An Approachability Perspective on Fair Online Learning

joint work with Christophe Giraud, Gilles Stoltz

November 5, 2025

Evgenii Chzhen

EU regulation for AI

PROHIBITED ARTIFICIAL INTELLIGENCE PRACTICES

Article 5

1. The following artificial intelligence practices shall be prohibited:

   (a) the placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm;

   (b) the placing on the market, putting into service or use of an AI system that exploits any of the vulnerabilities of a specific group of persons due to their age, physical or mental disability, in order to materially distort the behaviour of a person pertaining to that group in a manner that causes or is likely to cause that person or another person physical or psychological harm;

   (c) the placing on the market, putting into service or use of AI systems by public authorities or on their behalf for the evaluation or classification of the trustworthiness of natural persons over a certain period of time based on their social behaviour or known or predicted personal or personality characteristics, with the social score leading to either or both of the following:

      (i) detrimental or unfavourable treatment of certain natural persons or whole groups thereof in social contexts which are unrelated to the contexts in which the data was originally generated or collected;

      (ii) detrimental or unfavourable treatment of certain natural persons or whole groups thereof that is unjustified or disproportionate to their social behaviour or its gravity;

The talk

- A biased introduction to fairness in ML
- An approachability perspective on (adversarial) fair online learning
- Application: trade-off between group-wise calibration and demographic parity

1- A (biased) tour in the Fair-ML zoology

Different points of view

We can identify (at least) 3 main approaches for improving fairness in prediction

1. Individual fairness: aims to treat similar people similarly (individual notions)

2. Causal fairness: tries to identify causes of unfairness in order to act on them (causal notions)

3. Group fairness: seeks to comply to fairness criteria at the sub-population level (statistical notions)
   3.1 Stochastically defined subgroups;
   3.2 Deterministically defined subgroups, but with overlaps (a.k.a multi-group fairness)

Learning framework

Notation

- Outcome $Y \in \mathcal{Y}$
- Covariate/features $X \in \mathcal{X}$
- Sensitive attribute $S \in \mathcal{S}$
- Predictor: $\underbrace{f : \mathcal{X} \times \mathcal{S} \to \mathcal{Y}}_{\text{Awareness}}$ (possibly $\underbrace{f : \mathcal{X} \to \mathcal{Y}}_{\text{Unawareness}}$)
- Prediction: $F = f(X, S)$ (possibly $F = f(X)$)
- Some distribution on $\mathbb{P}$ on $(\mathcal{X}, \mathcal{S}, \mathcal{Y})$

Ex: binary classification with binary sensitive attribute

- Outcome: label $Y \in \{0, 1\}$
- Sensitive attribute: $S \in \{0, 1\}$

Statistical fairness: Demographic parity

Demographic parity

$$F \perp\!\!\!\perp S$$

(Kamiran and Calders, 2012)

Ex: (binary classification)

$$\mathbb{P}\left[F = 1 | S = 1\right] \cong \mathbb{P}\left[F = 1 | S = 0\right]$$

Demographic parity promotes *diversity* and can be related to affirmative action policies.

Statistical fairness: Equalized Odds

Equalized Odds

$$F \perp\!\!\!\perp S \mid Y$$

(Hardt, Price, and Srebro, 2016)

Ex: (binary classification)

$$\mathbb{P}\left[F = 1 | S = 1, Y\right] \cong \mathbb{P}\left[F = 1 | S = 0, Y\right]$$

Equalized Odds encodes a notion of *Meritocratic fairness*.

Performance fairness: Group-wise calibration

Group-wise calibration

$$\mathbb{E}\left[Y|S,F\right] \cong F$$

(Barocas, Hardt, and Narayanan, 2023)

Ex: (binary classification) for a score $F \in [0,1]$

$$\mathbb{P}\left[Y=1|S=1,F\right] \cong \mathbb{P}\left[Y=1|S=0,F\right] \cong F$$

The prediction are *calibrated for each group*.

Performance fairness: Equal group-wise risk

Equal group-wise risk

For a loss function ℓ

$$\mathbb{E}\left[\ell(Y,F)|S\right] \cong \mathbb{E}\left[\ell(Y,F)\right]$$

Statistical fairness: many different criteria

A large zoology

Demographic parity	$F \perp\!\!\!\perp S$
Equalized odds	$F \perp\!\!\!\perp S \vert Y$
Equal opportunity	$F \perp\!\!\!\perp S \vert Y \in \mathcal{Y}_+$
Predictive parity	$Y \in \mathcal{Y}_+ \perp\!\!\!\perp S \mid F \in \mathcal{Y}_+$
Group-wise calibration	$\mathbb{E}\left[Y \vert S, F\right] \cong F$
Equal group-wise risk	$\mathbb{E}\left[\ell(Y, F) \vert S\right] \cong \mathbb{E}\left[\ell(Y, F)\right]$

with some incompatible notions!

The famous COMPAS case

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a software which aims to predict recidivism risk.

ProPublica compared COMPAS predictions across ethinicity groups in the USA. It exhibits a large violation of the Equalized Odds criteria.

The COMPAS developers argue yet that COMPAS (almost) complies with Predictive parity.

Chouldechova (2017) and Kleinberg *et al.* (2017) show that it is impossible to comply simultaneously with Equalized Odds and Predictive parity, unless $Y \perp\!\!\!\perp S$.

Finding a balance between different notions

Relaxing fairness criteria

- Fairness criteria are imperfect mathematical transposition of qualitative ideas;
- Evaluations of fairness criteria are subjected to uncertainties;
- Some fairness criteria are incompatible;
- We can seek for a good trade-off between different fairness criteria and prediction performance.

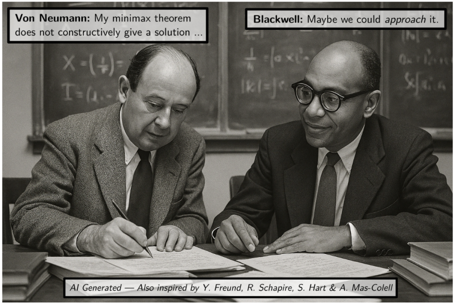
Instead of asking for an exact compliance to fairness criteria, maybe

- introduce (quantitive) measures of violation of the fairness criteria
- and seek for limited violation of fairness criteria?

2- An Approachability Perspective on Fair Online Learning

Our goals

- To investigate fairness in adversarial online learning
- To adopt a unified perspective
- To get benchmark algorithms
- To retrieve information on possible trade-offs between different objectives



Fair online learning via approachability

Informal description: for $t \geq 1$

- A request arrives with attributes (x_t, s_t)
- The Learner observes x_t and tries to predict the (adversarial) outcome y_t
- The goal of the Learner is to provide a prediction a_t which is both fair and accurate.

Encoding the objectives of the learner

We encode the objectives (no-regret, demographic parity, etc) via

- a payoff function $\mathbf{m}(a_t, y_t, x_t, s_t)$
- and a target set $\mathcal{C}$.

Goal: $$\frac{1}{T}\sum_{t=1}^{T}\mathbf{m}(a_t, y_t, x_t, s_t) \longrightarrow \mathcal{C}$$

Encoding learning and fairness constraints

The payoff function $\mathbf{m}(a, y, x, s)$ and the target set $\mathcal{C}$ encode the objectives of the learner (no-regret, Demographic parity, etc).

Example: Demographic Parity (DP)

⚠ As we are in an adversarial online setting, we replace distributional properties by empirical counterparts.

Aim: to have, for T large,

$$\left| \frac{1}{\gamma_0 T} \sum_{t=1}^{T} a_t \mathbf{1}_{s_t=0} - \frac{1}{\gamma_1 T} \sum_{t=1}^{T} a_t \mathbf{1}_{s_t=1} \right| \leq \delta,$$

where $\gamma_s = \mathbf{Q}(s_t = s)$.

DP payoff function: $\mathbf{m}_{\mathrm{DP}}(a, s) = \left(\frac{a}{\gamma_0} \mathbf{1}_{s=0}, \ \frac{a}{\gamma_1} \mathbf{1}_{s=1} \right)$

DP target set: $\mathcal{C}_{\mathrm{DP}}(\delta) = \left\{ (u, v) \in \mathbb{R}^2 : \ |u - v| \leq \delta \right\}$

Encoding learning and fairness constraints

The payoff function $\mathbf{m}(a, y, x, s)$ and the target set $\mathcal{C}$ encode the objectives of the learner (no-regret, Demographic parity, etc).

Example: Group Calibration (GrCal)

Aim: to have, for T large,

$$\sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \left| \frac{1}{\gamma_s T} \sum_{t=1}^{T} (a - y_t) \mathbf{1}_{a_t = a} \mathbf{1}_{s_t = s} \right| \leq \varepsilon,$$

where $\gamma_s = \mathbf{Q}(s_t = s)$.

GrCal payoff function: $\mathbf{m}_{\text{gr-cal}}(a, y, s) = \left(\frac{a' - y}{\gamma_{s'}} \mathbf{1}_{a = a'} \mathbf{1}_{s = s'} \right)_{\substack{a' \in \mathcal{A} \\ s' \in \mathcal{S}}}$

GrCal target set: $\mathcal{C}_{\text{gr-cal}}(\varepsilon) = \left\{ \mathbf{v} \in \mathbb{R}^{N|\mathcal{S}|} : \|\mathbf{v}\|_1 \leq \varepsilon \right\}$

similar to (Hart and Mas-Colell, 2000; Mannor and Stoltz, 2010)

Encoding learning and fairness constraints

The payoff function $\mathbf{m}(a, y, x, s)$ and the target set $\mathcal{C}$ encode the objectives of the learner (no-regret, Demographic parity, etc).

Criterion	Vector payoff function $\mathbf{m}$	Closed convex target set $\mathcal{C}$
Calibration	$\mathbf{m}_{\text{cal}}(a,y) = \big((a'-y)\,\mathbf{1}_{a=a'}\big)_{a'\in\mathcal{A}}$	$\mathcal{C}_{\text{cal}} = \big\{\mathbf{v}\in\mathbb{R}^N : \ \Vert\mathbf{v}\Vert_1 \le \varepsilon\big\}$
Group-calibration	$\mathbf{m}_{\text{gr-cal}}(a,y,s) = \big(\mathbf{m}_{\text{cal}}(a,y)\,\mathbf{1}_{s=s'}/\gamma_{s'}\big)_{s'\in\mathcal{S}}$	$\mathcal{C}_{\text{gr-cal}} = \big\{\mathbf{v}\in\mathbb{R}^{N\vert\mathcal{S}\vert} : \ \Vert\mathbf{v}\Vert_1 \le \varepsilon\big\}$
No-regret	$\mathbf{m}_{\text{reg}}(a,y,x,s) = \big(r(a,y,x,s) - r(a',y,x,s)\big)_{a'\in\mathcal{A}}$	$\mathcal{C}_{\text{reg}} = [0,+\infty)^N$
Group-no-regret	$\mathbf{m}_{\text{gr-reg}}(a,y,x,s) = \big(\mathbf{m}_{\text{reg}}(a,y,x,s)\,\mathbf{1}_{s'=s}\big)_{s'\in\mathcal{S}}$	$\mathcal{C}_{\text{gr-reg}} = [0,+\infty)^{N\vert\mathcal{S}\vert}$
Demographic parity	$\mathbf{m}_{\text{DP}}(a,s) = \big(\frac{a}{\gamma_0}\,\mathbf{1}_{s=0},\ \frac{a}{\gamma_1}\,\mathbf{1}_{s=1}\big)$	$\mathcal{C}_{\text{DP}} = \big\{(u,v)\in\mathbb{R}^2 : \ \vert u-v\vert \le \delta\big\}$
Equalized payoffs	$\mathbf{m}_{\text{eq-pay}}(a,y,x,s) = \left(\frac{r(a,y,x,s')}{\gamma_{s'}}\,\mathbf{1}_{s=s'}\right)_{s'\in\{0,1\}}$	$\mathcal{C}_{\text{eq-pay}} = \big\{(u,v)\in\mathbb{R}^2 : \ \vert u-v\vert \le \varepsilon\big\}$

N.B. See other examples in other contexts (Perchet, 2010)

Encoding learning and fairness constraints

Combining the learning goals

$$\begin{cases} \text{Performance goals} & \left(\boldsymbol{m}_{\text{perf}}, \mathcal{C}_{\text{perf}}\right) \\ \text{Fairness goals} & \left(\boldsymbol{m}_{\text{fair}}, \mathcal{C}_{\text{fair}}\right) \end{cases} \Longrightarrow \left((\boldsymbol{m}_{\text{perf}}, \boldsymbol{m}_{\text{fair}}), \mathcal{C}_{\text{perf}} \times \mathcal{C}_{\text{fair}}\right)$$

Online learning setting: formal description

We model our fair online learning problem as a
contextual learning game between the Learner and Nature.

Stochastic attributes (context)

At each time t, the attributes (x_t, s_t) are sampled according to $\mathbf{Q}$, independently from the past

---

Nature (un)awareness

Let G denotes Nature (un)awareness mapping

- Nature *awareness* $G(x, s) = (x, s)$,
- Nature *unawareness*: $G(x, s) = x$.

Online learning setting: formal description

Learning setting

For $t = 1, 2, \ldots$

1. Simultaneously,
   - the Learner chooses $(\mathbf{p}_t^x)_{x \in \mathcal{X}}$ based on $(\mathbf{m}_\tau, x_\tau, s_\tau)_{\tau \leq t-1}$
   - Nature chooses $\left(\mathbf{q}_t^{G(x,s)}\right)_{(x,s) \in \mathcal{X} \times \mathcal{S}}$ based on $(a_\tau, y_\tau, x_\tau, s_\tau)_{\tau \leq t-1}$
2. (x_t, s_t) are sampled according to $\mathbf{Q}$, independently from the past
3. Simultaneously
   - the Learner observes x_t, and picks an action $a_t \in \mathcal{A}$ according to $\mathbf{p}_t^{x_t}$
   - Nature observes $G(x_t, s_t)$, and picks $y_t \in \mathcal{Y}$ according to $\mathbf{q}_t^{G(x_t, s_t)}$
4. The Learner observes the payoff $\mathbf{m}_t = \mathbf{m}(a_t, y_t, x_t, s_t)$ and s_t, while Nature observes (a_t, y_t, x_t, s_t).

Aim: The Learner wants to ensure that $\displaystyle\frac{1}{T}\sum_{t=1}^{T} \mathbf{m}_t \to \mathcal{C}$ a.s. for some target set $\mathcal{C}$.

Online learning setting: formal description

Learning setting

For $t = 1, 2, \ldots$

1. Simultaneously,
   - the Learner chooses $(\mathbf{p}_t^x)_{x \in \mathcal{X}}$ based on $(\mathbf{m}_\tau, x_\tau, s_\tau)_{\tau \leq t-1}$
   - Nature chooses $\left(\mathbf{q}_t^{G(x,s)}\right)_{(x,s) \in \mathcal{X} \times \mathcal{S}}$ based on $(a_\tau, y_\tau, x_\tau, s_\tau)_{\tau \leq t-1}$
2. (x_t, s_t) are sampled according to $\mathbf{Q}$, independently from the past
3. Simultaneously
   - the Learner observes x_t, and picks an action $a_t \in \mathcal{A}$ according to $\mathbf{p}_t^{x_t}$
   - Nature observes $G(x_t, s_t)$, and picks $y_t \in \mathcal{Y}$ according to $\mathbf{q}_t^{G(x_t,s_t)}$
4. The Learner observes the payoff $\mathbf{m}_t = \mathbf{m}(a_t, y_t, x_t, s_t)$ and s_t, while Nature observes (a_t, y_t, x_t, s_t).

Aim: The Learner wants to ensure that $\displaystyle\frac{1}{T}\sum_{t=1}^{T} \mathbf{m}_t \to \mathcal{C}$ a.s. for some target set $\mathcal{C}$.

Online learning setting: formal description

Learning setting

For $t = 1, 2, \dots$

1. Simultaneously,
   - the Learner chooses $(\mathbf{p}_t^x)_{x \in \mathcal{X}}$ based on $(\mathbf{m}_\tau, x_\tau, s_\tau)_{\tau \le t-1}$
   - Nature chooses $\left(\mathbf{q}_t^{G(x,s)}\right)_{(x,s) \in \mathcal{X} \times \mathcal{S}}$ based on $(a_\tau, y_\tau, x_\tau, s_\tau)_{\tau \le t-1}$
2. (x_t, s_t) are sampled according to $\mathbf{Q}$, independently from the past
3. Simultaneously
   - the Learner observes x_t, and picks an action $a_t \in \mathcal{A}$ according to $\mathbf{p}_t^{x_t}$
   - Nature observes $G(x_t, s_t)$, and picks $y_t \in \mathcal{Y}$ according to $\mathbf{q}_t^{G(x_t,s_t)}$
4. The Learner observes the payoff $\mathbf{m}_t = \mathbf{m}(a_t, y_t, x_t, s_t)$ and s_t, while Nature observes (a_t, y_t, x_t, s_t).

Aim: The Learner wants to ensure that $\displaystyle\frac{1}{T}\sum_{t=1}^{T} \mathbf{m}_t \to \mathcal{C}$ a.s. for some target set $\mathcal{C}$.

3- Blackwell Approachability: a reminder

Blackwell approachability : the setup

Setup

1. For the Player: finite set of actions $\mathcal{A}$
2. For the Nature: finite set of actions $\mathcal{B}$
3. A vector-valued pay-off function $\mathbf{m} : \mathcal{A} \times \mathcal{B} \to \mathbb{R}^d$
4. A target set $\mathcal{C} \subset \mathbb{R}^d$

Game

For $t = 1, 2, \ldots$

1. Player and Nature simultaneously pick $\mathbf{p}_t \in \mathcal{P}(\mathcal{A})$ and $\mathbf{q}_t \in \mathcal{P}(\mathcal{B})$
2. $(a_t, b_t) \in \mathcal{A} \times \mathcal{B}$ is sampled according to $\mathbf{p}_t \otimes \mathbf{q}_t$
3. Player observes the payoff $\mathbf{m}_t := \mathbf{m}(a_t, b_t)$; Nature observes (a_t, b_t)

Goal of the Player: $\bar{m}_T := \dfrac{1}{T} \sum_{t=1}^{T} \mathbf{m}_t \longrightarrow \mathcal{C}$ a.s.

Blackwell's result

Approachable set

The target set $\mathcal{C}$ is $\mathbf{m}$-approachable if the Player manages to achieve the above for any strategy of the Nature

Average payoff

$$\mathbf{m}(\mathbf{p}, \mathbf{q}) := \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} \mathbf{p}(a)\mathbf{q}(b)\mathbf{m}(a, b), \quad \text{for } \mathbf{p} \in \mathcal{P}(\mathcal{A}),\ \mathbf{q} \in \mathcal{P}(\mathcal{B}).$$

Blackwell condition

If $\mathcal{C} \subset \mathbb{R}^d$ is closed convex, then $\mathcal{C}$ is $\mathbf{m}$-approachable iff

$$\forall \mathbf{q} \in \mathcal{P}(\mathcal{B}),\ \exists \mathbf{p} \in \mathcal{P}(\mathcal{A}) \quad \text{s.t.} \quad \mathbf{m}(\mathbf{p}, \mathbf{q}) \in \mathcal{C}$$

(Blackwell, 1956)

Proof of Blackwell approachability 1/3

Blackwell's algorithm

Set $\bar{\mathbf{m}}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{m}_t$. At stage $t+1$, choose

$$\mathbf{p}_{t+1} \in \operatorname*{argmin}_{\mathbf{p}\in\mathcal{P}(\mathcal{A})} \max_{\mathbf{q}\in\mathcal{P}(\mathcal{B})} \langle \bar{\mathbf{m}}_t - \Pi_{\mathcal{C}}\bar{\mathbf{m}}_t, \mathbf{m}(\mathbf{p},\mathbf{q})\rangle \tag{1}$$

L^2 convergence: proof sketch

Expanding the squares with $\bar{\mathbf{m}}_{t+1} = \frac{t}{t+1}\bar{\mathbf{m}}_t + \frac{1}{t+1}\mathbf{m}_{t+1}$

$$\begin{aligned} d(\bar{\mathbf{m}}_{t+1},\mathcal{C})^2 &\le \|\bar{\mathbf{m}}_{t+1} - \Pi_{\mathcal{C}}\bar{\mathbf{m}}_t\|^2 \\ &= \left(\frac{t}{t+1}\right)^2 \underbrace{\|\bar{\mathbf{m}}_t - \Pi_{\mathcal{C}}\bar{\mathbf{m}}_t\|^2}_{=d(\bar{\mathbf{m}}_t,\mathcal{C})^2} + \frac{\|\mathbf{m}_{t+1} - \Pi_{\mathcal{C}}\bar{\mathbf{m}}_t\|^2}{(t+1)^2} \\ &\quad + \frac{2t}{(t+1)^2} \underbrace{\langle \bar{\mathbf{m}}_t - \Pi_{\mathcal{C}}\bar{\mathbf{m}}_t, \mathbf{m}_{t+1} - \Pi_{\mathcal{C}}\bar{\mathbf{m}}_t\rangle}_{=:C_{t+1}} \end{aligned}$$

Proof of Blackwell approachability 2/3

According to min-max theorem for bilinear functions, Blackwell condition and the convexity of $\mathcal{C}$

$$\begin{aligned} C_{t+1} &= \langle \bar{\mathbf{m}}_t - \Pi_{\mathcal{C}} \bar{\mathbf{m}}_t, \mathbf{m}_{t+1} - \Pi_{\mathcal{C}} \bar{\mathbf{m}}_t \rangle \\ &\leq \underbrace{\langle \bar{\mathbf{m}}_t - \Pi_{\mathcal{C}} \bar{\mathbf{m}}_t, \mathbf{m}_{t+1} - \mathbf{m}(\mathbf{p}_{t+1}, \mathbf{q}_{t+1}) \rangle}_{=Z_{t+1}} \\ &\quad + \underbrace{\max_{\mathbf{q}} \langle \bar{\mathbf{m}}_t - \Pi_{\mathcal{C}} \bar{\mathbf{m}}_t, \mathbf{m}(\mathbf{p}_{t+1}, \mathbf{q}) - \Pi_{\mathcal{C}} \bar{\mathbf{m}}_t \rangle}_{=\max_{\mathbf{q}} \min_{\mathbf{p}} \langle \bar{\mathbf{m}}_t - \Pi_{\mathcal{C}} \bar{\mathbf{m}}_t, \mathbf{m}(\mathbf{p},\mathbf{q}) - \Pi_{\mathcal{C}} \bar{\mathbf{m}}_t \rangle \leq 0} \end{aligned}$$

The term Z_{t+1} is a martingale increment, i.e. $\mathbb{E}[Z_{t+1}|H_t] = 0$, so

$$\mathbb{E}\left[d(\bar{\mathbf{m}}_{t+1}, \mathcal{C})^2\right] \leq \left(\frac{t}{t+1}\right)^2 \mathbb{E}\left[d(\bar{\mathbf{m}}_t, \mathcal{C})^2\right] + \frac{K}{(t+1)^2}.$$

Hence,

$$\sqrt{\mathbb{E}\left[d(\bar{\mathbf{m}}_T, \mathcal{C})^2\right]} \leq \sqrt{\frac{K}{T}}.$$

4- Contextual Blackwell Approachability

Reminder: approachability for our online learning setting

Contextual approachability problem

For $t = 1, 2, \dots$

1. Simultaneously,
   - Nature chooses $\left(\mathbf{q}_t^{G(x,s)}\right)_{(x,s)\in\mathcal{X}\times\mathcal{S}}$ based on $(a_\tau, y_\tau, x_\tau, s_\tau)_{\tau\le t-1}$
   - the Learner chooses $(\mathbf{p}_t^x)_{x\in\mathcal{X}}$ based on $(\mathbf{m}_\tau, x_\tau, s_\tau)_{\tau\le t-1}$
2. (x_t, s_t) are sampled according to $\mathbf{Q}$, independent from the past
3. Simultaneously
   - Nature observes $G(x_t, s_t)$, and picks $y_t \in \mathcal{Y}$ according to $\mathbf{q}^{G(x_t,s_t)}$
   - the Learner observes x_t, and picks an action $a_t \in \mathcal{A}$ according to $\mathbf{p}^{x_t}$
4. The Learner observes the payoff $\mathbf{m}_t = \mathbf{m}(a_t, y_t, x_t, s_t)$ and s_t, while Nature observes (a_t, y_t, x_t, s_t).

Aim: The Learner wants to ensure that $\bar{m}_T := \dfrac{1}{T}\sum_{t=1}^{T} \mathbf{m}_t \to \mathcal{C}$ a.s.

Contextual Blackwell approachability

Assumption: fast enough sequential estimation of $\mathbf{Q}$

The Player can build estimators $(\hat{\mathbf{Q}}_t)_{t\geq 1}$ of the unknown distribution $\mathbf{Q}$ such that

$$\mathbb{E}\left[\mathrm{TV}^2(\hat{\mathbf{Q}}_t, \mathbf{Q})\right] = O\left(\big(\log t\big)^{-3}\right) \qquad \text{as } t \to \infty \tag{2}$$

Theorem

If $\mathcal{C} \subset \mathbb{R}^d$ is closed convex, $\mathbf{m}$ is bounded, and assumption (2) is satisfied, then
$\mathcal{C}$ is **m**-approachable iff $\forall(\mathbf{q}^{G(x,s)})_{(x,s)\in\mathcal{X}\times\{0,1\}}$ $\exists(\mathbf{p}^x)_{x\in\mathcal{X}}$ such that

$$\int_{\mathcal{X}\times\mathcal{S}} \mathbf{m}\big(\mathbf{p}^x, \mathbf{q}^{G(x,s)}, x, s\big) d\mathbf{Q}(x, s) \in \mathcal{C}$$

Proof of contextual Blackwell approachability 1/3

Contextual Blackwell algorithm

Set $\mathbf{m}(\mathbf{p},\mathbf{q},\hat{\mathbf{Q}}_t) := \int \mathbf{m}\big(\mathbf{p}^x, \mathbf{q}^{G(x,s)}, x, s\big) d\hat{\mathbf{Q}}_t(x,s)$. At stage $t+1$, choose

$$(\mathbf{p}^x_{t+1})_{x\in\mathcal{X}} \in \operatorname*{argmin}_{(\mathbf{p}^x)_x} \max_{(\mathbf{q}^{G(x,s)})_{x,s}} \langle \bar{\mathbf{m}}_t - \Pi_{\mathcal{C}}\bar{\mathbf{m}}_t, \mathbf{m}(\mathbf{p},\mathbf{q},\hat{\mathbf{Q}}_t)\rangle$$

As for classical Blackwell proof

$$\begin{aligned}\|\bar{\mathbf{m}}_{t+1} - \Pi_{\mathcal{C}}\bar{\mathbf{m}}_t\|^2 \le& \left(\frac{t}{t+1}\right)^2 \|\bar{\mathbf{m}}_t - \Pi_{\mathcal{C}}\bar{\mathbf{m}}_t\|^2 + \frac{K}{(t+1)^2} \\ &+ \frac{2t}{(t+1)^2}\langle \bar{\mathbf{m}}_t - \Pi_{\mathcal{C}}\bar{\mathbf{m}}_t, \mathbf{m}_{t+1} - \mathbf{m}(\mathbf{p}_{t+1},\mathbf{q}_{t+1},\hat{\mathbf{Q}}_t)\rangle \\ &+ \frac{2t}{(t+1)^2}\underbrace{\max_{\mathbf{q}}\langle \bar{\mathbf{m}}_t - \Pi_{\mathcal{C}}\bar{\mathbf{m}}_t, \mathbf{m}(\mathbf{p}_{t+1},\mathbf{q},\hat{\mathbf{Q}}_t) - \Pi_{\mathcal{C}}\bar{\mathbf{m}}_t\rangle}_{=\max_{\mathbf{q}}\min_{\mathbf{p}}\langle \bar{\mathbf{m}}_t - \Pi_{\mathcal{C}}\bar{\mathbf{m}}_t, \mathbf{m}(\mathbf{p},\mathbf{q},\hat{\mathbf{Q}}_t) - \Pi_{\mathcal{C}}\bar{\mathbf{m}}_t\rangle}\end{aligned}$$

If $\mathbf{Q}$ instead of $\hat{\mathbf{Q}}_t$, we could directly conclude as in the original proof.

Proof of contextual Blackwell approachability 2/3

We have yet

$$\left|\langle \bar{\mathbf{m}}_t - \Pi_{\mathcal{C}} \bar{\mathbf{m}}_t, \mathbf{m}(\mathbf{p}, \mathbf{q}, \hat{\mathbf{Q}}_t) - \mathbf{m}(\mathbf{p}, \mathbf{q}, \mathbf{Q})\rangle\right| \leq 2d(\bar{\mathbf{m}}_t, \mathcal{C}) \left\|\mathbf{m}\right\|_\infty \mathrm{TV}(\hat{\mathbf{Q}}_t, \mathbf{Q}).$$

Hence, with the same arguments as in the original proof, we get

$$\begin{aligned}\mathbb{E}\left[d(\bar{\mathbf{m}}_{t+1}, \mathcal{C})^2\right] \leq \left(\frac{t}{t+1}\right)^2 \mathbb{E}\left[d(\bar{\mathbf{m}}_t, \mathcal{C})^2\right] + \frac{K}{(t+1)^2} \\ + \frac{8t\|m\|_\infty}{(t+1)^2} \sqrt{\mathbb{E}\left[d(\bar{\mathbf{m}}_t, \mathcal{C})^2\right]} \sqrt{\mathbb{E}\left[\mathrm{TV}(\hat{\mathbf{Q}}_t, \mathbf{Q})\right]^2}.\end{aligned}$$

Hence, by induction,

$$\sqrt{\mathbb{E}\left[d(\bar{\mathbf{m}}_T, \mathcal{C})^2\right]} \leq \sqrt{\frac{K}{T}} + \frac{4\|m\|_\infty}{T} \sum_{t=1}^{T-1} \sqrt{\mathbb{E}\left[\mathrm{TV}(\hat{\mathbf{Q}}_t, \mathbf{Q})\right]^2}.$$

3- Application: Optimal Trade-off between Demographic Parity and Group-Calibration

Deriving optimal trade-offs from Blackwell condition

Why is it useful?

- Blackwell condition allows to investigate optimal trade-offs between learning and fairness objectives.
- Blackwell strategy provides an algorithm for achieving this optimal trade-off.

Contextual Blackwell condition

If $\mathcal{C} \subset \mathbb{R}^d$ is a closed convex, $\mathbf{m}$ is bounded, and assumption (2) is satisfied, then
$\mathcal{C}$ is **m**-approachable iff $\forall (\mathbf{q}^{G(x,s)})_{(x,s)\in\mathcal{X}\times\{0,1\}}\ \exists (\mathbf{p}^x)_{x\in\mathcal{X}}$ such that

$$\int_{\mathcal{X}\times\mathcal{S}} \mathbf{m}\big(\mathbf{p}^x, \mathbf{q}^{G(x,s)}, x, s\big) d\mathbf{Q}(x,s) \in \mathcal{C}$$

Objectives

Demographic Parity (DP) and Group Cal (GrCal)

Learning objective: in the learning problem with $\mathcal{S} = \{0, 1\}$ and $\mathcal{Y} = [0, 1]$, we want to have,

$$\limsup_{T\to\infty} \left| \frac{1}{\gamma_0 T} \sum_{t=1}^{T} a_t \mathbf{1}_{s_t=0} - \frac{1}{\gamma_1 T} \sum_{t=1}^{T} a_t \mathbf{1}_{s_t=1} \right| \leq \delta,$$

and

$$\limsup_{T\to\infty} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \left| \frac{1}{\gamma_s T} \sum_{t=1}^{T} (a - y_t) \mathbf{1}_{a_t=a} \mathbf{1}_{s_t=s} \right| \leq \varepsilon,$$

where $\gamma_s = \mathbf{Q}(s_t = s)$.

Question: What values of (ε, δ) are achievable?

Blackwell approachability condition

Blackwell condition

Approchable <u>iff</u> $\forall(\mathbf{q}^{G(x,s)})_{(x,s)\in\mathcal{X}\times\{0,1\}}$ $\exists(\mathbf{p}^x)_{x\in\mathcal{X}}$ such that

$$\left\| \int_{\mathcal{X}\times\mathcal{S}} \mathbf{m}_{\text{gr-cal}}\big(\mathbf{p}^x, \mathbf{q}^{G(x,s)}, x, s\big) d\mathbf{Q}(x,s) \right\|_1 \leq \varepsilon$$

$$\Delta\left(\int_{\mathcal{X}\times\mathcal{S}} \mathbf{m}_{\text{DP}}\big(\mathbf{p}^x, \mathbf{q}^{G(x,s)}, x, s\big) d\mathbf{Q}(x,s) \right) \leq \delta$$

with $\Delta(u_1, u_2) = |u_1 - u_2|$.

Maximal DP violation

We always have $\Delta(\ldots) \leq \mathrm{TV}(\mathbf{Q}^0, \mathbf{Q}^1)$, where $\mathbf{Q}^s = \mathbf{Q}(\cdot|s_t = s)$. So, we can restrict to

$$\delta_\tau = \tau \cdot \mathrm{TV}(\mathbf{Q}^0, \mathbf{Q}^1), \quad \text{with } \tau \in [0,1].$$

Pareto frontier

Pareto frontier

We identify $\varepsilon^*(\tau)$, the smallest ε such that $\mathcal{C}(\varepsilon, \delta_\tau)$ is approachable.

Nature awareness $G(x, s) = (x, s)$

$$\varepsilon^*(\tau) = 1 - \tau \cdot \mathrm{TV}(\mathbf{Q}^0, \mathbf{Q}^1)$$

Nature unawareness $G(x, s) = x$

$$\varepsilon^*(\tau) = (1 - \tau)\,\mathrm{TV}(\mathbf{Q}^0, \mathbf{Q}^1)$$

N.B. Optimal trade-offs (and hence $\mathcal{C}$) are not known beforehand!

Comments

Nature awareness

- Perfect group-calibration ($\varepsilon = 0$) is never possible, unless $\mathrm{TV}(\mathbf{Q}^0, \mathbf{Q}^1) = 1$ (and $\tau = 1$ is picked, i.e. no DP constraint).
- It corresponds to the case where the supports of $\mathbf{Q}^0$ and $\mathbf{Q}^1$ are disjoint, hence allowing the Player to infer the sensitive context s from the non-sensitive one x.

Nature unawareness

- Perfect group-calibration is always possible by setting $\tau = 1$, no matter the value of $\mathrm{TV}(\mathbf{Q}^0, \mathbf{Q}^1)$.
- If $\mathrm{TV}(\mathbf{Q}^0, \mathbf{Q}^1) = 0$, i.e., $x_t \perp\!\!\!\perp s_t$, then the Player is able to achieve perfect Group-calibration and demographic parity simultaneously.

An important extension

Limitation: the target set $\mathcal{C}$ has to be known

Case of unknown target set

The results can be extended (at the price of some technicalities) to the case where we only have a consistent super-estimate $\hat{\mathcal{C}}_t$ of $\mathcal{C}$.

Strategy unknown target set

The strategy is to work by phases, applying the Blackwell algorithm with $\mathcal{C}$ replaced by $\hat{\mathcal{C}}_{2^r}$ for $2^r \leq t \leq 2^{r+1} - 1$.

⚠ Some stats and probabilistic bounds are hidden there!

Thank you !

Take home message

- Adversarial fair online learning can be cast as an approachability problem
- Blackwell approachability theory can be adapted to a contextual setting with unknown approachability sets
- It provides (benchmark) algorithms
- It allows for a systematic investigation of the trade-offs between learning / fairness constraints (or some other constraints objectives?)

Some supplemental material

Examples of biased AI

Twitter cropping

Twitter automatically crops large images in order to fit the size of an average mobile screen.

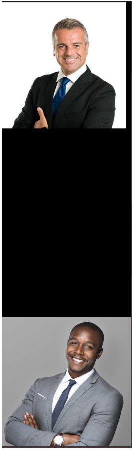
Original

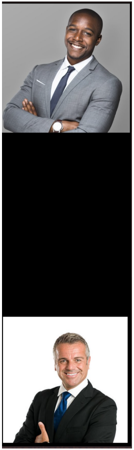
Cropped

Examples of biased AI

Question:

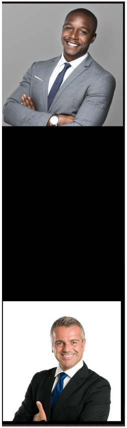
How will Twitter crop these two images?





Examples of biased AI

The two outomes

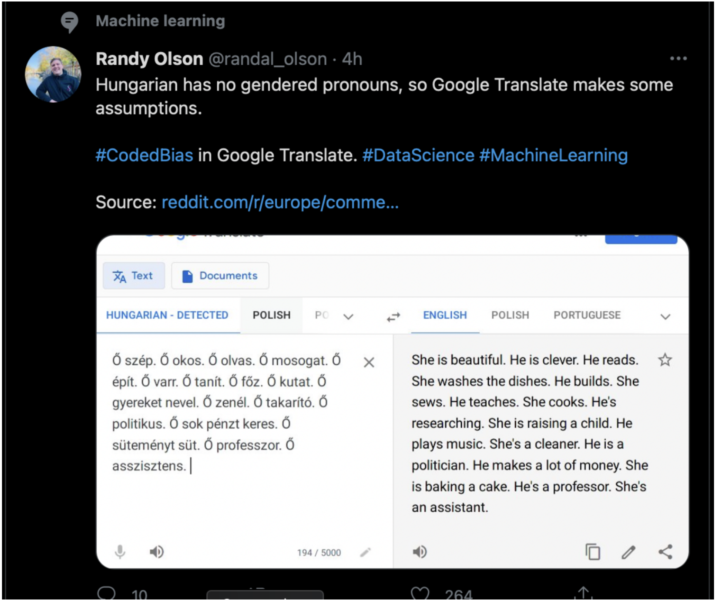


Examples of biased AI

With more famous people?

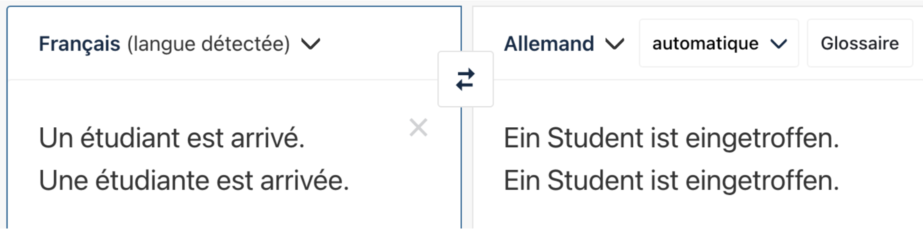
Examples of biased AI

Automatic translation reproduces gender stereotypes



Examples of biased AI

Automatic translation reproduces gender stereotypes



Nature awareness

Nature awareness $G(x, s) = (x, s)$

$$\varepsilon^*(\tau) = 1 - \tau \cdot \mathrm{TV}(\mathbf{Q}^0, \mathbf{Q}^1)$$

Worse Nature strategy: Set $\mathbf{q}^{(x,0)} = \delta_1$ and $\mathbf{q}^{(x,1)} = \delta_0$.

$$\begin{aligned}
\text{Gr-Cal} &= \sum_{a \in \mathcal{A}} \left| \int_{\mathcal{X}} \mathbf{p}^x(a)(a-1)\, d\mathbf{Q}^0(x) \right| + \sum_{a \in \mathcal{A}} \left| \int_{\mathcal{X}} \mathbf{p}^x(a) a \, d\mathbf{Q}^1(x) \right| \\
&= \int_{\mathcal{X}} \sum_{a \in \mathcal{A}} \mathbf{p}^x(a)\, d\mathbf{Q}^0(x) + \underbrace{\int_{\mathcal{X}} \sum_{a \in \mathcal{A}} \mathbf{p}^x(a) a \left(d\mathbf{Q}^1(x) - d\mathbf{Q}^0(x) \right)}_{\text{absolute value equals DP}} \\
&\geq 1 - \mathrm{DP}
\end{aligned}$$

Pareto p-strategy: with probability $1 - \tau$ play $a = 1/2$, with probability τ play $a = \mathbf{q}^{(x,0)}(1)\mathbf{1}_{\mathbf{Q}^0(x) > \mathbf{Q}^1(x)} + \mathbf{q}^{(x,1)}(1)\mathbf{1}_{\mathbf{Q}^1(x) \geq \mathbf{Q}^0(x)}$

Nature unawareness: lower bound

Nature awareness $G(x, s) = x$

$$\varepsilon^*(\tau) \geq (1 - \tau) \cdot \mathrm{TV}(\mathbf{Q}^0, \mathbf{Q}^1)$$

Worst Nature strategy:
Set $\mathbf{q}^x = \delta_1$ if "$\mathbf{Q}^1(x) \geq \mathbf{Q}^0(x)$" and $\mathbf{q}^x = \delta_0$ else.

Pareto p-strategy: with probability $1 - \tau$ play

$$a = \int_{u \in \mathcal{X}} \mathbf{q}^u(1) \frac{d\mathbf{Q}^0(u) + d\mathbf{Q}^1(u)}{2}$$

with probability τ play $a = \mathbf{q}^x(1)$