Trading-off Privacy for Fair Allocation

Vianney Perchet with M. Molina, N. Gast & P. Loiseau

ENSAE & Criteo AI Lab

Criteo's Trustworthy AI Symposium 2025

- Joint team between **Criteo**, **ENSAE** and **Inria**
- Led by, and since March 2022,
  - Patrick Loiseau (Inria) and
  - Vianney Perchet (Criteo & ENSAE)
- Working on "data-marketplace design"
  - **Matching** offer and demand
  - Combining datasets: **mecanism design**
  - **Ethical** questions
- Large and active group ($\simeq$ 10 permanent, $\simeq$ 20 juniors)

- **(Differential) Privacy** Protected attributes should be kept secret
- **Algorithmic Fairness** Users with different protected attributes should be treated the same

- **Example:** smoker/non-smoker dataset
  - Insurance companies need to know the proportion p of smokers
  - Dataset: $X_i = 1$ if user i smokes (vs $X_i = 0$) : $p = \frac{\sum_i X_i}{T}$
  - I do not want my insurance to know that I smoke (or not)
  - At least, I want to be able to deny it

- **A solution:** ε-differential privacy
  - Dataset: $\tilde{X}_i = X_i$ with probability $1 - \varepsilon$ and $\tilde{X}_i = 1 - X_i$
  - Noisy prop. $\tilde{p} = \frac{\sum_i \tilde{X}_i}{T} \simeq (1-\varepsilon)p + \varepsilon(1-p) = p(1-2\varepsilon) + \varepsilon$

- **The message:** We can add noise for privacy (and keep signal)
  - The **higher** the noise, the **more private**, but **less informative**.

- **Binary classification:** Predict credit (non-)failure $Y = 1$

  Based on feature $X_i \in \mathcal{X}$, predicts $Y_i \in \{0; 1\}$

  Sensible attribute $A \in \{a, b\}$ [gender, ethnicity]
- "**Fair**" algorithm w.r.t. the sensible variable A
  - **Many** different notions of fairness
  - Incompatible and/or irreconcilable
- First, natural (?) concept **Independence**

$$\mathbb{P}\{\hat{Y} = 1 | A = a\} = \mathbb{P}\{\hat{Y} = 1 | A = b\}$$

- What if Y is **correlated** to A ? (before or after "selection")

Two other, refined, concepts

- **Independence** (of $\hat{Y}$ and A) Pb if Y **correlated** to A

$$\mathbb{P}\{\hat{Y}=1| \quad A=a\}=\mathbb{P}\{\hat{Y}=1| \quad A=b\}$$

- **Separation:** Independence of $\hat{Y}$ and A conditionally to Y

$$\mathbb{P}\{\hat{Y}=1| \quad A=a, Y=y\}=\mathbb{P}\{\hat{Y}=1| \quad A=b, Y=y\}$$

- **Sufficiency** Independence of Y and A conditionally to $\hat{Y}$

$$\mathbb{P}\{Y=1| \quad A=a, \hat{Y}=y\}=\mathbb{P}\{Y=1| \quad A=b, \hat{Y}=y\}$$

- If 100% of women reimburse their credit and only 50% of men ?
  - Either predict 50% to women or 100% to men...

- **Independence** (of $\hat{Y}$ and A) Pb if Y **correlated** to A

$$\mathbb{P}\{\hat{Y}=1|X_i, A=a\} = \mathbb{P}\{\hat{Y}=1|X_i, A=b\}$$

- **Separation:** Independence of $\hat{Y}$ and A conditionally to Y

$$\mathbb{P}\{\hat{Y}=1|X_i, A=a, Y=y\} = \mathbb{P}\{\hat{Y}=1|X_i, A=b, Y=y\}$$

- **Sufficiency** Independence of Y and A conditionally to $\hat{Y}$

$$\mathbb{P}\{Y=1|X_i, A=a, \hat{Y}=y\} = \mathbb{P}\{Y=1|X_i, A=b, \hat{Y}=y\}$$

- If 100% of women reimburse their credit and only 50% of men ?
  - Either predict 50% to women or 100% to men...
  - Maybe, if lucky, additional features X_i ?

1. If A & Y not independent, then **independence** and **sufficiency** cannot hold simultaneously

2. If A & Y not independent and $\hat{Y}$ & Y not independent, then **independence** and **separation** cannot hold simultaneously

3. If A & Y not independent and all values of $(A, Y, \hat{Y})$ have positive proba, then **sufficiency** and **separation** cannot hold simultaneously

- **Algorithmic Fairness** Protected attributes should be **used** to treat patient the same

- **(Differential) Privacy** Protected attributes should be kept **secret**

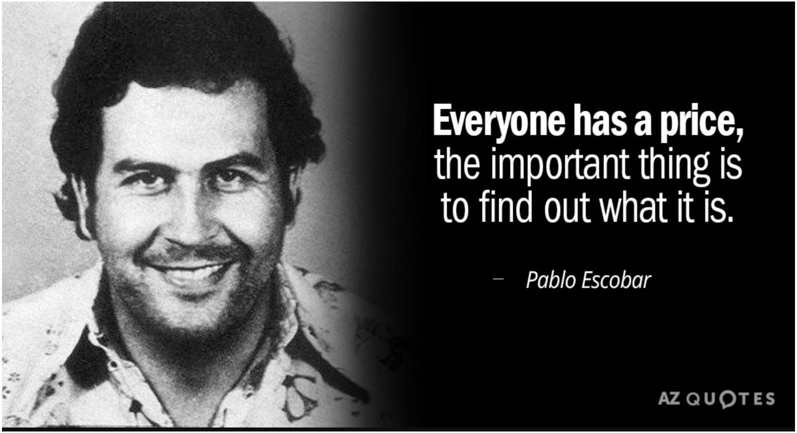
Can they be **reconciled**, and how?

Two “inspirational” (yet controversial) quotes

“You need ethics [but] we should not confuse ethics and intimacy. Young people are ready to share a lot of data”

Paul Hermelin, Chairman of the board of directors of Cap Gemini

Two "inspirational" (yet controversial) quotes



- Stream of users $t = 1, \ldots, T$
- 2 decisions include/not $x_t \in \{0, 1\}$ [or $x_t \in \mathbb{R}^m$]

  Utility $\sum_{t=1}^{T} u_t.x_t$

  u_t known or not (irrelevant to us)
- Protected attributes $a_t \in \{-1, +1\}$ [or $a_t \in \mathbb{R}^d$]

  Fairness measure $R\left(\frac{\sum_{t=1}^{T} a_t x_x}{T}\right)$ [or $R\left(\frac{\sum_{t=1}^{T} a_t x_t}{\sum_t x_t}\right)$]

  Any convex L-Lipchitz function.
- Stochastic data (u_t, a_t) iid

  or **adversarial** any sequence

- **Independence** $\mathbb{P}(X = x, A = i) = \mathbb{P}(X = x)\underbrace{\mathbb{P}(A = i)}_{=\alpha_i}$

  "Decisions are independent of the type"

  $$\text{one-hot encoding } \frac{\sum_t (a_t)_i x_t}{T} = \frac{\sum_t x_t}{T}\frac{\sum_t (a_t)_i}{T} \simeq \frac{\sum_t x_t}{T}\alpha_i$$

  Fairness measure $R\left(\frac{\sum_t (a_t - \alpha)x_t}{T}\right)$ with $R(\cdot) = \|\cdot\|^2$

- **Separation/Sufficiency** $\mathbb{P}(A = i|X = x) = \mathbb{P}(A = i)$

  Fairness measure $R\left(\frac{\sum_t (a_t - \alpha)x_t}{\sum_t x_t}\right)$ with $R(\cdot) = \|\cdot\|^2$

- Privacy Attributes a_t **not** observed
- Costly info. K sources of information

  more (or less) precise, for instance $a_t + \varepsilon_t^{(k)}$ (LDP)

  more (or less) costly. Pay $p^{(k_t)}$ to observe context $c_t^{(k)}$
- Past data $\mathbb{E}[u_t|c_t^{(k)}]$ and $\mathbb{E}[a_t|c_t^{(k)}]$ **known**

  Can be estimated [bandit techniques]
- Public covariates

  Can add $z_t \in \mathbb{R}^n$ and u_t, a_t, p_k functions of it

  [contextual bandit techniques]

Online Ojective

$$\max_{\vec{x},\vec{k}} \underbrace{\sum_{t=1}^{T} u_t x_t}_{\text{utility}} - \underbrace{\sum_{t=1}^{T} p^{(k_t)}}_{\text{cost}} - \underbrace{T.R\left(\frac{\sum_{t=1}^{T} a_t x_t}{T}\right)}_{\text{unfairness penalty}} =: \mathcal{U}(\vec{x},\vec{k})$$

- **Assumption** $[a_t, u_t, c_t^{(1)}, \ldots, c_t^{(k)}]$ are iid
- **Benchmark 1** Static-OPT

$$\max_{k\in[K]} \mathbb{E}\left\{ \max_{\vec{x}} \mathbb{E}\left[\mathcal{U}(k,\vec{x}) | c_1^{(k)}, \ldots, c_T^{(k)}\right]\right\}$$

- **Benchmark 2** Dynamic-OPT

$$\max_{\vec{k}\in[K]} \mathbb{E}\left\{ \max_{\vec{x}} \mathbb{E}\left[\mathcal{U}(\vec{k},\vec{x}) | c_1^{(k_1)}, \ldots, c_T^{(k_T)}\right]\right\}$$

Static-OPT is much worse than Dynamic-OPT !

- A simple balanced model
  - 2 attributes (man 1/woman 2)
  - 2 possible utilities (good $+1$/bad -1)
  - 25% of each pair utility/attribute
  - Fairness measure: **independence**
- Only two sources of information (but weird ones)
  - Source 1 tells if user is a good man
  - Source 2 tells if user is a good woman
  - no information on the sex of bad person
- A single source **cannot** ensure independence
- Using both sources (at random) **can** ensure independence

Objective Linearization

$$\mathbb{E}[\mathcal{U}(\vec{k},\vec{x})] = \mathbb{E}[\sum_t u_t x_t - p^{(k_t)} - T.R(\frac{\sum_t a_t x_t}{T})] \text{ with } \delta_t = \mathbb{E}[a_t | c_t^{(k_t)}] x_t$$

$$\leq \sum_t \mathbb{E}[u_t | c_t^{(k_t)}] x_t - p^{(k_t)} - T.R(\frac{\sum_t \delta_t}{T})$$

$$= \sum_t \mathbb{E}[u_t | c_t^{(k_t)}] x_t - p^{(k_t)} - T \sup_\lambda \left\{ \lambda^\top \frac{\sum_t \delta_t}{T} - R^*(\lambda) \right\}$$

$$= \inf_\lambda \left\{ \sum_t \mathbb{E}[u_t | c_t^{(k_t)}] x_t - p^{(k_t)} - \lambda^\top \delta_t - R^*(\lambda) \right\}$$

$$= \inf_\lambda \sum_{t=1}^{T} \mathcal{L}(\lambda, k_t) \leq T \sup_{\pi \in \mathcal{P}[K]} \inf_\lambda \pi^\top \mathcal{L}(\lambda) \simeq \text{OPT}$$

where $R^*(\lambda) = \sup_{\delta \in \Delta} \delta^\top \lambda - R(\delta)$ is the Fenchel conjugate

Regret Decomposition 1/3

$$\mathbb{E}[\mathcal{U}(\vec{k},\vec{x})] = \mathbb{E}[\sum_t u_t x_t - p^{(k_t)} - T.R(\frac{\sum_t a_t x_t}{T})]$$

$$= \sum_t \left\{ \mathbb{E}[u_t | c_t^{(k_t)}] x_t - p^{(k_t)} - \lambda_t^\top \delta_t + R^*(\lambda_t) \right\} -$$

$$\sum_t R^*(\lambda_t) + \sum_t \lambda_t^\top \delta_t - TR(\frac{\sum_t a_t x_t}{T})$$

$$= \sum_t \mathcal{L}(\lambda_t, k_t) - R^*(\lambda_t) + \lambda_t^\top \delta_t - R(\frac{\sum_t a_t x_t}{T})$$

$$\geq \sum_t \mathcal{L}(\lambda_t, k_t) + R(\gamma_t) + \lambda_t^\top (\delta_t - \gamma_t) - R(\frac{\sum_t a_t x_t}{T})$$

If $\gamma_t = \arg\max_{\gamma : \|\gamma - \delta_t\| \leq \mathrm{diam}(\Delta)} \lambda_t^\top \gamma - R(\gamma)$

$$\begin{aligned}\mathbb{E}[\mathcal{U}(\vec{k},\vec{x})] &\geq \sum_t \mathcal{L}(\lambda_t, k_t) + R(\gamma_t) + \lambda_t^\top(\delta_t - \gamma_t) - R(\frac{\sum_t a_t x_t}{T}) \\ &\geq \sum_t \mathcal{L}(\lambda_t, k_t) + R(\gamma_t) - R(\delta_t) + \lambda_t^\top(\delta_t - \gamma_t) + \\ &\qquad R(\delta_t) - R(\frac{\sum_t a_t x_t}{T}) \\ &\geq \sum_t \mathcal{L}(\lambda_t, k_t) - \hat{\lambda}^\top(\delta_t - \gamma_t) + \lambda_t^\top(\delta_t - \gamma_t) + \\ &\qquad R(\delta_t) - R(\frac{\sum_t a_t x_t}{T})\end{aligned}$$

where $\hat{\lambda} \in \partial R(\frac{\sum_t \delta_t}{T})$

Regret Decomposition 3/3

$$
\begin{aligned}
\mathbb{E}[\mathcal{U}(\vec{k},\vec{x})] \geq & \sum_t \mathcal{L}(\lambda_t, k_t) - \max_\pi \pi^\top \sum_t \mathcal{L}(\lambda_t) \\
& + \sum_t \hat{\lambda}^\top(\gamma_t - \delta_t) - \lambda_t^\top(\gamma_t - \delta_t) \\
& + R\left(\frac{\sum_t \delta_t}{T}\right) - R\left(\frac{\sum_t a_t x_t}{T}\right) \\
& + \max_\pi \pi^\top \sum_t \mathcal{L}(\lambda_t) - T \max_\pi \inf_\lambda \pi^\top \mathcal{L} \\
& + T \max_\pi \inf_\lambda \pi^\top \mathcal{L}
\end{aligned}
$$

- adversarial bandit (arms $k_t \in [K]$),
- Linear bandit (arms $\lambda_t \in \mathbb{R}$),
- Concentration $\geq -L\sqrt{dT}$,
- positive and $\geq$ OPT

Algorithm

- Linear bandit on λ_t with loss $\lambda_t^\top(\gamma_t - \delta_t)$

$$\gamma_t = \arg\max_{\gamma:\|\gamma-\delta_t\|\leq \mathrm{diam}(\Delta)} \lambda_t^\top \gamma - R(\gamma)$$

$$\lambda_{t+1} = \lambda_t + \eta(\delta_t - \gamma_t) \qquad \text{[Gradient Descent]}$$

Regret term in $L\sqrt{dT}$

- EXP3 bandit algo on $\mathcal{D}(\lambda_t, k_t)$

$$\pi_t \propto \exp(-\theta(\textstyle\sum_{s<t} \mathcal{D}(\lambda_s, k_s))) \qquad \text{[Mirror Descent]}$$

Regret term in $\|\lambda\|_\infty \sqrt{TK\log(K)}$

- Concentration $\leq L\sqrt{dT}$,
- **Total Regret** smaller than

$$\Big(L\sqrt{d} + \|\lambda\|_\infty \sqrt{K\log(K)}\Big)\sqrt{T}$$

Bounding λ

- Iteration $\lambda_{t+1} = \lambda_t + \eta(\delta_t - \gamma_t)$ and $\lambda_0 \in \Lambda = \text{conv} \cup_{\delta \in 2\Delta} \partial R(\delta)$ with $\gamma_t = \arg\max_{\gamma : \|\gamma - \delta_t\| \leq \text{diam}(\Delta)} \lambda_t^\top \gamma - R(\gamma)$
- **KKT:** $0 \in -\lambda_t + \partial R(\gamma_t) + \mu(\gamma_t - \delta_t)$, for some $\mu \geq 0$,
- $\lambda_{t+1} = \lambda_t + \alpha(\lambda_{\delta_t} - \lambda_t)$ with $\lambda_{\delta_t} \in \partial R(\gamma_t) \in \Lambda$ and $\alpha \geq 0$
    - If $\alpha \leq 1$, $d(\lambda_{t+1}, \Lambda) \leq d(\lambda_t, \Lambda)$
    - If $\alpha > 1$, $\lambda_{t+1} \in \Lambda + B(0, 2\eta\text{diam}(\Delta)))$
- **Conclusion**

$$\|\lambda_t\|_2 \leq L + 2\eta\text{diam}(\Delta)$$

1. It is possible to **reconcile fairness and privacy** !
   Because Privacy is different from Intimacy.

1. It is possible to **reconcile fairness and privacy** !
   Because Privacy is different from Intimacy.
2. **Sublinear** regret bound

$$\mathbb{E}[\mathcal{U}(\vec{k},\vec{x})] \geq \text{Dynamic-OPT} - \left(L\sqrt{d} + L\sqrt{K\log(K)}\right)\sqrt{T}$$

Three Conclusions

1. It is possible to **reconcile fairness and privacy** !
   Because Privacy is different from Intimacy.
2. **Sublinear** regret bound

$$\mathbb{E}[\mathcal{U}(\vec{k},\vec{x})] \geq \text{Dynamic-OPT} - \left(L\sqrt{d} + L\sqrt{K\log(K)}\right)\sqrt{T}$$

3. I do not know how to handle page counters in Beamer.