Shapley Effects for Global Sensitivity Analysis: Theory and Computation. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1060–1083, 2016.

[103] H. Song, T. Diethe, M. Kull, and P. Flach. Distribution calibration for regression. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, pages 5897–5906, 2019.

[104] A. Stevens, P. Deruyck, Z. Van Veldhoven, and J. Vanthienen. Explainability and Fairness in Machine Learning: Improve Fair End-to-end Lending for Kiva. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1241–1248, 2020.

[105] T. Sullivan. *Introduction to Uncertainty Quantification*. Springer, 2017.

[106] Y. S. Taspinar, M. Koklu, and M. Altin. Classification of flame extinction based on acoustic oscillations using artificial intelligence methods. *Case Studies in Thermal Engineering*, 28:101561, December 2021.

[107] Y. S. Taspinar, M. Koklu, and M. Altin. Acoustic-Driven Airflow Flame Extinguishing System Design and Analysis of Capabilities of Low Frequency in Different Fuels. *Fire Technology*, 58(3):1579–1597, May 2022.

[108] E. Torre, S. Marelli, P. Embrechts, and B. Sudret. A general framework for data-driven uncertainty quantification under complex input dependencies using vine copulas. *Probabilistic Engineering Mechanics*, 55:1–16, 2019.

[109] N. Tripuraneni, B. Adlam, and J. Pennington. Overparameterization improves robustness to covariate shift in high dimensions. In *35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[110] G. Valle-Pérez and A.A. Louis. Generalization bounds for deep learning. *arXiv:2012.04115*, 2020.

[111] V. N. Vapnik and A. Y. Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, January 1971. Publisher: Society for Industrial and Applied Mathematics.

[112] C. Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, March 2003.

[113] G. Visani, E. Bagli, F. Chesani, A. Poluzzi, and D. Capuzzo. Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models. *Journal of the Operational Research Society*, 73(1):91–101, 2022.

[114] J. Wang, C. Lan, C. Liu, Y. Ouyang, and T. Qin. Generalizing to unseen domains: A survey on domain generalization. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4627–4635, 8 2021.

[115] X. Wang and F. Li. Isotonic Smoothing Spline Regression. *Journal of Computational and Graphical Statistics*, 17(1):21–37, 2008.

[116] R. Yousefzadeh. Deep Learning Generalization and the Convex Hull of Training Sets. *arXiv:2101.09849*, 2021.

[117] M. Zondervan-Zwijnenburg, W. van de Schoot-Hubeek, K. Lek, H. Hoijtink, and R. van de Schoot. Application and Evaluation of an Expert Judgment Elicitation Procedure for Correlations. *Frontiers in Psychology*, 8:90, 2017.

Appendix A. Proofs

Proof of Remark 1. First, recall the following result from [33]:

Theorem 2. *If G is a real-valued non-decreasing, left-continuous function with domain $(0,1)$, then there is a unique distribution function F such that $G = F^{\leftarrow}$.*

Noticing that distribution functions (i.e., functions in $\mathcal{F}$) uniquely induce probability measures as their Lebegue-Stieltjes measures (see, [64], Thm 2.14, p42) ultimately proves the proposition. □

Proof of Lemma 1. Notice that if (8) is respected, then the constraints are non-decreasing. Then, there exists at least a function $F^{\leftarrow}$ in $\mathcal{F}^{\leftarrow}$ such that the constraints are respected (e.g., the linear interpolant of the constraints). So that, using Proposition 1, there exists a probability measure with $F^{\leftarrow}$ as a generalized quantile function. □

Proof of Lemma 2. Since $[\eta_0, \eta_1]$ is bounded, one can define a standardized intensity parameter $\theta \in \Theta = [-1, 1]$ as:

$$\theta(b) = \frac{p_\alpha - b}{p_\alpha - \eta_1} \mathbb{1}_{\{b > p_\alpha\}}(b) + \frac{b - p_\alpha}{p_\alpha - \eta_0} \mathbb{1}_{\{b < p_\alpha\}}(b).$$

This intensity level can be interpreted as follows:

- $-1 \leq \theta < 0 \Leftrightarrow b < p_\alpha$: the (perturbed) distribution Q such that $F_Q^{\leftarrow}(\alpha) = b$ is constrained to have a lower α-quantile than P, down to η_0;
- $\theta = 0 \Leftrightarrow b = p_\alpha$: Q and P share the same α-quantile;
- $0 < \theta \leq 1 \Leftrightarrow b > p_\alpha$: Q is constrained to have a higher α-quantile than P, up to η_1.

Equivalently, one can express b in terms of θ, which directly provides the expression of $b_\alpha(\boldsymbol{\eta}, \theta)$. □

Proof of Lemma 3. Preserving the midpoint of Ω_X while perturbing its width requires that, for any couple $(b_0, b_1) \in \mathbb{R}^2$, that

$$\begin{cases} \dfrac{b_0 + b_1}{2} = \dfrac{\omega_0 + \omega_1}{2} \\ b_1 - b_0 = \kappa(\omega_1 - \omega_0) \end{cases} \iff \begin{cases} b_1 = \dfrac{\omega_1(\kappa + 1) - \omega_0(\kappa - 1)}{2} \\ b_0 = \dfrac{\omega_0(\kappa + 1) - \omega_1(\kappa - 1)}{2} \end{cases}$$

where $\kappa \in [\frac{1}{\eta}, \eta]$. Using the transformation

$$\theta(\kappa) = \begin{cases} -\dfrac{\kappa - 1}{\frac{1}{\eta} - 1} & \text{if} \quad \dfrac{1}{\eta} \leq \kappa < 1 \\ 0 & \text{if} \quad \kappa = 1 \\ \dfrac{\kappa - 1}{\eta - 1} & \text{if} \quad 1 < \kappa < \eta \end{cases}$$

allows to define the formulas for b_0 and b_1 provided in the lemma statement. The perturbation intensity θ can be interpreted as follows:

- If $-1 \leq \theta < 0$, the application domain is narrowed, its width being divided by up to η;
- If $\theta = 0$, the application domain boundaries are not perturbed;
- If $0 < \theta \leq 1$, the application domain is widened, its width being multiplied by up to η.

□

Proof of Remark 2. **(i)** Consider the empirical situation, typically encountered in ML, where the marginal distributions of the inputs, respectively $P_1, \ldots, P_d$, are purely atomic. For $j \in \{1, \ldots, d\}$, denote $\{x_{j,i}\}_{1 \leq i \leq n}$ the jth marginal sample of observations. From [77] the dependence structure of the joint measure P is known by its empirical copula $\hat{C}_P : [0, 1]^d \to [0, 1]$ defined as

$$\hat{C}_P(u_1, \ldots, u_d) = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{d} \mathbb{1}_{\left\{\frac{R_{j,i}}{n} \leq u_j\right\}}(u_j), \tag{A.1}$$

where $R_{j,k}$ denotes the rank of $x_{j,k}$ in $\{x_{j,i}\}_{1\leq i\leq n}$.

Perturbing only the marginals of P when solving the optimal projection problem (2) amounts apply to each marginal sample $\{x_{j,i}\}_{1\leq i\leq n}$ the transportation maps

$$T_j = (F_{Q_j}^{\leftarrow} \circ F_{P_j}), \quad j = 1, \ldots, d. \tag{A.2}$$

and consequently the marginal empirical probability measure of the perturbed samples are defined as

$$\widetilde{Q}_j = \frac{1}{n}\sum_{i=1}^{n} \delta_{\tilde{x}_{j,i}}$$

where $\tilde{x}_{j,i} = T_j(x_{j,i})$ are the perturbed datapoints. Since the marginal cdf and gqf in (A.2) are non-decreasing (ie., monotonic) functions, their composition is non-decreasing as well. Since monotonic functions preserve orders, the ranks of $x_{j,i}$ and $\tilde{x}_{j,i}$ are invariant, for $(i,j) \in \{1,\ldots,n\} \times \{1,\ldots,d\}$. From (A.1), it implies that the empirical copula of Q is $\hat{C}_Q = \hat{C}_P$. This traduces, for instance, the invariance of Spearman correlation matrices between the initial and perturbed datasets.

The transportation map T_j defined by (10) is not the unique non-decreasing transportation map between P and Q ([96], Chap. 2). However, T_j is indeed an optimal perturbation plan (in Monge's sense) for a wide variety of transportation costs between the marginals P_j and $\widetilde{Q}_j$ ([87], Remark 2.30). Moreover notice that for any $j = 1, \ldots, d$

$$\widetilde{Q}_j \underset{n\to\infty}{\longrightarrow} Q_j$$

since the distribution of $F_{P_j}(X_j)$ converges toward a uniform distribution on $[0,1]$.

(ii) Now, for the sake of conciseness, let P denote any univariate probability measure and Q its optimally perturbed counterpart. If $F_Q^{\leftarrow}$ is strictly increasing then from [33], for all $u \in [0,1]$

$$(F_Q \circ F_Q^{\leftarrow})(u) = u$$

Now denote the transportation map

$$\begin{array}{rccl} T : \mathcal{X} & \to & \mathcal{X} \\ x & \mapsto & (F_Q^{\leftarrow} \circ F_P)(x) \end{array}$$

Let $X \sim P$ and define $U_P = F_P(X)$. Then, one has, if P is atomless, that $T(X) \sim Q$ and

$$U_Q = F_Q(T(X)) = (F_Q \circ F_Q^{\leftarrow} \circ F_P)(X) = F_P(X) = U_P \text{ a.s.}$$

Now, for a multivariate probability measure P with marginals $P_1, \ldots, P_d$, and respective optimally perturbed probability measures $Q_1, \ldots, Q_d$, this entails that, for $i = 1, \ldots, d$:

$$U_{Q_i} = F_{Q_i}(T(X_i)) = F_{P_i}(X_i) = U_{P_i} \text{ a.s.}$$

and hence, the random vectors U_Q and U_P are equal almost surely. Subsequently, for any $u \in [0,1]^d$, $C_P(u) = C_Q(u)$. Hence the X and $T(X)$ have the same copula. □

Proof of Proposition 1. First, one can notice that, for any probability measure $G \in \mathcal{P}(\mathbb{R})$:

$$F_G^{\leftarrow} = F_G^{\rightarrow} \quad \mu - \text{almost everywhere (a.e.)}$$

where μ denotes the Lebesgue measure on $[0,1]$ (see, [30]). This entails that:

$$\begin{aligned} \int_0^1 (F_G^{\leftarrow} - F_P^{\rightarrow})^2 \, d\mu &= \int_0^1 (F_G^{\rightarrow} - F_P^{\rightarrow})^2 \, d\mu \\ &= W_2^2(P, G) \end{aligned}$$

by a continuous composition of μ-a.e. equal functions, and by identification with the definition of the W_2 distance for probability measures supported on $\mathbb{R}$.

In the following of this proof, one refers to the projection problem (12) as the "Wasserstein projection", and the projection problem (13) as the "L^2 projection".

The proposition can be proven in two steps. First, if Q is the solution of the Wasserstein projection, then its quantile function $F_Q^{\leftarrow}$ is necessarily the solution of the L^2 projection. This is due to the fact that $F_Q^{\leftarrow}$ is uniquely characterized by Q, as well as the particular case of the 2-Wasserstein distance for measures supported on $\mathbb{R}$ (see Definition 4).

Second, let Q be any probability measure in $\mathcal{P}_2(\mathbb{R})$, and denote $F_Q^{\leftarrow}$ its characterizing quantile function. Assume that $F_Q^{\leftarrow}$ is the solution of the L^2 projection. Since $F_Q^{\leftarrow}$ characterizes uniquely Q, thanks to Proposition 1, one has that Q is necessarily the solution of the Wasserstein projection. □

Proof of Proposition 2. First, note that the intervals $A_i, i = 1, \dots, K$ are disjoint. Moreover for any $i = 1, \dots, K-1$, consider the four cases:

1. If $\alpha_i < \beta_i < \alpha_{i+1}$ and, then $A_i = (\alpha_i, \beta_i]$;
2. If $\beta_i < \alpha_i < \beta_{i+1}$ and, then $A_i = (\beta_i, \alpha_i]$;
3. If $\alpha_i < \beta_i$ and assume that $\alpha_{i+j} < \beta_{i+j-1}$ for $j = 1, \dots, m$ where $m \leq K-i$ is some non-negative integer, then $A_i = (\alpha_i, \alpha_{i+1}]$, additionally for $j = i+1, \dots, i+m-1$, $A_j = (\alpha_j, \alpha_{j+1}]$ and finally $A_{i+m} = (\alpha_{i+m}, \beta_{i+m}]$;
4. If $\beta_i < \alpha_i$ and assume that $\alpha_{i+j} < \beta_{i+j+1}$ for $j = 1, \dots, m$ where $m \leq K-i-1$ is some non-negative integer, then $A_i = (\beta_i, \alpha_i]$ and for $j = i+1, \dots, i+m$, $A_j = (\alpha_{j-1}, \alpha_j]$.

The integral can be decomposed as follows:

$$\int_0^1 \left(L(x) - F_P^{\rightarrow}(x)\right)^2 dx = \int_{\overline{A}} \left(L(x) - F_P^{\rightarrow}(x)\right)^2 dx + \sum_{i=1}^{K} \int_{A_i} \left(L(x) - F_P^{\rightarrow}(x)\right)^2 dx$$

where

$$\int_{\overline{A}} \left(L(x) - F_P^{\rightarrow}(x)\right)^2 dx \geq 0.$$

Since the quantile constraints are of the form:

$$L(\alpha_i) \leq b_i \leq L\left(\alpha_i^+\right).$$

we can always write $L(y) = b_i + h(y)$ for $y \in A_i$, and where h is an non-decreasing, left-continuous function. Moreover, note that:

- $h(y)$ is non-negative, and $F_P^{\rightarrow}(y) - b_i \leq 0$ if A_i falls in cases 2. and 4.
- $h(y)$ is non-positive, and $F_P^{\rightarrow}(y) - b_i \geq 0$ if A_i falls in cases 1. and 3.

Then we have:

$$\begin{aligned}
\int_{A_i} \left(L(x) - F_P^{\rightarrow}(x)\right)^2 dx &= \int_{A_i} \left(L(x) - b_i - h(y)\right)^2 dx \\
&= \int_{A_i} \left(F_P^{\rightarrow}(x) - b_i\right)^2 dx + \int_{A_i} h(x)^2 dx \\
&\quad - 2 \int_{A_i} h(x) \left(F_P^{\rightarrow}(x) - b_i\right) dx \\
&\geq \int_{A_i} \left(F_P^{\rightarrow}(x) - b_i\right)^2 dx
\end{aligned}$$

since $h(x)$ and $F_P^{\rightarrow}(x) - b_i$ have different sign. Due to the constraint and the left-continuous non-decreasing nature of L, this bound is tight and is attained if and only if $h(y) = 0$ for all $y \in A_i$. Globally, this entails that

$$\int_0^1 \left(L(x) - F_P^{\rightarrow}(x)\right)^2 dx \geq \sum_{i=1}^{K} \int_{A_i} \left(F_P^{\rightarrow}(x) - b_i\right)^2 dx$$

and this tight bound is uniquely attained by the left-continuous non-decreasing function defined as

$$F_Q^{\leftarrow}(y) = \begin{cases} F_P^{\rightarrow}(y) & \text{if } y \in \overline{A} \\ b_i & \text{if } y \in A_i, \quad i = 1, \ldots, K. \end{cases}$$

□

Proof of Theorem 1 (ingredients). This proof relies on the following results that can be found in [82, 83, 68], and further recalled in [101]. They involve sum-of-squares (SOS) polynomials, which can be defined as follows.

Definition 5 (SOS polynomials)**.** *A polynomial S of even degree p is said to be a SOS polynomial if, for $m \in \mathbb{N}^*$, there exists $s_1, \ldots, s_m$ polynomials of degree at most equal to $\frac{d}{2}$, and such that, $\forall x \in \mathbb{R}$:*

$$S(x) = \sum_{i=1}^{m} \big(s_i(x)\big)^2.$$

Theorem 3. *Let $t_0, t_1 \in \mathbb{R}$ such that $t_0 < t_1$, and let $p \in \mathbb{N}^*$.*

(i) *A univariate polynomial S of even degree $d = 2p$ is non-negative on $[t_0, t_1]$ if and only if it can be written as, $\forall x \in [t_0, t_1]$*

$$S(x) = Z(x) + (x - t_0)(t_1 - x)W(x)$$

where Z is a SOS polynomial of degree at most equal to d, and W is an SOS polynomial of degree at most equal to $d - 2$.

(i) *An univariate polynomial S of odd degree $d = 2p+1$ is non-negative on $[t_0, t_1]$ if and only if it can be written as, $\forall x \in [t_0, t_1]$*

$$S(x) = (x - t_0)Z(x) + (t_1 - x)W(x)$$

where Z, W are SOS polynomials of degree at most equal to d.

It is important to note that Theorem 3 is quite general, in the sense that it allows for extensions to multivariate polynomials (i.e., polynomials taking values from $\mathbb{R}^d$). As pointed out in [31] (Thm. 1.4.2), nonnegative polynomials on compact intervals can also be defined as a linear combination of squared polynomials. It may facilitate the identification of the nonnegative polynomials' coefficients, as done in [75] in the context of statistical learning. However, for the sake of potential future genericity, we chose to leverage the direct powerful link between SOS polynomials and semi definite positive matrices, as expressed in the following theorem.

Theorem 4. *Let S be an univariate polynomial of even degree $d = 2p$, with coefficients $s = (s_0, \ldots, s_d)$, and denote x_p the usual monomial basis of polynomials of degree at most equal to p, i.e., $x_p = (1, x, x^2, \ldots, x^{p-1}, x^p)^\top$. S is an SOS polynomial if and only if there exists a $(p \times p)$ symmetric semi definite positive (SDP) matrix*

$$\Gamma = \big[\Gamma_{ij}\big]_{i,j=1,\ldots,p}$$

that satisfies, $\forall x \in \mathbb{R}$,

$$S(x) = x_p^\top \Gamma x_p.$$

Moreover, for $k = 0, \dots, d$, *let* $\mathbb{I}_k^p$ *be the* $(p \times p)$ *matrix defined by, for* $i, j = 1, \dots, p$:

$$\left[\mathbb{I}_k^p\right]_{i,j} = \mathbb{1}_{\{i+j=k+2\}}(i,j).$$

Then one additionally has that, for $i = 0, \dots, d$

$$s_i = \langle \mathbb{I}_i^p, \Gamma \rangle_F = \sum_{j+k=i+2} \Gamma_{j,k} \tag{A.3}$$

where, $\langle ., . \rangle_F$ *denotes the Frobenius norm on matrices.*

Theorem 5. *Let* $\mathbb{S}_n$ *the subspace of real-valued symmetric matrices, in the vector space of square matrices. The set of symmetric SDP matrices* Σ_N *is a proper cone in* $\mathbb{S}_n$, *and thus is a closed convex set.*

A few results on the preservation of convexity of sets under transformations are also required. These lemmas can be found in [15].

Lemma 4 (Linear maps preserve convexity)**.** *Let* V, W *be two vector spaces over the same field* F. *Let* $T : V \to W$ *be a linear map, and let* $\mathcal{C} \subset V$ *be a convex set. Then the image of* $\mathcal{C}$ *under* T, *i.e., :*

$$T(C) = \{T(x) \in W \mid x \in C \subset V\}$$

is also a convex set.

Lemma 5 (Cartesian product of convex sets is a convex set)**.** *Let* C_1 *be a subset of* $\mathbb{R}^m$ *and* C_2 *be a convex subset of* $\mathbb{R}^n$. *Then, the Cartesian product* $C_1 \times C_2$ *is a convex subset of* $\mathbb{R}^m \times \mathbb{R}^n$.

Two additional results, proven beneath, are required before proceeding to the proof of Theorem 1.

Lemma 6. *The mapping in (A.3),* $V : \mathbb{S}_p \to \mathbb{R}^{2p}$, *defined, for any* $\Gamma \in \mathbb{S}_p$, *as:*

$$V(\Gamma) = \left(\sum_{j+k=i+2} \Gamma_{j,k} \right)_{i=0,\dots,2p}$$

is linear.

Proof of Lemma 6. We need to show that:

- For $A, B \in \mathbb{S}_p$, $T(A+B) = T(A) + T(B)$;
- For $\alpha \in \mathbb{R}$, $\Gamma \in \mathbb{S}_p$, $T(\alpha\Gamma) = \alpha T(\Gamma)$.

First, we have, for $i = 0, \dots, 2p$:

$$\begin{aligned}
\Big[T(A+B)\Big]_i &= \sum_{j+k=2p-i} \Big[A+B\Big]_{jk} \\
&= \sum_{j+k=i+2} A_{jk} + B_{jk} \\
&= \sum_{j+k=i+2} A_{jk} + \sum_{j+k=i+2} B_{jk} \\
&= \Big[T(A)\Big]_i + \Big[T(B)\Big]_i
\end{aligned}$$

since it holds for $i = 0, \dots, 2p$, it entails:

$$T(A + B) = T(A) + T(B).$$

Moreover, we have, for $i = 0, \dots, 2p$:

$$\begin{aligned}\Big[T(\alpha\Gamma)\Big]_i &= \sum_{j+k=i+2} \alpha\Gamma_{jk} \\ &= \alpha\Big[T(\Gamma)\Big]_i\end{aligned}$$

and since it holds for $i = 0, \dots, 2p$, it entails:

$$T(\alpha\Gamma) = \alpha T(\Gamma).$$

Hence T is a linear map between $\mathbb{S}_p$ and $\mathbb{R}^{2p}$. □

Lemma 7. *Let S be a univariate polynomial of degree d and $s = (s_0, \dots, s_d)^\top \in \mathbb{R}^{d+1}$ its coefficients. Let S' be its derivative, i.e., a polynomial of degree $d-1$, with coefficients $\breve{s} = (s_1, \dots, s_d)^\top \in \mathbb{R}^d$. Let Z and W be SOS polynomials, with coefficients z and w, and assume that S' is non-negative on $[t_0, t_1]$ as a combination of Z and W as in Theorem 3. Moreover, let*

$$D = \operatorname{diag}(1, 2, \dots, d)$$

be the $(d \times d)$ diagonal matrix with $(1, \dots, d)$ as a diagonal elements and denote the bloc-matrices

$$\overline{\mathcal{I}}_{i,d} = \begin{pmatrix} I_d \\ \mathbf{0}_{i,d} \end{pmatrix}, \quad \underline{\mathcal{I}}_{i,d} = \begin{pmatrix} \mathbf{0}_{i,d} \\ I_d \end{pmatrix}, \quad \overline{\underline{\mathcal{I}}}_{i,d} = \begin{pmatrix} \mathbf{0}_{i,d} \\ I_d \\ \mathbf{0}_{i,d} \end{pmatrix}$$

where $\mathbf{0}_{i,d}$ denotes the $(i \times d)$ matrix of zeros, and I_d be the $(d \times d)$ identity matrix. If d is odd, then $z \in \mathbb{R}^d$ and $w \in \mathbb{R}^{d-2}$ and furthermore

$$\breve{s} = Az + Bw$$

where A and B are $(d \times d)$ and $(d \times d-2)$ matrices, respectively. If the degree d of S is even, one has that $z, w \in \mathbb{R}^{d-1}$ and furthermore:

$$\breve{s} = Cz + Dw.$$

where C and D are $(d \times d-1)$ matrices. More specifically,

$$\begin{aligned} A &= \mathcal{D}_d^{-1}, & B &= \mathcal{D}_d^{-1}\left((t_0 + t_1)\overline{\underline{\mathcal{I}}}_{1,d-2} - \underline{\mathcal{I}}_{2,d-2} - t_0 t_1 \overline{\mathcal{I}}_{2,d-2}\right), \\ C &= \mathcal{D}_d^{-1}\left(\underline{\mathcal{I}}_{1,d-1} - t_0 \overline{\mathcal{I}}_{1,d-1}\right), & D &= \mathcal{D}_d^{-1}\left(t_1 \overline{\mathcal{I}}_{1,d-1} - \underline{\mathcal{I}}_{1,d-1}\right). \end{aligned}$$

Proof of Lemma 7. First, assume that S is a polynomial of odd degree $d = 2p + 1$, meaning that its derivative, S', is a polynomial of even degree $2p$. From Theorem 3, one has that $S'(x)$ is positive on an interval $[t_0, t_1]$ if and only if it can be expressed as :

$$S'(x) = Z(x) + (x - t_0)(t_1 - x)W(x)$$

where Z is an SOS polynomial of degree at most equal to $d-1$ and W is an SOS polynomial of degree at most equal to $d-3$. Denote $\breve{s} = (s_1, \dots, s_d) \in R^d$ the coefficients of S' and $z = (z_1, \dots, z_d) \in \mathbb{R}^d$ and $w = (w_1, \dots, w_{d-2)} \in \mathbb{R}^{d-2}$ the coefficients of Z and W respectively. One has that :

$$\begin{aligned} S'(x) &= \sum_{i=1}^{d} i s_i x^{i-1} \\ &= \sum_{j=0}^{d-1} (j+1) s_{i+1} x^i \end{aligned}$$

and if S' is assumed to be non-negative on $[t_0, t_1]$

$$\begin{aligned} S'(x) &= Z(x) + (x - t_0)(t_1 - x)W(x) \\ &= \sum_{j=0}^{d-1} z_{j+1} x^j + (-x^2 + (t_0 + t_1)x - t_0 t_1) \sum_{j=0}^{d-3} w_{j+1} x^j \end{aligned}$$

leading to the following identification :

$$\begin{cases} s_1 = z_1 - t_0 t_1 w_1 \\ s_2 = \frac{1}{2}\left(z_2 - t_0 t_1 w_2 + (t_0 + t_1) w_1\right) \\ s_i = \frac{1}{i}\left(z_i - t_0 t_1 w_i + (t_0 + t_1) w_{i-1} - w_{i-2}\right), \quad \text{for } i = 3, \ldots, d-2 \\ s_{d-1} = \frac{1}{d-1}\left(z_{d-1} + (t_0 + t_1) w_{d-2} - w_{d-3}\right) \\ s_d = \frac{1}{d}\left(z_{d-1} - w_{d-2}\right), \end{cases}$$

or, written in a matrix form:

$$\breve{s} = \mathcal{D}_d^{-1}\left(z + \left((t_0 + t_1)\overline{\mathcal{I}}_{1,d-2} - \underline{\mathcal{I}}_{2,d-2} - t_0 t_1 \overline{\mathcal{I}}_{2,d-2}\right) w\right).$$

If S is assumed to be a polynomial of even degree $d = 2p$, S' is necessarily odd degree. From Theorem 3, one has that $S'(x)$ is positive on an interval $[t_0, t_1]$ if and only if it can be expressed as :

$$S'(x) = (x - t_0)Z(x) + (t_1 - x)W(x)$$

where Z and W are SOS polynomials of degree at most equal to $d-2$ with $z = (z_1, \ldots, z_{d-1}) \in \mathbb{R}^{d-1}$ and $w = (w_1, \ldots, w_{d-1)} \in \mathbb{R}^{d-1}$ as coefficients, respectively. It leads to the following identification:

$$\begin{cases} s_1 = -t_0 z_1 + t_1 w_1 \\ s_i = \frac{1}{i}\left(z_{i-1} - t_0 z_i + t_1 w_i - w_{i-1}\right) \quad \text{for } i = 2, \ldots, d-1 \\ s_d = \frac{1}{d}\left(z_{d-1} - w_{d-1}\right), \end{cases}$$

which can be written in matrix form as

$$\breve{s} = \mathcal{D}_d^{-1}\left(\left(\underline{\mathcal{I}}_{1,d-1} - t_0 \overline{\mathcal{I}}_{1,d-1}\right) z + \left(t_1 \overline{\mathcal{I}}_{1,d-1} - \underline{\mathcal{I}}_{1,d-1}\right) w\right).$$

□

We can now proceed to prove Theorem 1.

Proof of Theorem 1 (rationale). This rationale can be broken down in two steps: **(a)** proving that the objective function (18) can indeed be written in a quadratic form, and:**(b)** proving that the problem constraints form a feasible set in $\mathbb{R}^{d+1}$ which is closed and convex.

(a) Notice first that the initial objective function

$$\int_{t_0}^{t_1} (L(x) - F_P^{\rightarrow}(x))^2 dx$$

where $L \in \mathbb{R}[x]_{\leq d}$ with coefficients $s \in \mathbb{R}^{d+1}$, can be rewritten as:

$$\begin{aligned}
\int_{t_0}^{t_1} (F_P^{\rightarrow}(x) - L(x))^2 dx &= \int_{t_0}^{t_1} (\sum_{i=0}^{d} s_i x^i - F_P^{\rightarrow}(x))^2 dx \\
&= \int_{t_0}^{t_1} \left(\left(\sum_{i=0}^{d} s_i x^i \right)^2 + (F_P^{\rightarrow}(x))^2 - 2 \sum_{i=0}^{d} s_i x^i F_P^{\rightarrow}(x) \right) dx \\
&= \int_{t_0}^{t_1} \left(\sum_{i=0}^{d} s_i x^i \right)^2 dx - 2 \sum_{i=0}^{d} s_i \int_{t_0}^{t_1} x^i F_P^{\rightarrow}(x) dx \\
&\quad + \int_{t_0}^{t_1} (F_P^{\rightarrow}(x))^2 dx.
\end{aligned}$$

Note that

$$\begin{aligned}
\int_{t_0}^{t_1} \left(\sum_{i=0}^{d} s_i x^i \right)^2 dx &= \sum_{i=0}^{d} \sum_{j=0}^{d} s_i s_j \int_{t_0}^{t_1} x^{i+j} dx \\
&= s^\top M s
\end{aligned}$$

where M is the moment matrix of the Lebesgue measure on $[t_0, t_1]$, i.e., defined entry-wise, for $i, j = 1, \ldots, d+1$ as

$$M_{ij} = \int_{t_0}^{t_1} x^{i+j-2} dx = \frac{(t_1)^{i+j-1} - (t_0)^{i+j-1}}{i+j-1}.$$

and further notice that M is thus positive definite since, for any $u \in \mathbb{R}^{d+1}$,

$$u^\top M u = \int_{t_0}^{t_1} \left(\sum_{i=0}^{d} u_{i+1} x^i \right)^2 dx \geq 0$$

is always non-negative, and equal to 0 if and only if $u_i = 0, i = 1, \ldots, d+1$. Moreover, note that:

$$\sum_{i=0}^{d} s_i \int_{t_0}^{t_1} x^i F_P^{\rightarrow}(x) dx = s^\top r$$

where $r \in \mathbb{R}^{d+1}$ is the moment vector of G with respect to the Lebesgue measure on $[t_0, t_1]$, defined for $i = 0, \ldots, d$ as:

$$r_i = \int_{t_0}^{t_1} x^i F_P^{\rightarrow}(x) dx$$

Since a polynomial is completely characterized by its coefficients, searching for:

$$S^* = \operatorname*{argmin}_{L \in \mathbb{R}[x]_{\leq d}} \int_{t_0}^{t_1} (L(x) - F_P^{\rightarrow}(x))^2 dx$$

is equivalent to finding the coefficients s^* of S^*, i.e.,

$$s^* = \operatorname*{argmin}_{s \in \mathbb{R}^{p+1}} s^\top M s - 2 s^\top r$$

and thus proving the first part of the proposition.

(b) Notice that the interpolation constraints

$$\begin{cases} S(t_0) = b_0 \\ S(t_1) = b_1 \end{cases}$$

can be written as

$$\begin{cases} s^\top \mathbf{t_0}^d = b_0 \\ s^\top \mathbf{t_1}^d = b_1 \end{cases}$$

where, for $a \in \mathbb{R}$, one denote $\mathbf{a}^d$ the vector of powers of a up to d, i.e., $\mathbf{a}^d = (1, a, \dots, a^{d-1}, a^d) \in \mathbb{R}^{d+1}$. Moreover, by letting:

$$\mathbf{T} = \begin{pmatrix} \mathbf{t_0}^d \\ \mathbf{t_1}^d \end{pmatrix}, \quad b = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix},$$

where $\mathbf{T}$ is a $(2 \times d + 1)$ bloc-matrix, the constraint can be written as:

$$Ts = b.$$

Furthermore, notice that

$$\mathcal{C}_0 = \{s \in \mathbb{R}^{d+1} \mid Ts = b\}$$

is a convex subset of $\mathbb{R}^{d+1}$, since the equality constraints are linear. Concerning the monotonicity constraint

$$S'(x) \geq 0, \quad \forall x \in [t_0, t_1],$$

from Lemma 7 we can quite generically write

$$\begin{pmatrix} s_d \\ \vdots \\ s_1 \end{pmatrix} = T_0(z, w) := Az + Bw$$

where z and w are the coefficient of SOS polynomials of degrees depending on d. Additionally, notice that the mapping $T_0 : \mathbb{R}^d \times \mathbb{R}^{d-2} \to \mathbb{R}^d$ is linear. Next, let $V_1 : \mathbb{S}_p \to \mathbb{R}^{2p}$, and $V_2 : \mathbb{S}_q \to \mathbb{R}^{2q}$ be defined as in (A.3), where $p = d - 1/2$ and $q = d - 3/2$ if d is odd, or $p = d - 2/2$ and $q = d - 2/2$ if d is even, and note that both mappings are linear thanks to Lemma 6.

Moreover, denote the following sets:

$$\mathcal{Z} = \{V_1(E) \mid E \in \Sigma_p\}, \quad \mathcal{W} = \{V_2(E) \mid E \in \Sigma_{p-1}\}$$

and notice the polynomial Z (resp. W) is SOS if and only its coefficients z (resp. w) are in $\mathcal{Z}$ (resp. $\mathcal{W}$) thanks to Theorem 5. In addition again, notice that, thanks to Lemma 4, and due to the fact that Σ_p is a closed convex set in $\mathbb{S}_p$ as per Theorem 5, both $\mathcal{Z}$ and $\mathcal{W}$ are convex subsets of $\mathbb{R}^{2p}$ and $\mathbb{R}^{2q}$ respectively. Besides, thanks to Lemma 5, the set $\mathcal{Z} \times \mathcal{W}$ is a convex subset of $\mathbb{R}^{2p} \times \mathbb{R}^{2q}$ as well. Moreover, let

$$\mathcal{C}_1 = \left\{ \begin{pmatrix} T_0(w, z) \\ x \end{pmatrix} \in \mathbb{R}^{d+1} \mid x \in \mathbb{R}, \quad (z, w) \in \mathcal{Z} \times \mathcal{W} \right\}$$

and note that it is a convex subset of $\mathbb{R}^{d+1}$ due to the fact that T_0 is a linear map.

Finally, since both $\mathcal{C}_0$ and $\mathcal{C}_1$ are convex sets, their intersection:

$$\mathcal{K} = \mathcal{C}_0 \cap \mathcal{C}_1$$

is as well, and note that any element $s \in \mathcal{K}$ are the coefficients of a polynomial respecting both equality and monotonicity constraints. In other words, $\mathcal{K}$ is the feasible set of coefficients of the initial optimization problem. □

Appendix B. Computing moment vector of arbitrary quantile functions

One wishes here at computing the vector described in (20). In the case where P is an empirical measure built on a n-sample, one has that for $[t_0, t_1] \in [0,1]$, $i = 0, \dots, p$:

$$\begin{aligned} r_i = & \frac{1}{i+1}\left[\sum_{j\in J} \frac{X_{(j)}}{n^{i+1}}\left((j+1)^{i+1} - j^{i+1}\right)\right. \\ & + X_{(\bar{j})}\left(t_1^{i+1} - \left(\frac{\bar{j}}{n}\right)^{i+1}\right) \\ & \left. + X_{(\underline{j}-1)}\left(\left(\frac{\underline{j}}{n}\right)^{i+1} - t_0^{i+1}\right)\right] \end{aligned}$$

where $J = \{i \in \mathbb{N} \mid \lfloor nt_0 \rfloor < i < \lfloor nt_1 \rfloor\}$, $\bar{j} = \lfloor t_1 n \rfloor$, $\underline{j} = \lfloor t_0 n \rfloor + 1$, and where $X_{(j)}$ denotes the j-th order statistic of the observe sample. In cases where $F_P^{\leftarrow}$ is continuous, it is possible to use numerical quadrature methods in order to evaluate each integral composing the elements r_i of r.