ROBUSTNESS ASSESSMENT OF BLACK-BOX MODELS

QUANTILE-CONSTRAINED WASSERSTEIN PROJECTIONS AND ISOTONIC POLYNOMIAL APPROXIMATIONS

1EDF Lab Chatou - Département PRISME
2Institut de Mathématiques de Toulouse
3SINCLAIR AI Lab

École Thématique sur les Incertitudes en Calcul Scientifique
GdR MASCOT-NUM
Golfe de Lozari, Corsica - October 2022

Marouane Il Idrissi123, Nicolas Bousquet13, Fabrice Gamboa2, Bertrand Iooss123, Jean-Michel Loubes2.
Introduction

**Goal:** Enhance the confidence in the practical usage of a black-box model, by assessing its robustness to input perturbations.

**Challenges:**

1. Define *generic*, but *understandable* input perturbations.
2. Unify ML interpretability and sensitivity analysis (SA)
   - ML: Features are modelled as *empirical probability measures*
   - SA: Inputs are modelled as *probability measures admitting a positive density*.
3. Local/Global robustness assessment of a model, or some of its key characteristics.

**Illustrative example:** Epistemic uncertainty on a riverbed’s roughness near an industrial site.
Let $P \in \mathcal{P}(\mathbb{R}^d)$ be an initial probability measure. We seek the solution of the projection problem

$$Q = \arg\min_{G \in \mathcal{P}(\mathbb{R}^d)} \mathcal{D}(P, G)$$

s.t. $G \in \mathcal{C}$, and $C_P = C_Q$

where $\mathcal{C} \subseteq \mathcal{P}(\mathbb{R}^d)$ is a perturbation class, and $\mathcal{D}$ a discrepancy between probability measures. Ideally, $P$ and $Q$ must have the same copula.

ML interpretability (Bachoc et al. 2020) and SA (Lemaître et al. 2015) work focus on the Kullback-Leibler divergence (KL) as a discrepancy, and generalized moments perturbations.
Context

Let \( P \in \mathcal{P}(\mathbb{R}^d) \) be an initial probability measure. We seek the solution of the projection problem

\[
Q = \arg\min_{G \in \mathcal{P}(\mathbb{R}^d)} \mathcal{D}(P, G)
\]

s.t. \( G \in \mathcal{C} \), and \( C_P = C_Q \)

where \( \mathcal{C} \subseteq \mathcal{P}(\mathbb{R}^d) \) is a perturbation class, and \( \mathcal{D} \) a discrepancy between probability measures. Ideally, \( P \) and \( Q \) must have the same copula.

ML interpretability (Bachoc et al. 2020) and SA (Lemaître et al. 2015) work focus on the Kullback-Leibler divergence (KL) as a discrepancy, and generalized moments perturbations.

Drawbacks:

- Generalized moments may not exist.
- Different results depending on \( P \) due to KL.
Let $P \in \mathcal{P}(\mathbb{R}^d)$ be an initial probability measure. We seek the solution of the projection problem

$$Q = \arg\min_{G \in \mathcal{P}(\mathbb{R}^d)} \mathcal{D}(P, G)$$

s.t. $G \in \mathcal{C}$, and $C_P = C_Q$

where $\mathcal{C} \subseteq \mathcal{P}(\mathbb{R}^d)$ is a perturbation class, and $\mathcal{D}$ a discrepancy between probability measures. Ideally, $P$ and $Q$ must have the same copula.

ML interpretability (Bachoc et al. 2020) and SA (Lemaître et al. 2015) work focus on the Kullback-Leibler divergence (KL) as a discrepancy, and generalized moments perturbations.

**Drawbacks:**
- Generalized moments may not exist.
- Different results depending on $P$ due to KL.

**Solutions:**
- Quantile perturbation class.
- 2-Wasserstein: does not depend on the nature of $P$. 
**Why quantiles?**

**Generalized quantile functions** are the generalized inverses (de la Fortelle 2015) of the cdf of random variables.

\[
F_P^{-}(a) = \sup \{ t \in \mathbb{R} \mid F_P(t) < a \} = \inf \{ t \in \mathbb{R} \mid F_P(t) \geq a \},
\]

\[
F_P^{+}(a) = \sup \{ t \in \mathbb{R} \mid F_P(t) \leq a \} = \inf \{ t \in \mathbb{R} \mid F_P(t) > a \},
\]

- They **characterize** probability measures (Dufour 1995)
- Univariate quantiles **always exist**.
The **quantile perturbation class** $Q_V$ is defined using constraints of the form

$$F_Q^\leftarrow(\alpha) \geq b \geq F_Q^\rightarrow(\alpha).$$

with $b \in \mathbb{R}$, and leading to the set

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

included in $\mathcal{P}(\mathbb{R})$, and where $V \subseteq \mathcal{F}^\rightarrow$ is a **(smoothing) restriction** on the **space of quantile functions**.
Quantile perturbation class

The quantile perturbation class $Q_V$ is defined using constraints of the form

$$F_Q^\leftarrow (\alpha) \geq b \geq F_Q^\rightarrow (\alpha).$$

with $b \in \mathbb{R}$, and leading to the set

$$Q_V = \{ Q \in \mathcal{P}(\mathbb{R}) | F_Q^\leftarrow \in V, \quad F_Q^\leftarrow (\alpha_i) \geq b_i \geq F_Q^\rightarrow (\alpha_i), \quad i = 1, \ldots, K \}.$$

included in $\mathcal{P}(\mathbb{R})$, and where $V \subseteq \mathcal{F}^\leftarrow$ is a (smoothing) restriction on the space of quantile functions.

Collections of perturbations can be driven by an intensity parameter $\theta \in [-1, 1]$

- **Quantile shift**: shifting the $\alpha$-quantile of $P$ between two values.
- **Operating domain dilatation**: widewing or narrowing the bounds of the support of $P$ w.r.t. a scaling parameter $\eta \in \mathbb{R}$.

Additional pointual modelling constraints can also be added (e.g., preservation of empirical quantiles, expert knowledge).
The Wasserstein distance

For two probability measure $P, Q \in \mathcal{P}(\mathbb{R}^d)$ having the same copula (Alfonsi and Jourdain 2014):

$$W^p_P(P, Q) = \sum_{i=1}^{d} W^p_{P_i, Q_i}.$$  \hfill (1)

where each $P_i, Q_i \in \mathcal{P}(\mathbb{R})$ is a marginal distribution. Each element of the sum reduces to (Santambrogio 2015):

$$W^p_{P_i, Q_i} = \int_{0}^{1} |F_{P_i}^{-1}(x) - F_{Q_i}^{-1}(x)|^p \, dx$$

whatever the “nature” of $P$ (empirical, continuous...).
The Wasserstein distance

For two probability measure \( P, Q \in \mathcal{P}(\mathbb{R}^d) \) having the same copula (Alfonsi and Jourdain 2014):

\[
W^p_P(P, Q) = \sum_{i=1}^{d} W^p_{P_i, Q_i}.
\]  

(1)

where each \( P_i, Q_i \in \mathcal{P}(\mathbb{R}) \) is a marginal distribution. Each element of the sum reduces to (Santambrogio 2015):

\[
W^p_{P_i, Q_i} = \int_0^1 \left| F^{-1}_{P_i}(x) - F^{-1}_{Q_i}(x) \right|^p dx
\]

whatever the “nature” of \( P \) (empirical, continuous...).

In particular, the 2-Wasserstein distance metricizes weak convergence on the set of probability measure with finite 2nd order moments \( \mathcal{P}_2(\mathbb{R}) \) (Villani 2003).
The Wasserstein distance

For two probability measure $P, Q \in \mathcal{P}(\mathbb{R}^d)$ having the same copula (Alfonsi and Jourdain 2014):

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

(1)

where each $P_i, Q_i \in \mathcal{P}(\mathbb{R})$ is a marginal distribution. Each element of the sum reduces to (Santambrogio 2015):

$$W_p^p(P_i, Q_i) = \int_{0}^{1} |F_{P_i}^{-1}(x) - F_{Q_i}^{-1}(x)|^p \, dx$$

whatever the “nature” of $P$ (empirical, continuous...).

In particular, the 2-Wasserstein distance metricizes weak convergence on the set of probability measure with finite 2nd order moments $\mathcal{P}_2(\mathbb{R})$ (Villani 2003).

- **Solving $d$ univariate perturbation problems.**
- **Optimal transportation map preserves the copula:** $T_i = (F_{Q_i}^{-1} \circ F_{P_i})$
Wasserstein and $L^2$ projections

Hence, one focuses on the marginal perturbation problem:

$$Q = \arg\min_{G \in \mathcal{P}(\mathbb{R})} \ W_2(P, G)$$

subject to \( G \in Q \mathcal{V} \)

(2)

**Proposition**

The solution \( Q \) of the problem in Eq. (2) is uniquely characterized by its quantile function being the solution

$$F_Q^\leftarrow = \arg\min_{L \in L^2([0,1])} \int_0^1 (L(x) - F_P^\rightarrow (x))^2$$

subject to

\( L(\alpha_i) \leq b_i \leq L(\alpha_i^+) \), \( i = 1, \ldots, K \), \( L \in \mathcal{V} \)
Solving the perturbation problem

If $V = F^\leftarrow$, there exists a unique analytical solution $Q$ to the problem:

$Q$ is the same as $P$, except on the intervals between $F_P^\leftarrow(\alpha_i)$ and $b_i$ which have no mass, and an atom is added at $b_i$, taking the initial mass of the interval.
Solving the perturbation problem

If $\nu = F^\leftarrow$, there exists a **unique analytical solution** $Q$ to the problem:

$Q$ is the same as $P$, except on the intervals between $F_P^\leftarrow(\alpha_i)$ and $b_i$ which have no mass, and an atom is added at $b_i$, taking the initial mass of the interval.

How to explicitly enforce “smoothness” to the resulting perturbed quantile function?
Isotonic interpolating piece-wise continuous polynomials

**Idea:** Using piece-wise continuous polynomials of degree $p$ to ensure continuity.

Partition $[0, 1]$ according into interval $[t_j, t_{j+1}], i = 0, \ldots, K$ with $t_0 = 0$, $t_{K+1} = 1$, and $t_i = \alpha_i$ (ordered increasingly), and solve for

$$S = \arg\min_{G \in \mathbb{R}[x]_{\leq p}} \int_{t_i}^{t_{i+1}} (F_P(x) - G(x))^2 \, dx$$

s.t.

$$G(t_i) = b_i, G(t_{i+1}) = b_{i+1}$$

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

(3)

**Proposition**

The polynomial solution of Eq. (3) admits as coefficients

$$s^* = \arg\min_{s \in \mathbb{R}^{p+1}} s^\top Ms - 2s^\top r$$

s.t.

$$s \in K$$

where $M$ is the moment matrix of the Lebesgue measure on $[t_i, t_{i+1}]$, $r$ is the moment vector of $F_P$, and $K$ is a closed convex subset of $\mathbb{R}^{p+1}$. 8/17
It is a **Convex Constrained Quadratic Problem** which can be solved using numerical solvers (e.g., CVXR (Fu, Narasimhan, and Boyd 2020)).

Each marginal input $X_i \sim P_i$ can be perturbed using the optimal monotone perturbation map

$$\tilde{X}_i = T_i(X_i) = (F_{Q_i}^{-1} \circ F_{P_i})(X_i)$$

preserving the (empirical) copula between all the inputs.
SIPA framework for model-agnostic interpretation

Our methodology follows the SIPA framework (Scholbeck et al. 2020):

1. **Sampling**: Observed (ML) or simulated (UQ) values of $P$.
2. **Intervention**: Define optimal perturbations under quantile constraints and apply the perturbation map, resulting in perturbed inputs $\tilde{X} = T(X)$ with the same dependence structure.
3. **Prediction**: Evaluate the model $G$ (numerical in UQ, learned in ML) on the perturbed inputs.
4. **Aggregation**: Estimate local or global statistics on the perturbed output $\tilde{Y} = G(\tilde{X})$. 
Simplified hydrological model

Model of the water level of a river. Simplification of the one-dimensional Saint-Venant equation, with a uniform and constant flow rate (Iooss and Lemaître 2015; Fu, Couplet, and Bousquet 2017)

- $Q$: River maximum annual water flow rate.
- $K_s$: Strickler riverbed roughness coefficient.
- $Z_v$: Downstream river level.
- $Z_m$: Upstream river level.
- $L$: River length.
- $B$: River width.

Model:

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

Gaussian copula with covariance matrix:

$$R_P = \begin{pmatrix}
1 & 0.5 & 0 & 0 & 0 & 0 & 0 \\
0.5 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0.3 & 0 & 0 & 0 \\
0 & 0 & 0.3 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0.3 & 0 \ \\
0 & 0 & 0 & 0 & 0.3 & 1 & 0
\end{pmatrix}$$
Perturbation strategy

Ponctual perturbations

**Q:**
- 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 $125$.

**L:**
- Shift the application domain from $[4990, 5010]$ to $[4988, 5012]$.
- Preserve the median of the distribution.

**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 $0.05$.

Optimal perturbation problems are solved with polynomial smoothing (arbitrary degree equal to 12).
Shapley effects

Double Monte Carlo estimation with $N_v = 10^5$, $N_o = 3 \times 10^3$ and $N_i = 300$. 
Conclusion & perspectives

Generic and interpretable marginal perturbation scheme.

Local and global robustness assessment of black-box numerical (SA) and predictive models (ML).

Perspectives:

- Optimal degree selection, and derivability of the resulting polynomial.
- Multivariate quantile perturbation.
- More general smoothing spaces (monotone Sobolev functions, RKHS).
Conclusion & perspectives

Generic and interpretable marginal perturbation scheme.

Local and global robustness assessment of black-box numerical (SA) and predictive models (ML).

Perspectives:

- Optimal degree selection, and derivability of the resulting polynomial.
- Multivariate quantile perturbation.
- More general smoothing spaces (monotone Sobolev functions, RKHS).

More details and ML application (Acoustic Fire Extinguisher) in our pre-print (HAL/arXiv) (I. et al. 2022):

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

Marouane Il Idrissi\textsuperscript{a,b,c,e}, Nicolas Bousquet\textsuperscript{a,b,d}, Fabrice Gamboa\textsuperscript{e}, Bertrand Iooss\textsuperscript{a,b,c}, Jean-Michel Loubès\textsuperscript{c}


Thank you for your attention!

Any questions?
River water level punctual perturbations

- **Perturbation of $Q$**
- **Perturbation of $L$**
- **Perturbation of $Z_m$**

**Wasserstein Projection**
**Initial Quantile Function**
**Interpolation points**
**Initial Application Domain**
Let $X \subseteq \mathbb{R}^d$, for $d$ a positive integer, and $P \in \mathcal{P}(\mathcal{X})$. Let $Q_i$ be the solution of the optimal projection problem with $C = C_V$, for every marginal distribution $P_i$ of $P$, $i = 1, \ldots, d$, and where $V \subseteq \otimes_{j=1}^d F_j^{-1}$. Let the random vectors

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

where

$$T : \mathcal{X} \rightarrow \mathcal{X}$$

$$\begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \mapsto \begin{pmatrix} T_1(x_1) \\ \vdots \\ T_d(x_d) \end{pmatrix}$$

(4)

where

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

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

2. If $P$ is atomless, and assuming additionally that $V$ is such that every $F_{Q_j}^{-1}$, $i = 1, \ldots, d$ is strictly increasing, then the random vectors $X$ and $\tilde{X}$ have the same copula. Moreover, each perturbed marginal $\tilde{X}_i \sim Q_i$. 

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

\[ c_1 = \min(\beta_1, \alpha_1), \quad c_i = \min \left[ \max(\alpha_{i-1}, \beta_i), \alpha_i \right], i = 2, \ldots, K, \]
\[ d_K = \max(\beta_K, \alpha_K), \quad d_j = \max \left[ \min(\beta_j, \alpha_{j+1}), \alpha_j \right], j = 1, \ldots, K - 1. \]

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

\[ F_{Q}^{-} (y) = \begin{cases} F_P^{-}(y) & \text{if } y \in \overline{A}, \\ b_i & \text{if } y \in A_i, \quad i = 1, \ldots, K. \end{cases} \]
Theorem (Non-negativity of polynomials on closed intervals)

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

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,
\[
\forall x \in [t_0, t_1] \quad S(x) = Z(x) + (x - t_0)(t_1 - x)W(x)
\]
where \( Z \) is an SOS polynomial of degree at most equal to \( d \), and \( W \) is an SOS polynomial of degree at most equal to \( d - 2 \).

A 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,
\[
\forall x \in [t_0, t_1] \quad S(x) = (x - t_0)Z(x) + (t_1 - x)W(x)
\]
where \( Z, W \) are SOS polynomials of degree at most equal to \( d \).
SDP representation of SOS polynomials

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

$$
\Gamma = \begin{bmatrix} 
\Gamma_{ij} 
\end{bmatrix}_{i,j=1,\ldots,p}
$$

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

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

Moreover, for $k = 0, \ldots, d$, let $\Pi^p_k$ be the $(p \times p)$ matrix defined by, for $i, j = 1, \ldots, p$:

$$
\left[ \Pi^p_k \right]_{i,j} = 1_{\{i+j=k+2\}}(i,j).
$$

If there exists a matrix $\Gamma$ such that $S$ is SOS, then one has that, for $i = 0, \ldots, d$

$$
s_i = \langle \Pi^p_i, \Gamma \rangle_F = \sum_{j+k=i+2} \Gamma_{j,k}
$$

where, $\langle ., . \rangle_F$ denotes the Frobenius norm on matrices.
Equivalent optimization formulation

Let \([t_0, t_1] \subset [0, 1]\), and let \(s = (s_0, \ldots, s_d)^\top \in \mathbb{R}^{d+1}\), \(M\) be the symmetric \((d+1 \times d+1)\) moment matrix of the Lebesgue measure on \([t_0, t_1]\), i.e. for \(i, j = 1, \ldots, d + 1\),

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

and denote \(r \in \mathbb{R}^{d+1}\) the moment vector of \(A(x)\), i.e., for \(i = 0, \ldots, d\)

\[
r_i = \int_{t_0}^{t_1} x^i F^{-\to}_p(x) dx
\]

Then, the optimization problem can be equivalently solved by finding \(s\) as being the solution of the following convex constrained quadratic program,

\[
s^* = \arg\min_{s \in \mathbb{R}^{p+1}} s^\top Ms - 2s^\top r
\]
\[
\text{s.t. } s \in \mathcal{K}
\]

where \(\mathcal{K}\) is a closed convex subset of \(\mathbb{R}^{p+1}\).