Mascot-Num Presentation

PhD student : Mathis Deronzier
Supervised by : François Bachoc, Olivier Roustant, and Andrés F. López-Lopera

Institut de Mathématiques de Toulouse, Université Toulouse

Regression under Monotonicity and Fairness Constraints

(1) Block-Additive Gaussian Processes under Monotonicity Constraints
Statistics and Computing - M. D., Andrés F. López-Lopera, François Bachoc, Olivier Roustant, and Jérémy Rohmer

(2) On the Nonconvexity of Push-Forward Constraints and Its Consequences in Machine Learning
SIMODS - Lucas De Lara, M. D., Alberto González-Sanz, and Virgil Foy

(3) Gaussian Regression under Fairness Constraints
M. D., Lucas De Lara, Élie Odin, François Bachoc

(4) Characterization of the dual Cone of Monotone Functions
M. D., Fabrice Gamboa, Olivier Roustant

Gaussian Process Regression under Functional Constraints

Constrained GPs & prediction

Goal. Given observations $\{x_i, y_i\}_{i=1}^n := (\mathbf{x}, \mathbf{y})$, a functional constraint $\mathcal{F}$ (e.g. monotonicity), and a GP $\{W(x) : x \in \mathcal{X}\} \sim \mathcal{GP}(0, k)$, construct:

$$W_{\mathcal{F}} := (W \mid \underbrace{W(\mathbf{x}) + \boldsymbol{\epsilon} = \mathbf{y}}_{\text{observations}}, \quad \underbrace{W \in \mathcal{F}}_{\text{functional constraints}} \quad)$$

Constrained GPs & prediction

Goal. Given observations $\{x_i, y_i\}_{i=1}^n := (\mathbf{x}, \mathbf{y})$, a functional constraint $\mathcal{F}$ (e.g. monotonicity), and a GP $\{W(x) : x \in \mathcal{X}\} \sim \mathcal{GP}(0, k)$, construct:

$$W_{\mathcal{F}} := (W \mid W(\mathbf{x}) + \boldsymbol{\epsilon} = \mathbf{y}, \, W \in \mathcal{F})$$

Key idea. Project W onto some finite dim-space $\mathrm{Span}\,\{\phi_i\}_{i=1}^m$:

$$\widetilde{W} = \Pi_{\boldsymbol{\Phi}} W = \sum_{i=1}^{m} \xi_i \phi_i, \qquad \boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Constrained GPs & prediction

Goal. Given observations $\{x_i, y_i\}_{i=1}^n := (\mathbf{x}, \mathbf{y})$, a functional constraint $\mathcal{F}$ (e.g. monotonicity), and a GP $\{W(x) : x \in \mathcal{X}\} \sim \mathcal{GP}(0, k)$, construct:

$$W_{\mathcal{F}} := (W \mid W(\mathbf{x}) + \boldsymbol{\epsilon} = \mathbf{y}, \, W \in \mathcal{F})$$

Key idea. Project W onto some finite dim-space $\operatorname{Span}\{\phi_i\}_{i=1}^m$:

$$\widetilde{W} = \Pi_{\boldsymbol{\Phi}} W = \sum_{i=1}^{m} \xi_i \phi_i, \qquad \boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

- $\boldsymbol{\xi}$ captures the randomness of $\widetilde{W}$

Constrained GPs & prediction

Goal. Given observations $\{x_i, y_i\}_{i=1}^n := (\mathbf{x}, \mathbf{y})$, a functional constraint $\mathcal{F}$ (e.g. monotonicity), and a GP $\{W(x) : x \in \mathcal{X}\} \sim \mathcal{GP}(0, k)$, construct:

$$W_{\mathcal{F}} := (W \mid W(\mathbf{x}) + \boldsymbol{\epsilon} = \mathbf{y},\, W \in \mathcal{F})$$

Key idea. Project W onto some finite dim-space $\mathrm{Span}\,\{\phi_i\}_{i=1}^m$:

$$\widetilde{W} = \Pi_{\boldsymbol{\Phi}} W = \sum_{i=1}^m \xi_i \phi_i, \qquad \boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

- $\boldsymbol{\xi}$ captures the randomness of $\widetilde{W}$

Prediction. Compute the law of the random vector:

$$\boldsymbol{\xi}_{\mathcal{F}} := (\boldsymbol{\xi} \mid \widetilde{W}(\mathbf{x}) + \boldsymbol{\epsilon} = \mathbf{y},\, \widetilde{W} \in \mathcal{F}),$$

and construct a predictor $\widehat{\boldsymbol{\xi}}_{\mathcal{F}} := \mathrm{mod}\, \boldsymbol{\xi}_{\mathcal{F}}$.

Constrained GPs & prediction

Goal. Given observations $\{x_i, y_i\}_{i=1}^n := (\mathbf{x}, \mathbf{y})$, a functional constraint $\mathcal{F}$ (e.g. monotonicity), and a GP $\{W(x) : x \in \mathcal{X}\} \sim \mathcal{GP}(0, k)$, construct:

$$W_{\mathcal{F}} := (W \mid W(\mathbf{x}) + \boldsymbol{\epsilon} = \mathbf{y},\, W \in \mathcal{F})$$

Key idea. Project W onto some finite dim-space $\mathrm{Span}\,\{\phi_i\}_{i=1}^m$:

$$\widetilde{W} = \Pi_{\boldsymbol{\Phi}} W = \sum_{i=1}^m \xi_i \phi_i, \qquad \boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

- $\boldsymbol{\xi}$ captures the randomness of $\widetilde{W}$

Prediction. Compute the law of the random vector:

$$\boldsymbol{\xi}_{\mathcal{F}} := (\boldsymbol{\xi} \mid \widetilde{W}(\mathbf{x}) + \boldsymbol{\epsilon} = \mathbf{y},\, \widetilde{W} \in \mathcal{F}),$$

and construct a predictor $\widehat{\boldsymbol{\xi}}_{\mathcal{F}} := \mathrm{mod}\, \boldsymbol{\xi}_{\mathcal{F}}$. ⚠ computation in $\mathcal{O}(m^3)$.

Block-Additive Gaussian Processes under Monotonicity Constraints - Statistics and Computing

Joint work with: Andrés F. López-Lopera, François Bachoc, Olivier Roustant, and Jérémy Rohmer

Contributions:

- Extending the framework of constrained GPs to the case of block-additive monotone functions:

$$y(x_1, \cdots, x_D) = y_1(\mathbf{x}_{\mathcal{B}_1}) + \cdots + y_B(\mathbf{x}_{\mathcal{B}_B}).$$

Block-Additive Gaussian Processes under Monotonicity Constraints - Statistics and Computing

Joint work with: Andrés F. López-Lopera, François Bachoc, Olivier Roustant, and Jérémy Rohmer

Contributions:

- Extending the framework of constrained GPs to the case of block-additive monotone functions:

$$y(x_1, \cdots, x_D) = y_1(\mathbf{x}_{\mathcal{B}_1}) + \cdots + y_B(\mathbf{x}_{\mathcal{B}_B}).$$

- A new algorithm *MaxMod*, that, given observations $\{\mathbf{x}_i, y(\mathbf{x}_i)\}_{i=1}^n$,
  - constructs a partition $\mathcal{P} := (\mathcal{B}'_1, \cdots, \mathcal{B}'_{B'})$, together with
  - a monotone block-additive predictor $\widehat{y}$ of y.

Block-Additive Gaussian Processes under Monotonicity Constraints - Statistics and Computing

Joint work with: Andrés F. López-Lopera, François Bachoc, Olivier Roustant, and Jérémy Rohmer

Contributions:

- Extending the framework of constrained GPs to the case of block-additive monotone functions:

$$y(x_1, \cdots, x_D) = y_1(\mathbf{x}_{\mathcal{B}_1}) + \cdots + y_B(\mathbf{x}_{\mathcal{B}_B}).$$

- A new algorithm *MaxMod*, that, given observations $\{\mathbf{x}_i, y(\mathbf{x}_i)\}_{i=1}^n$,
  - constructs a partition $\mathcal{P} := (\mathcal{B}'_1, \cdots, \mathcal{B}'_{B'})$, together with
  - a monotone block-additive predictor $\widehat{y}$ of y.
- Implementation of MaxMod for baGPs on the lineqGPR package: `https://github.com/anfelopera/lineqGPR`.

Introduction to Fairness

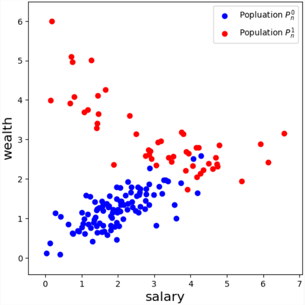
Example

Let $P_n = \{z_1, \ldots, z_n\}$ be a population of n individuals, composed of the blue group P_n^0 and the red group P_n^1 s.t. $P_n = P_n^0 \cup P_n^1$.

Example

Let $P_n = \{z_1, \ldots, z_n\}$ be a population of n individuals, composed of the blue group P_n^0 and the red group P_n^1 s.t. $P_n = P_n^0 \cup P_n^1$.

Each individual $z_i = (\mathbf{x}_i, s_i)$ is described by its **(wealth, salary)**$=$ $\mathbf{x} \in \mathbb{R}^2$, and **color** $s \in \{0, 1\}$

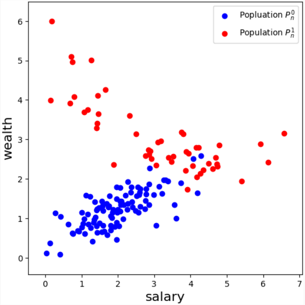


Plot of P_n

Example

Let $P_n = \{z_1, \ldots, z_n\}$ be a population of n individuals, composed of the blue group P_n^0 and the red group P_n^1 s.t. $P_n = P_n^0 \cup P_n^1$.

Each individual $z_i = (\mathbf{x}_i, s_i)$ is described by its **(wealth, salary)**= $\mathbf{x} \in \mathbb{R}^2$, and **color** $s \in \{0, 1\}$ applies for a housing loan.

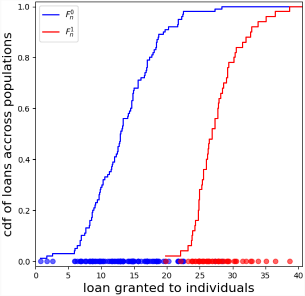


Plot of P_n

each z_i is attributed a loan y_i

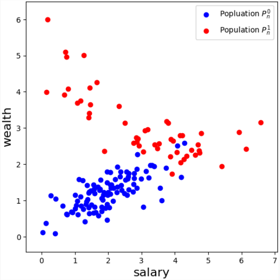
$$y_i = f^\star(z_i) + \xi_i$$

ξ_i is some noise



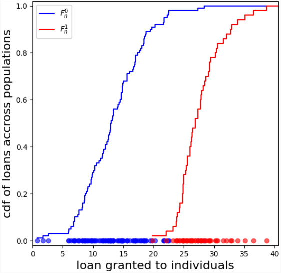
Plot of $\{y_i\}_{i=1}^n$ with CDFs

Ideas with examples for fairness



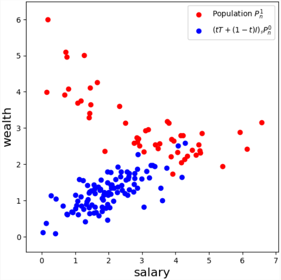
attributing of a loan
from f

$$z_i \mapsto f(z_i)$$



Constructing f : (wealth, salary, color) $\mapsto$ loan

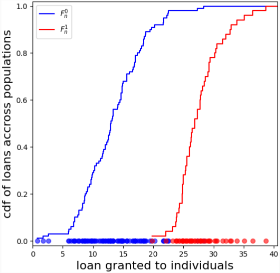
Ideas with examples for fairness



attributing of a loan from f

$$f(\cdot,\cdot,0) = f(T,1)$$

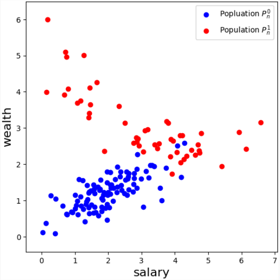
T s.t. $T(P_n^0) \simeq P_n^1$



Constructing f : (wealth, salary, color) $\mapsto$ loan

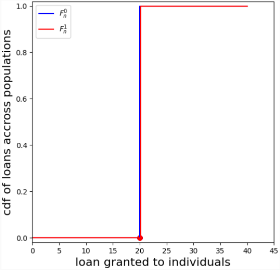
- **Statistical Parity.** The distribution of the loan $F_n^0 \simeq F_n^1$

Ideas with examples for fairness



attributing of a loan
from f

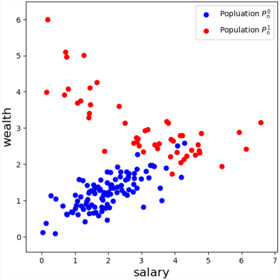
$$f : z_i \mapsto 20$$



Constructing f : (wealth, salary, color) $\mapsto$ loan

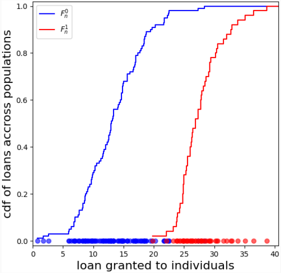
- **Statistical Parity.** The distribution of the loan $F_n^0 \simeq F_n^1$
- **No disparate treatment.** $f(\cdot, \cdot, 0) \simeq f(\cdot, \cdot, 1)$

Ideas with examples for fairness



attributing of a loan from f

$$z_i \mapsto f(z_i)$$



Constructing f : (wealth, salary, color) $\mapsto$ loan

- **Statistical Parity.** The distribution of the loan $F_n^0 \simeq F_n^1$
- **No disparate treatment.** $f(\cdot, \cdot, 0) \simeq f(\cdot, \cdot, 1)$
- **Accuracy.** $f(z_i) \simeq y_i, \forall i \in \{1, \cdots, n\}$

Formalization of Fairness Notions

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $\mathcal{X}$ a Borel space, and $\mathcal{S} = \{0, 1\}$ a binary set (protected attribute), we call $\mathcal{Z} = \mathcal{X} \times \mathcal{S}$.

Let $Z := (X, S) : \Omega \to \mathcal{Z}$ be a random variable of law P and $Y : \Omega \to \mathbb{R}$ a real random variable.

We denote $P^s = \mathcal{L}((X, S) \mid S = s)$, $s \in \mathcal{S} = \{0, 1\}$.

Formalization of Fairness Notions

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $\mathcal{X}$ a Borel space, and $\mathcal{S} = \{0, 1\}$ a binary set (protected attribute), we call $\mathcal{Z} = \mathcal{X} \times \mathcal{S}$.

Let $Z := (X, S) : \Omega \to \mathcal{Z}$ be a random variable of law P and $Y : \Omega \to \mathbb{R}$ a real random variable.

We denote $P^s = \mathcal{L}((X, S) \mid S = s)$, $s \in \mathcal{S} = \{0, 1\}$.

Formalization of fairness

Let $f : \mathcal{Z} \to \mathbb{R}$, we say that f satisfies:

Statistical parity whenever $f(X, S) \perp\!\!\!\perp S \iff f_\sharp P^0 = f_\sharp P^1$

No disparate treatment whenever $f(\cdot, 0) = f(\cdot, 1)$

Formalization of Fairness Notions

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $\mathcal{X}$ a Borel space, and $\mathcal{S} = \{0, 1\}$ a binary set (protected attribute), we call $\mathcal{Z} = \mathcal{X} \times \mathcal{S}$.

Let $Z := (X, S) : \Omega \to \mathcal{Z}$ be a random variable of law P and $Y : \Omega \to \mathbb{R}$ a real random variable.

We denote $P^s = \mathcal{L}((X, S) \mid S = s)$, $s \in \mathcal{S} = \{0, 1\}$.

Formalization of fairness

Let $f : \mathcal{Z} \to \mathbb{R}$, we say that f satisfies:

Statistical parity whenever $f(X, S) \perp\!\!\!\perp S \iff f_\sharp P^0 = f_\sharp P^1$

No disparate treatment whenever $f(\cdot, 0) = f(\cdot, 1)$

Accuracy through quadratic risk.

We are looking at functions minimizing the *quadratic* risk

$$\mathbb{E}\left[|f(X, S) - Y|^2\right], \quad \text{when } f \text{ is fair.}$$

Related Works

Statistical Parity (S.P.): [Chzhen et al., 2020, Le Gouic et al., 2020]

$$\min \mathbb{E}\left[|f(X,S)-Y|^2\right], \quad f \text{ satisfies S.P.}$$

S.P. & No Disparate Treatment (N.D.T.): [Divol and Gaucher, 2024]

$$\min \mathbb{E}\left[|f(X,S)-Y|^2\right], \quad f \text{ satisfies S.P. and N.D.T.}$$

Fair GP for S.P. in Expectation: [Fitzsimons et al., 2019]

$$\text{Linear constraint: } \mathbb{E}[f(X,S=0)] = \mathbb{E}[f(X,S=1)]$$

Kernel Fair Regression: [Pérez-Suay et al., 2017, Li et al., 2019]

$$\min_{f\in\mathcal{H}} \tau\|f\|_{\mathcal{H}}^2 + \mathbb{E}\left[|f(X,S)-Y|^2\right] + \mathrm{HSIC}(f(X,S),S)$$

Existence of a Convex C-loss for the Statistical Parity Constraint?

On the Nonconvexity of Push-Forward Constraints and Its Consequences in Machine Learning - SIMODS

Joint work with: Lucas De Lara, Alberto González-Sanz, and Virgil Foy

Contributions:

C is a subset of $\mathcal{G} := \left\{ f : \mathbb{R}^d \to \mathbb{R}^p \text{ measurable} \right\}$
$J : \mathcal{G} \to [0, +\infty]$ is a C-*loss* if: $J^{-1}(\{0\}) = C$.

On the Nonconvexity of Push-Forward Constraints and Its Consequences in Machine Learning - SIMODS

Joint work with: Lucas De Lara, Alberto González-Sanz, and Virgil Foy

Contributions:

C is a subset of $\mathcal{G} := \left\{ f : \mathbb{R}^d \to \mathbb{R}^p \text{ measurable} \right\}$
$J : \mathcal{G} \to [0, +\infty]$ is a C-*loss* if: $J^{-1}(\{0\}) = C$.

- If $J : \mathcal{G} \to [0, +\infty]$ is a convex function, then $f^{-1}(\{0\})$ is convex. Hence, there exists no convex C-loss J if C is not convex.

On the Nonconvexity of Push-Forward Constraints and Its Consequences in Machine Learning - SIMODS

Joint work with: Lucas De Lara, Alberto González-Sanz, and Virgil Foy

Contributions:

C is a subset of $\mathcal{G} := \left\{ f : \mathbb{R}^d \to \mathbb{R}^p \text{ measurable} \right\}$
$J : \mathcal{G} \to [0, +\infty]$ is a C-*loss* if: $J^{-1}(\{0\}) = C$.

- If $J : \mathcal{G} \to [0, +\infty]$ is a convex function, then $f^{-1}(\{0\})$ is convex. Hence, there exists no convex C-loss J if C is not convex.

- In general, the set of *equalizing maps* betweeen P and $Q \in \mathcal{P}(\mathbb{R}^d)$, $\mathcal{E}(P, Q) := \left\{ f : \mathbb{R}^d \to \mathbb{R}^p : f_\sharp P = f_\sharp Q \right\}$ is not convex.

On the Nonconvexity of Push-Forward Constraints and Its Consequences in Machine Learning - SIMODS

Joint work with: Lucas De Lara, Alberto González-Sanz, and Virgil Foy

Contributions:

C is a subset of $\mathcal{G} := \left\{ f : \mathbb{R}^d \to \mathbb{R}^p \text{ measurable} \right\}$
$J : \mathcal{G} \to [0, +\infty]$ is a C-*loss* if: $J^{-1}(\{0\}) = C$.

- If $J : \mathcal{G} \to [0, +\infty]$ is a convex function, then $f^{-1}(\{0\})$ is convex. Hence, there exists no convex C-loss J if C is not convex.

- In general, the set of *equalizing maps* betweeen P and $Q \in \mathcal{P}(\mathbb{R}^d)$, $\mathcal{E}(P, Q) := \left\{ f : \mathbb{R}^d \to \mathbb{R}^p : f_\sharp P = f_\sharp Q \right\}$ is not convex.

Conclusion: In general, there exists no convex C-loss for the sets of *equalizing maps*.

Gaussian Process Regression under Fairness Constraints

Joint work with: François Bachoc, Lucas De Lara, and Élie Odin

Fair Gaussian Process

Empirical information. Let $(\mathbf{x}, \mathbf{s}, \mathbf{y}) = \{x_i, s_i, y_i\}_{i=1}^n$ be n i.i.d. realizations of (X, S, Y). We write $Z = (X, S)$ and $z_i = (x_i, s_i)$. We split $\mathbf{z} = \{(x_i, s_i)\}_{i=1}^n$ in two: $\mathbf{z} = \mathbf{z}^0 \cup \mathbf{z}^1$, $\mathbf{z}^s = \{(x_i, s_i)\}_{i: s_i = s}$, $s \in \{0, 1\}$. Call $\widehat{\mathbb{P}}_{\mathbf{z}^0}, \widehat{\mathbb{P}}_{\mathbf{z}^1}$ the empirical measures associated to $\mathbf{z}^0$ and $\mathbf{z}^1$.

Fair Gaussian Process

Empirical information. Let $(\mathbf{x}, \mathbf{s}, \mathbf{y}) = \{x_i, s_i, y_i\}_{i=1}^n$ be n i.i.d. realizations of (X, S, Y). We write $Z = (X, S)$ and $z_i = (x_i, s_i)$. We split $\mathbf{z} = \{(x_i, s_i)\}_{i=1}^n$ in two: $\mathbf{z} = \mathbf{z}^0 \cup \mathbf{z}^1$, $\mathbf{z}^s = \{(x_i, s_i)\}_{i:s_i=s}$, $s \in \{0, 1\}$. Call $\widehat{\mathbb{P}}_{\mathbf{z}^0}, \widehat{\mathbb{P}}_{\mathbf{z}^1}$ the empirical measures associated to $\mathbf{z}^0$ and $\mathbf{z}^1$.

δ-Fair Gaussian Process.

Let $\{W(z) : z \in \mathcal{Z}\} \sim \mathcal{GP}(0, k)$ and $\mathcal{D}$ a distance on $\mathcal{P}(\mathbb{R})$.
We define the δ-fair process:

$$\widehat{W}_\delta := \left(W \mid \underbrace{W(\mathbf{z}) = \mathbf{y} + \boldsymbol{\epsilon}}_{\text{Accuracy}}, \underbrace{\mathcal{D}\left(W_\sharp \widehat{\mathbb{P}}_{\mathbf{z}^0}, W_\sharp \widehat{\mathbb{P}}_{\mathbf{z}^1}\right) \leq \delta}_{\text{Statistical Parity}} \right),$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \tau \mathbf{I}_n)$ is a Gaussian noise independent of W.

Fair Gaussian Process

Empirical information. Let $(\mathbf{x}, \mathbf{s}, \mathbf{y}) = \{x_i, s_i, y_i\}_{i=1}^n$ be n i.i.d. realizations of (X, S, Y). We write $Z = (X, S)$ and $z_i = (x_i, s_i)$. We split $\mathbf{z} = \{(x_i, s_i)\}_{i=1}^n$ in two: $\mathbf{z} = \mathbf{z}^0 \cup \mathbf{z}^1$, $\mathbf{z}^s = \{(x_i, s_i)\}_{i:s_i=s}$, $s \in \{0, 1\}$. Call $\widehat{\mathbb{P}}_{\mathbf{z}^0}, \widehat{\mathbb{P}}_{\mathbf{z}^1}$ the empirical measures associated to $\mathbf{z}^0$ and $\mathbf{z}^1$.

δ-Fair Gaussian Process.

Let $\{W(z) : z \in \mathcal{Z}\} \sim \mathcal{GP}(0, k)$ and $\mathcal{D}$ a distance on $\mathcal{P}(\mathbb{R})$. We define the δ-fair process:

$$\widehat{W}_\delta := \left(W \mid \underbrace{W(\mathbf{z}) = \mathbf{y} + \boldsymbol{\epsilon}}_{\text{Accuracy}}, \underbrace{\mathcal{D}\left(W_\sharp \widehat{\mathbb{P}}_{\mathbf{z}^0}, W_\sharp \widehat{\mathbb{P}}_{\mathbf{z}^1}\right) \leq \delta}_{\text{Statistical Parity}} \right),$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \tau \mathbf{I}_n)$ is a Gaussian noise independent of W.

Remark. $\mathcal{D}\left(W_\sharp \widehat{\mathbb{P}}_{\mathbf{z}^0}, W_\sharp \widehat{\mathbb{P}}_{\mathbf{z}^1}\right)$ is a function of $W(\mathbf{z})$.

Construction of Fair Predictions

Definition. Maximum a Posteriori (MAP) prediction

The MAP prediction, $\mathscr{Z}^p \to \mathbb{R}^p$ is defined for every element $\mathbf{t} \in \mathscr{Z}^p$ as

$$\mathrm{pred}(\mathbf{t}) := \pi_{[p]}(\mathrm{mod}\ \widehat{W}_\delta(\mathbf{t}, \mathbf{z})).$$

Construction of Fair Predictions

Definition. Maximum a Posteriori (MAP) prediction

The MAP prediction, $\mathcal{Z}^p \to \mathbb{R}^p$ is defined for every element $\mathbf{t} \in \mathcal{Z}^p$ as

$$\mathrm{pred}(\mathbf{t}) := \pi_{[p]}(\mathrm{mod}\ \widehat{W}_\delta(\mathbf{t}, \mathbf{z})).$$

Theorem. Consistency of the MAP

There exists $\widehat{f}_\delta : \mathcal{Z} \to \mathbb{R}$ s.t. $\forall \mathbf{t} \in \mathcal{Z}^p$:

$$\mathrm{pred}(\mathbf{t}) = \widehat{f}_\delta(\mathbf{t}) = k(\mathbf{t}, \mathbf{z}) k(\mathbf{z}, \mathbf{z})^{-1} \widehat{\mathbf{w}}_\delta, \tag{1}$$

where k is the kernel of W

Construction of Fair Predictions

Definition. Maximum a Posteriori (MAP) prediction

The MAP prediction, $\mathcal{Z}^p \to \mathbb{R}^p$ is defined for every element $\mathbf{t} \in \mathcal{Z}^p$ as

$$\mathrm{pred}(\mathbf{t}) := \pi_{[p]}(\mathrm{mod}\ \widehat{W}_\delta(\mathbf{t}, \mathbf{z})).$$

Theorem. Consistency of the MAP

There exists $\widehat{f}_\delta : \mathcal{Z} \to \mathbb{R}$ s.t. $\forall \mathbf{t} \in \mathcal{Z}^p$:

$$\mathrm{pred}(\mathbf{t}) = \widehat{f}_\delta(\mathbf{t}) = k(\mathbf{t}, \mathbf{z}) k(\mathbf{z}, \mathbf{z})^{-1} \widehat{\mathbf{w}}_\delta, \tag{1}$$

where k is the kernel of W and $\widehat{\mathbf{w}}_\delta$ is a solution of

$$\begin{aligned} \min_{\mathbf{w} = (\mathbf{w}^0, \mathbf{w}^1) \in \mathbb{R}^n} \quad & \|\mathbf{w} - \boldsymbol{\mu}\|^2_{\boldsymbol{\Gamma}^{-1}} \\ \text{subject to} \quad & \mathcal{D}(\widehat{\mathbb{P}}_{\mathbf{w}^0}, \widehat{\mathbb{P}}_{\mathbf{w}^1}) \le \delta, \end{aligned} \tag{Pb$_\delta$}$$

$(W(\mathbf{z}) \mid W(\mathbf{z}) = \mathbf{y} + \boldsymbol{\epsilon}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Gamma})$ and $\widehat{\mathbb{P}}_{\mathbf{w}^s}$ is the empirical probability measure associated to $\mathbf{w}^s$ for $s \in \{0, 1\}$.

Computation of the MAP predictor

The MAP problem. We aim to solve:

$$\begin{aligned} \min_{\mathbf{w}=(\mathbf{w}^0,\mathbf{w}^1)\in\mathbb{R}^n} \quad & \|\mathbf{w}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Gamma}^{-1}} \\ \text{subject to} \quad & \mathcal{D}(\widehat{\mathbb{P}}_{\mathbf{w}^0}, \widehat{\mathbb{P}}_{\mathbf{w}^1}) \leq \delta, \end{aligned} \tag{Pb$_\delta$}$$

Computation of the MAP predictor

The MAP problem. We aim to solve:

$$\begin{aligned} \min_{\mathbf{w}=(\mathbf{w}^0,\mathbf{w}^1)\in\mathbb{R}^n} \quad & \|\mathbf{w}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Gamma}^{-1}} \\ \text{subject to} \quad & \mathcal{D}(\widehat{\mathbb{P}}_{\mathbf{w}^0}, \widehat{\mathbb{P}}_{\mathbf{w}^1}) \leq \delta, \end{aligned} \tag{Pb$_\delta$}$$

Penalized problem. In practice, we solve the penalized problem:

$$\min_{\mathbf{w}=(\mathbf{w}^0,\mathbf{w}^1)\in\mathbb{R}^n} \quad ||\mathbf{w}-\boldsymbol{\mu}||^2_{\boldsymbol{\Gamma}^{-1}} + \lambda\mathcal{D}(\widehat{\mathbb{P}}_{\mathbf{w}^0}, \widehat{\mathbb{P}}_{\mathbf{w}^1}). \tag{Pb$_\lambda$}$$

Computation of the MAP predictor

The MAP problem. We aim to solve:

$$\begin{aligned} \min_{\mathbf{w}=(\mathbf{w}^0,\mathbf{w}^1)\in\mathbb{R}^n} \quad & \|\mathbf{w}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Gamma}^{-1}} \\ \text{subject to} \quad & \mathcal{D}(\widehat{\mathbb{P}}_{\mathbf{w}^0}, \widehat{\mathbb{P}}_{\mathbf{w}^1}) \leq \delta, \end{aligned} \tag{Pb$_\delta$}$$

Penalized problem. In practice, we solve the penalized problem:

$$\min_{\mathbf{w}=(\mathbf{w}^0,\mathbf{w}^1)\in\mathbb{R}^n} \quad ||\mathbf{w}-\boldsymbol{\mu}||^2_{\boldsymbol{\Gamma}^{-1}} + \lambda\mathcal{D}(\widehat{\mathbb{P}}_{\mathbf{w}^0}, \widehat{\mathbb{P}}_{\mathbf{w}^1}). \tag{Pb$_\lambda$}$$

Correspondence between (Pb$_\lambda$) and (Pb$_\delta$)

For all $\lambda > 0$:

- $\widehat{\mathbf{w}}_\lambda = (\widehat{\mathbf{w}}^0_\lambda, \widehat{\mathbf{w}}^1_\lambda)$ sol (Pb$_\lambda$) $\implies$ $\widehat{\mathbf{w}}_\lambda$ sol (Pb$_\delta$) if $\boxed{\delta = \mathcal{D}(\widehat{\mathbb{P}}_{\widehat{\mathbf{w}}^0_\lambda}, \widehat{\mathbb{P}}_{\widehat{\mathbf{w}}^1_\lambda})}$

Computation of the MAP predictor

The MAP problem. We aim to solve:

$$\begin{aligned} \min_{\mathbf{w}=(\mathbf{w}^0,\mathbf{w}^1)\in\mathbb{R}^n} \quad & \|\mathbf{w}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Gamma}^{-1}} \\ \text{subject to} \quad & \mathcal{D}(\widehat{\mathbb{P}}_{\mathbf{w}^0}, \widehat{\mathbb{P}}_{\mathbf{w}^1}) \leq \delta, \end{aligned} \tag{Pb$_\delta$}$$

Penalized problem. In practice, we solve the penalized problem:

$$\min_{\mathbf{w}=(\mathbf{w}^0,\mathbf{w}^1)\in\mathbb{R}^n} \quad ||\mathbf{w}-\boldsymbol{\mu}||^2_{\boldsymbol{\Gamma}^{-1}} + \lambda\mathcal{D}(\widehat{\mathbb{P}}_{\mathbf{w}^0}, \widehat{\mathbb{P}}_{\mathbf{w}^1}). \tag{Pb$_\lambda$}$$

Correspondence between (Pb$_\lambda$) and (Pb$_\delta$)

For all $\lambda > 0$:

- $\widehat{\mathbf{w}}_\lambda = (\widehat{\mathbf{w}}^0_\lambda, \widehat{\mathbf{w}}^1_\lambda)$ sol (Pb$_\lambda$) $\implies$ $\widehat{\mathbf{w}}_\lambda$ sol (Pb$_\delta$) if $\boxed{\delta = \mathcal{D}(\widehat{\mathbb{P}}_{\widehat{\mathbf{w}}^0_\lambda}, \widehat{\mathbb{P}}_{\widehat{\mathbf{w}}^1_\lambda})}$
- $\mathcal{D}(\widehat{\mathbb{P}}_{\widehat{\mathbf{w}}^0_\lambda}, \widehat{\mathbb{P}}_{\widehat{\mathbf{w}}^1_\lambda}) \leq ||\boldsymbol{\mu}||^2_{\boldsymbol{\Gamma}^{-1}}/\lambda$

Hence, the empirical statistical parity is controlled by λ.

A Kimeldorf-Wahba Correspondence

Let $\widehat{\mathbf{w}}_\lambda$ be a solution of (Pb_λ), define

$$\widehat{f}_\lambda(\mathbf{t}) = k(\mathbf{t}, \mathbf{z}) k(\mathbf{z}, \mathbf{z})^{-1} \widehat{\mathbf{w}}_\lambda.$$

A Kimeldorf-Wahba Correspondence

Let $\widehat{\mathbf{w}}_\lambda$ be a solution of (Pb$_\lambda$), define

$$\widehat{f}_\lambda(\mathbf{t}) = k(\mathbf{t}, \mathbf{z})k(\mathbf{z}, \mathbf{z})^{-1}\widehat{\mathbf{w}}_\lambda.$$

Theorem

The function $\widehat{f}_\lambda$, is solution of the Tikhonov-regularized minimization problem penalized with fairness constraint:

$$\min_{f \in \mathcal{H}} \underbrace{\tau \|f\|^2_{\mathcal{H}}}_{\text{noise } \epsilon} + \underbrace{\sum_{i=1}^{n} |f(z_i) - y_i|^2}_{\text{empirical } L^2 \text{ risk}} + \underbrace{\lambda \mathcal{D}(f_\sharp \widehat{\mathbb{P}}_{\mathbf{z}^0}, f_\sharp \widehat{\mathbb{P}}_{\mathbf{z}^1})}_{\text{fairness penalty}}. \qquad (\text{Pb}_{\mathcal{H},\lambda})$$

where $\mathcal{H}$ is the RKHS generated by the kernel k of the GP W.

A Kimeldorf-Wahba Correspondence

Let $\widehat{\mathbf{w}}_\lambda$ be a solution of (Pb$_\lambda$), define

$$\widehat{f}_\lambda(\mathbf{t}) = k(\mathbf{t}, \mathbf{z}) k(\mathbf{z}, \mathbf{z})^{-1} \widehat{\mathbf{w}}_\lambda.$$

Theorem

The function $\widehat{f}_\lambda$, is solution of the Tikhonov-regularized minimization problem penalized with fairness constraint:

$$\min_{f \in \mathcal{H}} \underbrace{\tau \|f\|^2_{\mathcal{H}}}_{\text{noise } \epsilon} + \underbrace{\sum_{i=1}^{n} |f(z_i) - y_i|^2}_{\text{empirical } L^2 \text{ risk}} + \underbrace{\lambda \mathcal{D}(f_\sharp \widehat{\mathbb{P}}_{\mathbf{z}^0}, f_\sharp \widehat{\mathbb{P}}_{\mathbf{z}^1})}_{\text{fairness penalty}}. \qquad (\text{Pb}_{\mathcal{H},\lambda})$$

where $\mathcal{H}$ is the RKHS generated by the kernel k of the GP W.

Remark. This is the fair kernel regression problem in [Pérez-Suay et al., 2017].

Asymptotic properties of the regressions functions

Let $(z_i, y_i)_{i\geq 1}$ be a sequence of i.i.d. realizations of $(Z, Y) : \Omega \to \mathcal{Z} \times \mathbb{R}$.
For all n, let $\mathbf{z}_n = \{z_i\}_{i=1}^n = \mathbf{z}_n^0 \cup \mathbf{z}_n^1 \in \mathcal{Z}^n$, $\tau_n,\ \lambda_n > 0$

Let $\mathcal{G}_n$ be the set of solutions of:

$$\min_{f\in\mathcal{H}}\ \tau_n\|f\|_{\mathcal{H}}^2 + \frac{1}{n}\sum_{i=1}^{n}|f(z_i) - y_i|^2 + \lambda_n\mathcal{D}\left(f_\sharp\widehat{\mathbb{P}}_{\mathbf{z}_n^0}, f_\sharp\widehat{\mathbb{P}}_{\mathbf{z}_n^1}\right) \qquad (\mathrm{Pb}_{\mathcal{H},n})$$

Asymptotic properties of the regressions functions

Let $(z_i, y_i)_{i\geq 1}$ be a sequence of i.i.d. realizations of $(Z, Y) : \Omega \to \mathcal{Z} \times \mathbb{R}$.
For all n, let $\mathbf{z}_n = \{z_i\}_{i=1}^n = \mathbf{z}_n^0 \cup \mathbf{z}_n^1 \in \mathcal{Z}^n$, τ_n, $\lambda_n > 0$

Let $\mathcal{G}_n$ be the set of solutions of:

$$\min_{f\in\mathcal{H}} \tau_n \|f\|_{\mathcal{H}}^2 + \frac{1}{n}\sum_{i=1}^n |f(z_i) - y_i|^2 + \lambda_n \mathcal{D}\left(f_\sharp \widehat{\mathbb{P}}_{\mathbf{z}_n^0}, f_\sharp \widehat{\mathbb{P}}_{\mathbf{z}_n^1}\right) \qquad (\mathrm{Pb}_{\mathcal{H},n})$$

Theorem. Convergence of the solution set

$\mathcal{G}_n$ converges a.s. to $\mathcal{G}^\star$ in Hausdorff distance.

where $\mathcal{G}^\star$ is the set of solutions of the problem

$$\min_{f\in\mathcal{H}} \tau \|f\|_{\mathcal{H}}^2 + \mathbb{E}\left[|f(Z) - Y|^2\right] + \lambda \mathcal{D}\left(f_\sharp P^0, f_\sharp P^1\right). \qquad (\mathrm{Pb}_{\mathcal{H},\star})$$

If $\tau_n \to \tau > 0$ and $\lambda_n \to \lambda$. Recall $P^s = \mathcal{L}(Z \mid S = s)$.

Parametrization for controlling correlation of GPs

Parametrization choices

Assume $\{W(x,s) : (x,s) \in \mathcal{X} \times \mathcal{S}\} \sim \mathcal{GP}(0, k^\rho)$ where

$$k^\rho((x,s),(x',s')) = k_\mathcal{X}(x,x') \cdot k_\mathcal{S}^\rho(s,s').$$

Where $k_\mathcal{S}^\rho$ is defined as:

$$\begin{bmatrix} k_\mathcal{S}^\rho(0,0) & k_\mathcal{S}^\rho(0,1) \\ k_\mathcal{S}^\rho(1,0) & k_\mathcal{S}^\rho(1,1) \end{bmatrix} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

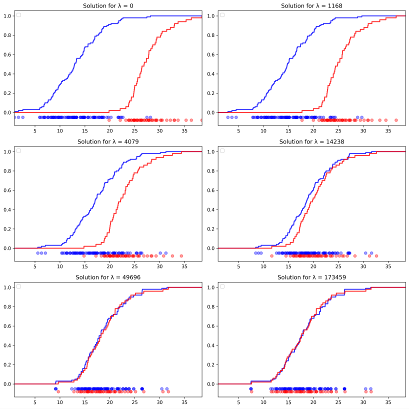
then, $\rho = \mathbb{C}\text{or}(W(x,0), W(x,1)) \quad \forall x \in \mathcal{X}$.

Consequences of the choice of ρ.

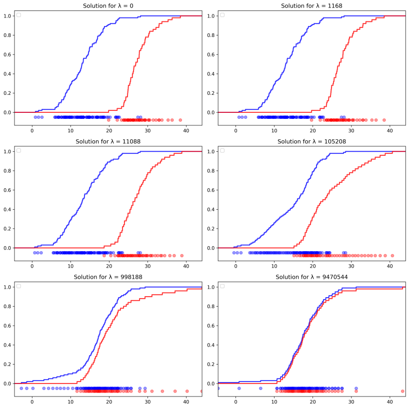
- When $\rho = 0, \quad W(x,0) \perp\!\!\!\perp W(y,1), \quad \forall x,y \in \mathcal{X}$.
- When $\rho = 1, \quad W(x,0) = W(x,1), \quad \forall x \in \mathcal{X}$ a.s.
- Control on the N.D.T. : $\left\| \widehat{f}_\rho(\cdot,0) - \widehat{f}_\rho(\cdot,1) \right\|_{\mathcal{H}_\mathcal{X}} \leq |1-\rho| A$.

Numerical Results

Numerical results on the example for different value of λ



Example of fair predictor with Sinkhorn divergence



Example of fair predictors with MMD distance

Numerical results of CRIME dataset

Methodology. Split the dataset (70% training, 30% testing). Computation of Mean Square Error (MSE) and Kolmogorov-Smirnov (KS): $\left\|\mathcal{F}^0 - \mathcal{F}^1\right\|_\infty$ on the test set and comparison of the results with the state of the art [Chzhen et al., 2020].

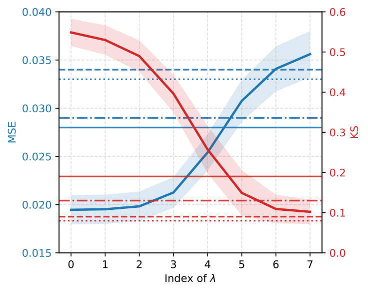


Figure 1:
$\rho = 0 \iff W(\cdot, 0) \perp\!\!\!\perp W(\cdot, 1)$

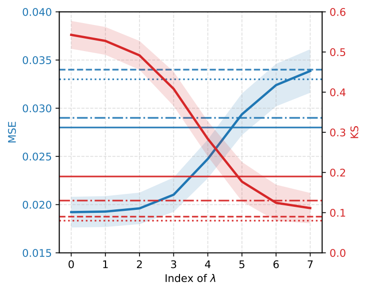


Figure 2:
$\rho = 1 \iff W(\cdot, 0) = W(\cdot, 1)$

Summary of Results & Conclusion

Summary of Results

MAP/RKHS correspondence

1. We developed a method to construct a “fair” predictor with a MAP method from a δ-fair GP.
2. The *fair GP regression* corresponds to *kernel fair regression*.
3. Asymptotically, the MAP predictors converge to the solutions of a regularized constrained problem.

Summary of Results

MAP/RKHS correspondence

1. We developed a method to construct a “fair” predictor with a MAP method from a δ-fair GP.
2. The *fair GP regression* corresponds to *kernel fair regression*.
3. Asymptotically, the MAP predictors converge to the solutions of a regularized constrained problem.

Control on the fairness constraint

4 We can control the **empirical statistical parity** constraint with parameter λ in (Pb_λ).

5 We can control the **no disparate treatment** through kernel parametrization $k_{\mathcal{S}}^{\rho}$.

Summary of Results

MAP/RKHS correspondence

1. We developed a method to construct a "fair" predictor with a MAP method from a δ-fair GP.
2. The *fair GP regression* corresponds to *kernel fair regression*.
3. Asymptotically, the MAP predictors converge to the solutions of a regularized constrained problem.

Control on the fairness constraint

4 We can control the **empirical statistical parity** constraint with parameter λ in (Pb_λ).

5 We can control the **no disparate treatment** through kernel parametrization $k_{\mathcal{S}}^{\rho}$.

Uncertainty Quantification

On our predictions? On the fairness?

Bibliography

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. (2020).
Fair regression with Wasserstein barycenters.
Advances in Neural Information Processing Systems, 33:7321–7331.

Divol, V. and Gaucher, S. (2024).
Demographic parity in regression and classification within the unawareness framework.
arXiv preprint arXiv:2409.02471.

Fitzsimons, J., Ali, A. A., Osborne, M., and Roberts, S. (2019).
A General Framework for Fair Regression.
Entropy, 21(8):741.

Le Gouic, T., Loubes, J.-M., and Rigollet, P. (2020).
Projection to Fairness in Statistical Learning.

Li, Z., Perez-Suay, A., Camps-Valls, G., and Sejdinovic, D. (2019).
Kernel dependence regularizers and Gaussian processes with applications to algorithmic fairness.
arXiv preprint arXiv:1911.04322.

Thank you! Questions?