Robustness Assessment of Black-box Models

Quantile-constrained Wasserstein Projections and Isotonic Polynomial Approximations

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

53èmes Journées de la Statistique de la SFdS
Lyon, Jeudi 16 Juin 2022

Marouane Il Idrissi[123], Nicolas Bousquet[13], Fabrice Gamboa[2], Bertrand Iooss[123], Jean-Michel Loubes[2].

Introduction

Goal: Optimally perturb a black-box model's input under **distributional constraints**, and **assess its robustness** w.r.t. to it.

Challenges:

1. Define **generic and interpretable** black-box model 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. Produce robustness to perturbation diagnostics.

Application: Classification task (neural network) of an acoustic fire extinguisher.

Context

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

$$Q = \underset{G \in \mathcal{P}(\mathbb{R})}{\operatorname{argmin}} \quad \mathcal{D}\left(P, G\right)$$
$$\text{s.t.} \quad G \in \mathcal{C}$$

where $\mathcal{C} \subseteq \mathcal{P}(\mathbb{R})$ is a **perturbation class**, and $\mathcal{D}$ a discrepancy between probability measures.

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

$$Q = \underset{G \in \mathcal{P}(\mathbb{R})}{\text{argmin}} \quad \mathcal{D}\left(P, G\right) \qquad \text{s.t.} \quad G \in \mathcal{C}$$

where $\mathcal{C} \subseteq \mathcal{P}(\mathbb{R})$ is a **perturbation class**, and $\mathcal{D}$ a discrepancy between probability measures.

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**
- KL divergence **implicitly smoothes results**

Context

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

$$Q = \underset{G \in \mathcal{P}(\mathbb{R})}{\text{argmin}} \quad \mathcal{D}\left(P, G\right)$$
$$\text{s.t.} \quad G \in \mathcal{C}$$

where $\mathcal{C} \subseteq \mathcal{P}(\mathbb{R})$ is a **perturbation class**, and $\mathcal{D}$ a discrepancy between probability measures.

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**
- KL divergence **implicitly smoothes results**

Solutions:

- **Quantile** perturbation class
- 2-Wasserstein distance and **explicit smoothing**

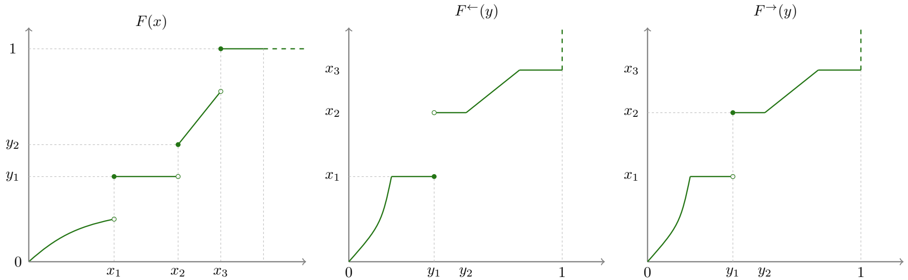
Why quantiles ?

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

$$F_P^{\leftarrow}(a) = \sup\{t \in \mathbb{R} \mid F_P(t) < a\} = \inf\{t \in \mathbb{R} \mid F_P(t) \geq a\}.$$

$$F_P^{\rightarrow}(a) = \sup\{t \in \mathbb{R} \mid F_P(t) \leq a\} = \inf\{t \in \mathbb{R} \mid F_P(t) > a\},$$

- They **characterize** probability measures (Dufour 1995)
- $\mathcal{F}^{\leftarrow}$ the space of left-continuous, non-decreasing functions on $[0, 1]$ is **uniquely linked** to $\mathcal{P}(\mathbb{R})$.



Quantile perturbation class

The **quantile perturbation class** $\mathcal{Q}_\mathcal{V}$ 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

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

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

Quantile perturbation class

The **quantile perturbation class** $\mathcal{Q}_{\mathcal{V}}$ 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

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

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

Perturbations can be driven by an **intensity parameter** $\theta \in [-1, 1]$

- **Quantile shift:** shifting the α-quantile of P between two values.
- **Operating domain dilatation:** widewing or narrowing the bounds of the support of P.

Additional **modelling constraints** can also be added (e.g., preservation of empirical quantiles, expert knowledge).

The Wasserstein distance

The p-Wasserstein distance between P and Q is the quantity defined by

$$W_p(P,Q) = \left(\inf_{\pi \in \Pi(P,Q)} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} \|x-y\|_p^p d\pi(x,y) \right\} \right)^{1/p}$$

where $\Pi(P,Q)$ is the set of probability couplings, with P and Q as its marginals, i.e.,

$$\Pi(P,Q) = \left\{ \pi \in \mathcal{P}(\mathcal{X}\times\mathcal{Y}) \mid \int_{\mathcal{Y}} \pi(dx,dy) = P(dx), \int_{\mathcal{X}} \pi(dx,dy) = Q(dy) \right\}.$$

For two probability measures P and Q in $\mathcal{P}_p(\mathbb{R})$, it simplifies to (Santambrogio 2015)

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

The Wasserstein distance

The p-Wasserstein distance between P and Q is the quantity defined by

$$W_p(P,Q) = \left(\inf_{\pi \in \Pi(P,Q)} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} \|x-y\|_p^p d\pi(x,y) \right\} \right)^{1/p}$$

where $\Pi(P,Q)$ is the set of probability couplings, with P and Q as its marginals, i.e.,

$$\Pi(P,Q) = \left\{ \pi \in \mathcal{P}(\mathcal{X}\times\mathcal{Y}) \mid \int_{\mathcal{Y}} \pi(dx,dy) = P(dx), \int_{\mathcal{X}} \pi(dx,dy) = Q(dy) \right\}.$$

For two probability measures P and Q in $\mathcal{P}_p(\mathbb{R})$, it simplifies to (Santambrogio 2015)

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

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).

Wasserstein and L^2 projections

The perturbation problem becomes

$$Q = \underset{G \in \mathcal{P}(\mathbb{R})}{\operatorname{argmin}} \quad W_2\left(P, G\right) \qquad \text{s.t.} \quad G \in \mathcal{Q}_{\mathcal{V}} \tag{1}$$

Proposition

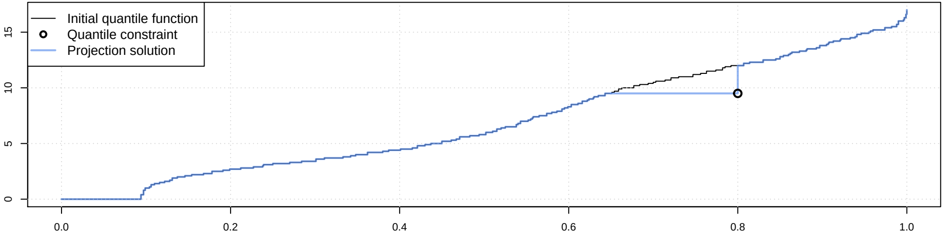
The solution Q of the problem in Eq. (1) is uniquely characterized by its quantile function being the solution

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

Solving the perturbation problem

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

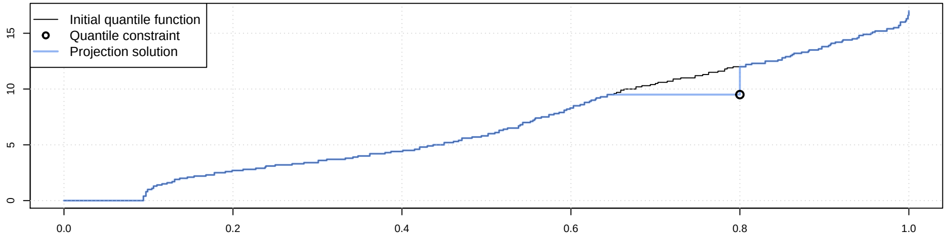
Q is the same as P, except on the intervals between $F_P^{\leftarrow}(\alpha_i)$ and b_i which have no mass, and an atom is added at b_i, taking the initial mass of the interval.



Solving the perturbation problem

If $\mathcal{V} = \mathcal{F}^{\leftarrow}$, 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 add 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

$$
\begin{aligned}
S = \underset{G \in \mathbb{R}[x]_{\leq p}}{\text{argmin}} \quad & \int_{t_i}^{t_{i+1}} (F_P^{\rightarrow}(x) - G(x))^2 dx \\
\text{s.t.} \quad & G(t_i) = b_i, G(t_{i+1}) = b_{i+1} \\
& G'(x) \geq 0, \quad \forall x \in [t_0, t_1]
\end{aligned}
\tag{2}
$$

Proposition

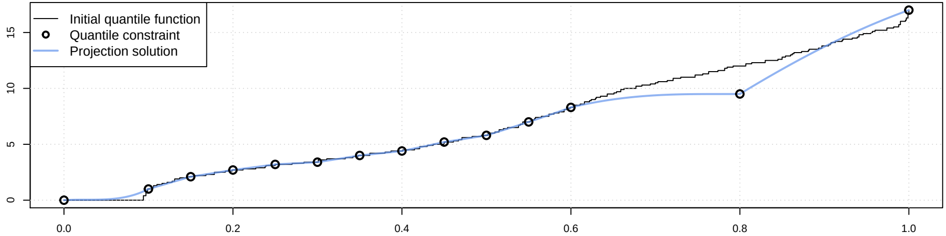
The polynomial solution of Eq. (2) admits as coefficients

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

where M is the moment matrix of the Lebesgue measure on $[t_i, t_{i+1}]$, r is the moment vector of $F_P^{\rightarrow}$, and $\mathcal{K}$ is a closed convex subset of $\mathbb{R}^{p+1}$.

Isotonic interpolation piece-wise continuous polynomials

It is a **Convex Constraint Quadratic Problem** which can be solved numerically using the `CVXR` solver (Fu, Narasimhan, and Boyd 2020).



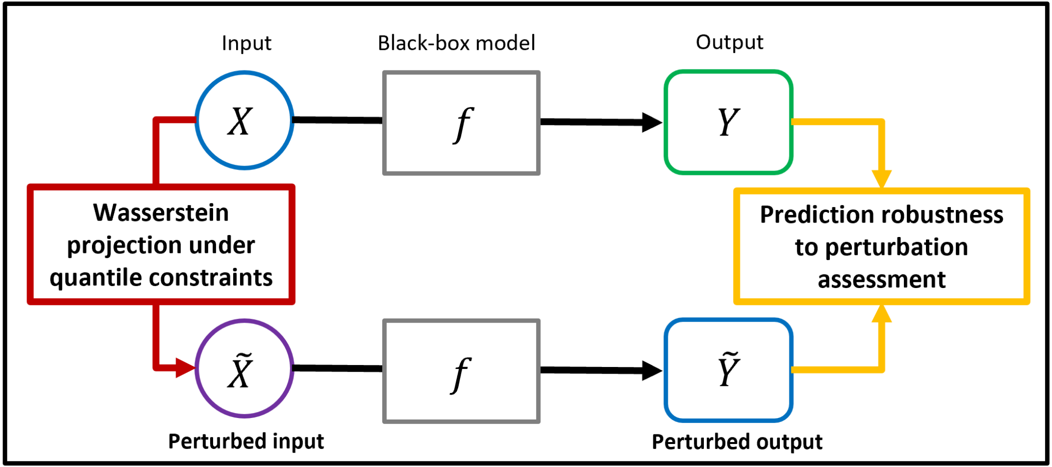
Each marginal input $X_i \sim P_i$ can be perturbed using the monotone perturbation map

$$T_i = (F_{Q_i}^{\leftarrow} \circ F_{P_i})$$

where $\widetilde{X}_i = T_i(X_i) \sim Q_i$, and **preserves the empirical copula** between the model's inputs .

Input perturbation and robustness assessment

Methodology



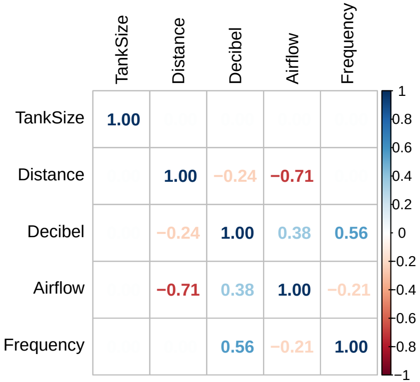
15390 experiments of sound wave fire extinguishing.

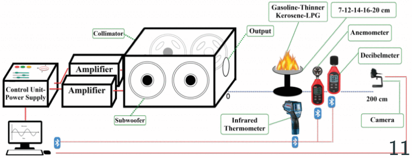
Classification task on 6 variables measured during the experiments.

- Tank Size (L)
- Fuel (Kerosene, Gasoline, Thinner)
- Fire source distance (m)
- Decibel
- Airflow
- Sound frequency

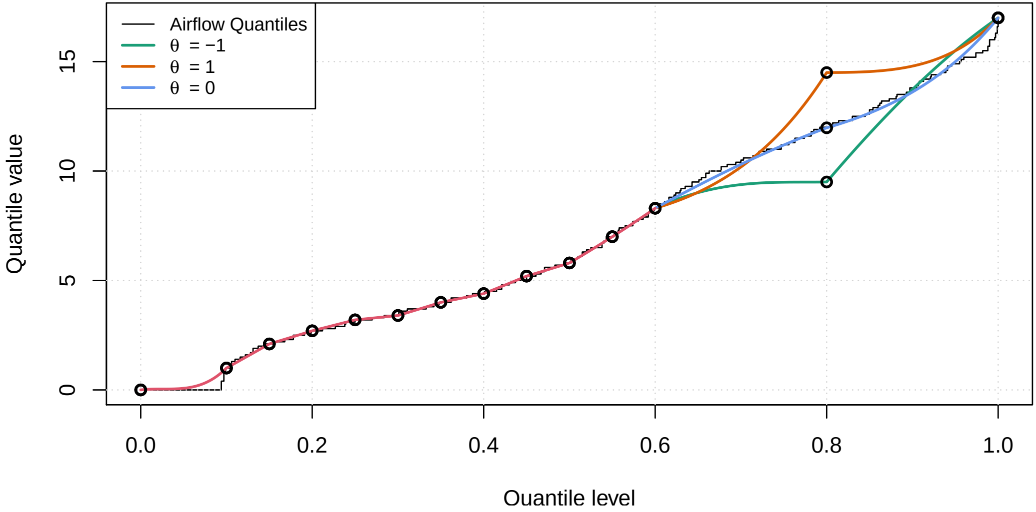
Black-box model: 1-layer neural network (Koklu and Taspinar 2021) trained with an accuracy of 95.15% (validation accuracy of 94.26%).

Perturbation scheme: shift of the Airflow 0.8-quantile: initial value at 12, shift between 9.5 ($\theta = -1$) and 14.5 ($\theta = 1$) by polynomial perturbation approximation of degree 9.

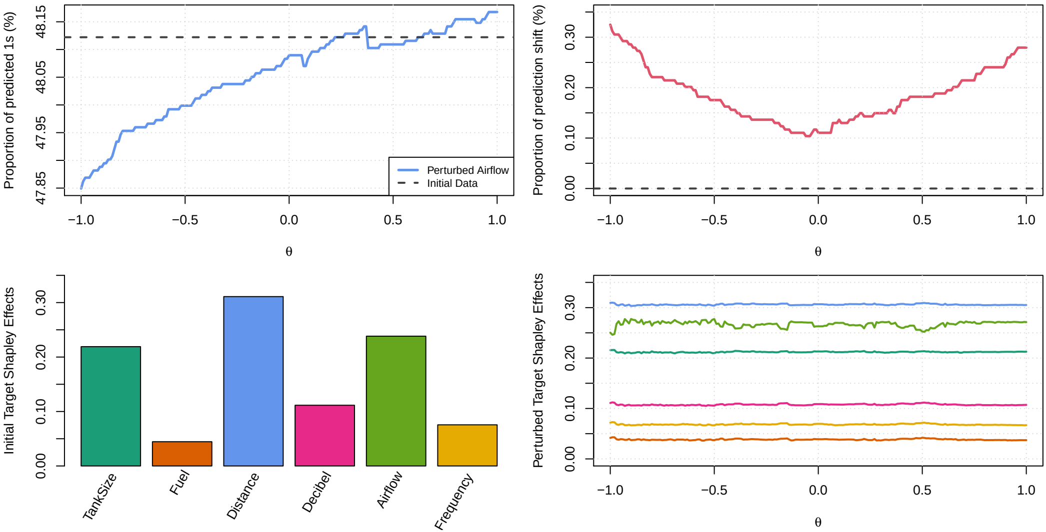




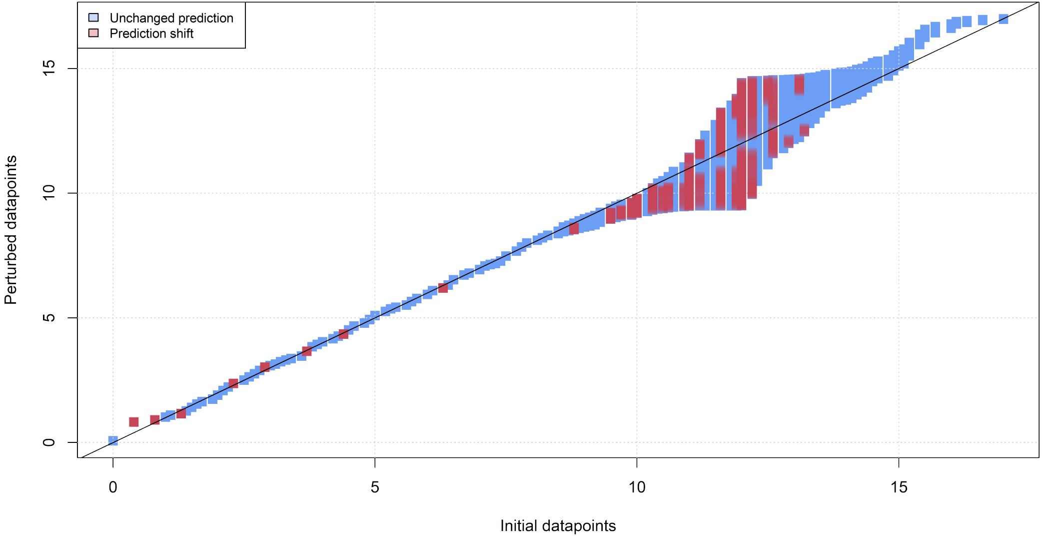
Airflow perturbations



Global robustness



Local robustness



Conclusion & perspectives

Generic and interpretable **marginal perturbation scheme**.

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

Perspectives:

- Parallel and efficient computation in R (soon to be published).
- Optimal degree selection, and derivability of the resulting polynomial.
- Multivariate quantile perturbation (including the copulas), and other discrepancies (Prokhorov).
- More general smoothing spaces (monotone Sobolev functions, monotone RKHS).
- Super-quantile constraints.

References i

Bachoc, F., F. Gamboa, M. Halford, J-M. Loubes, and L. Risser. 2020. "Explaining Machine Learning Models using Entropic Variable Projection" [in en]. ArXiv: 1810.07924, *arXiv:1810.07924* [*cs, stat*] (December). http://arxiv.org/abs/1810.07924.

de la Fortelle, A. 2015. "A study on generalized inverses and increasing functions Part I: generalized inverses" [in en], 14. https://hal-mines-paristech.archives-ouvertes.fr/hal-01255512.

Dufour, J-M. 1995. *Distribution and quantile functions* [in en]. https://jeanmariedufour.github.io/ResE/Dufour_1995_C_Distribution_Quantile_W.pdf.

Fu, A., B. Narasimhan, and S. Boyd. 2020. "CVXR: An R Package for Disciplined Convex Optimization." *Journal of Statistical Software* 94 (14): 1–34. https://doi.org/10.18637/jss.v094.i14.

Koklu, M., and Y. S. Taspinar. 2021. "Determining the Extinguishing Status of Fuel Flames With Sound Wave by Machine Learning Methods." Conference Name: IEEE Access, *IEEE Access* 9:86207–86216. ISSN: 2169-3536. https://doi.org/10.1109/ACCESS.2021.3088612.

Lemaître, P., E. Sergienko, A. Arnaud, N. Bousquet, F. Gamboa, and B. Iooss. 2015. "Density modification-based reliability sensitivity analysis." *Journal of Statistical Computation and Simulation* 85 (6): 1200–1223. https://doi.org/10.1080/00949655.2013.873039. eprint: https://doi.org/10.1080/00949655.2013.873039. https://doi.org/10.1080/00949655.2013.873039.

Santambrogio, F. 2015. *Optimal Transport for Applied Mathematicians.* Vol. 87. Progress in Nonlinear Differential Equations and Their Applications. Cham: Springer International Publishing. ISBN: 978-3-319-20827-5 978-3-319-20828-2. https://doi.org/10.1007/978-3-319-20828-2. http://link.springer.com/10.1007/978-3-319-20828-2.

References ii

Villani, C. 2003. *Topics in Optimal Transportation* [in en]. Vol. 58. Graduate Studies in Mathematics. ISSN: 1065-7339. American Mathematical Society, March. ISBN: 978-0-8218-3312-4 978-0-8218-7232-1 978-1-4704-1804-5, accessed June 23, 2021. https://doi.org/10.1090/gsm/058. http://www.ams.org/gsm/058.

Thank you for your attention!

Any questions?

Projecting without smoothing

Let P be a probability measure in $\mathcal{P}_2(\mathbb{R})$. Let $\mathcal{C}$ be a non-empty perturbation class, defined by a set of quantile constraints $\mathcal{Q}$. Furthermore, assume, without loss of generality, that, for $i = 1, \ldots, K$,

$$\alpha_1 < \cdots < \alpha_K, \quad \text{along with ,} \quad b_1 < \ldots b_k$$

and let $\beta_i = F_P(b_i)$ for $i = 1, \ldots, K$. Denote the following intervals:

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

Furthermore, let $A_i = [c_i, d_i)$ for $i = 1, \ldots, K$, $A = \bigcup_{i=1}^K A_i$ and $\overline{A} = [0,1] \setminus A$.

The solution of the perturbation problem

$$Q = \underset{G \in \mathcal{P}_2(\mathbb{R})}{\text{argmin}}\ W_2(P, G) \qquad \text{s.t. } G \in \mathcal{C} \tag{3}$$

admits, as a characterizing quantile function :

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

Non-negativity of polynomials on closed intervals

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, $\forall x \in [t_0, t_1]$

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

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, $\forall x \in [t_0, t_1]$

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

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

$$\Gamma = \left[\Gamma_{ij}\right]_{i,j=1,\dots,p}$$

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 $i, j = 1, \dots, p$:

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

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

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

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 $i = 0, \ldots, d$

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

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

$$s^* = \underset{s \in \mathbb{R}^{p+1}}{\operatorname{argmin}}\ s^\top M s - 2 s^\top r$$
$$\text{s.t. } s \in \mathcal{K}$$

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