Active Fairness Auditing

Tom Yan
Carnegie Mellon University

Chicheng Zhang
University of Arizona

ICML 2022

Auditing machine learning models

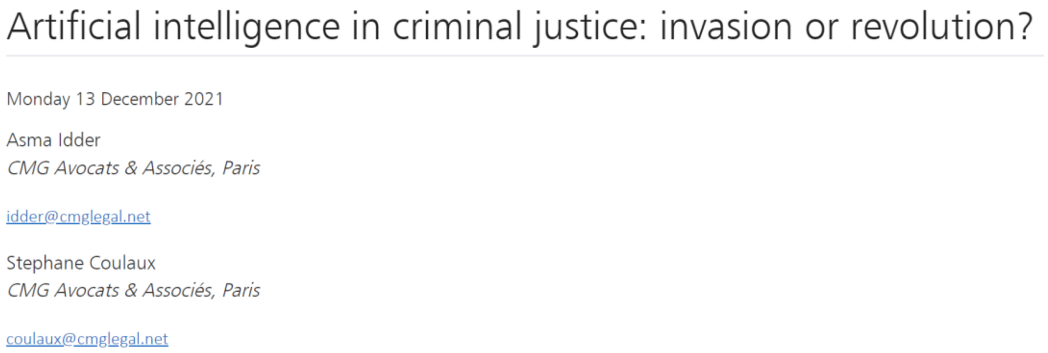
- Machine learning models are increasingly being used for consequential decisions

ARTIFICIAL INTELLIGENCE

Bias isn't the only problem with credit scores—and no, AI can't help

The biggest-ever study of real people's mortgage data shows that predictive tools used to approve or reject loans are less accurate for minorities.

By Will Douglas Heaven June 17, 2021



Artificial intelligence in criminal justice: invasion or revolution?

Monday 13 December 2021

Asma Idder
CMG Avocats & Associés, Paris

idder@cmglegal.net

Stephane Coulaux
CMG Avocats & Associés, Paris

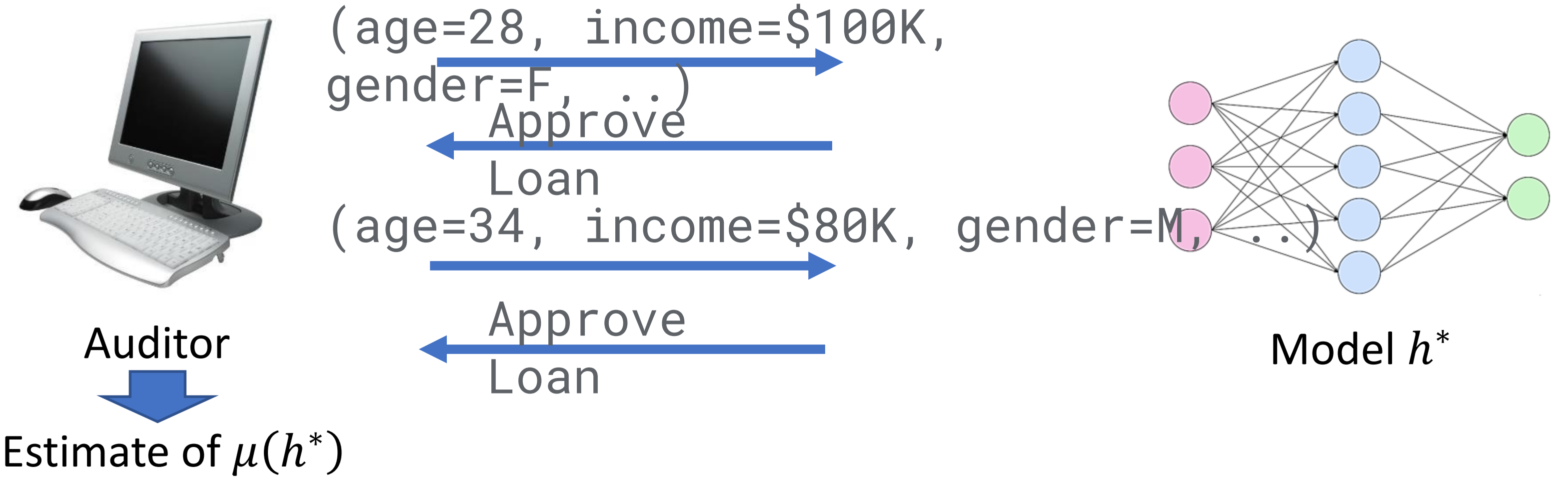
coulaux@cmglegal.net

- How can we efficiently audit the risks of machine learning models?
  - See e.g. Supreme Audit Institutions of Finland, Germany, the Netherlands, Norway and the UK, *Auditing machine learning algorithms: a white paper for public auditors*

This work: active fairness auditing

- Model h^* from a known class $\mathcal{H}$
- Known joint distribution D over feature x and sensitive attribute $x_A \in \{0,1\}$
- With adaptive black-box query access to h^*, how can we efficiently estimate its demographic parity

$$\mu(h^*) = \Pr(h^*(x) = +1 \mid x_A = 1) - \Pr(h^*(x) = +1 \mid x_A = 0)?$$



- Performance measure:
  - Query efficiency
  - Computational efficiency

Related work

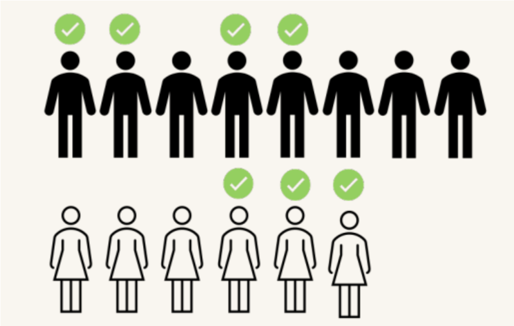
- (Tan et al'18, Rastegarpanah et al'21): auditing model's feature usage
- (Xue et al'20): auditing model's individual fairness
- (Sabato & Yom-Tov'20): bounding model's fairness using its population statistics
- …
- This work: auditing model h^*'s group fairness by assuming access to a hypothesis class that contains h^*

Baselines

- Estimate demographic parity:

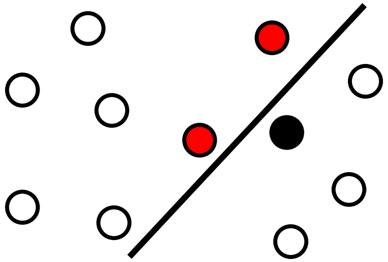
$$\mu(h^*) = \underbrace{\Pr(\, h^*(x) = +1 \mid x_A = 1\,)}_{\gamma_1(h^*)} - \underbrace{\Pr(\, h^*(x) = +1 \mid x_A = 0\,)}_{\gamma_0(h^*)} \text{ to precision } \epsilon$$

Assume that $\min(\Pr(x_A = 1), \Pr(x_A = 0)) = \Omega(1)$



- Baseline 1: i.i.d. sampling
  - Estimate $\gamma_b(h^*)$ using iid draws $D \mid x_A = b$
  - Query complexity: $O(1/\epsilon^2)$

- Baseline 2: PAC active learning
  - Learn $\hat{h}$ such that $\Pr\left(\hat{h}(x) \neq h^*(x)\right) \leq O(\epsilon)$, return $\mu(\hat{h})$
  - Query complexity: active learning's label complexity (e.g. Hanneke'14)



Main results

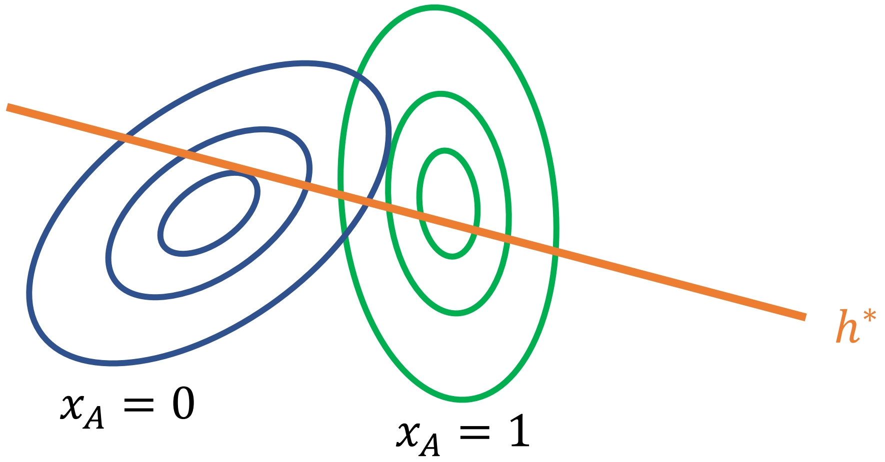
- Separation between active fairness auditing and active learning
  - Two examples: choosing between iid sampling and active learning is information-theoretically optimal

- Algorithms for general $(\mathcal{H}, D)$:
  - Optimal deterministic algorithm
  - Oracle-efficient algorithm with competitive guarantees
  - Manipulation-proof auditing and empirical evaluation

Separation example: linear classification

$D \mid x_A = b : \mathcal{N}(\mu_b, \Sigma_b)$

$\mathcal{H} = \{\mathrm{sign}(\langle w, x \rangle + b) : w \in \mathrm{R}^d, b \in \mathrm{R}\}$



- i.i.d. sampling: $O(1/\epsilon^2)$
- Active learning: $\widetilde{\Theta}(d)$
- $\epsilon \gg \frac{1}{\sqrt{d}} \Rightarrow$ i.i.d. sampling has much lower query complexity
- Information-theoretic lower bound: $\Omega\left(\min(1/\epsilon^2, d)\right)$
- Similar phenomenon happens in another discrete-domain example (see paper)

Main results

- Separation between active fairness auditing and active learning
  - Two examples: choosing between iid sampling and active learning is information-theoretically optimal

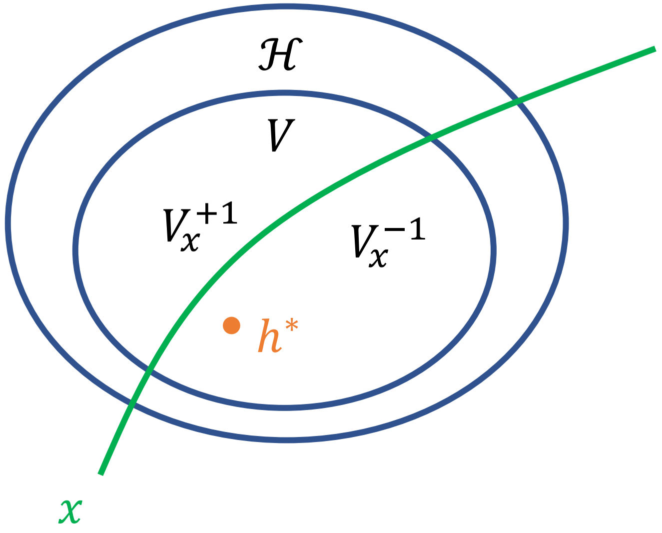
- Algorithms for general $(\mathcal{H}, D)$:
  - Optimal deterministic algorithm
  - Oracle-efficient algorithm with competitive guarantees
  - Manipulation-proof auditing and empirical evaluation

Optimal deterministic algorithm

- Cost complexity of active fairness auditing with version space V:

$$\mathrm{Cost}(V) = \begin{cases} 0, & \mathrm{diam}_\mu(V) := \max_{h,h' \in V} \mu(h) - \mu(h') \le 2\epsilon \\ 1 + \min_{x} \max_{y} \mathrm{Cost}(V_x^y), & \text{otherwise} \end{cases}$$

- Dynamic programming (DP) (cf. Hanneke'06):
  - Maintain V based on current information
  - Query x by minimizing worst-case future costs



Optimal deterministic algorithm

- Theorem (optimality):
  - DP-based algorithm makes at most $\mathrm{Cost}(\mathcal{H})$ queries
  - Any deterministic active fairness auditing algorithm must make $\mathrm{Cost}(\mathcal{H})$ queries

- Comparison with baselines:
  - i.i.d. sampling: $\mathrm{Cost}(\mathcal{H}) \leq O(\ln|\mathcal{H}|/\epsilon^2)$
  - active learning: $\mathrm{Cost}(\mathcal{H}) \leq$ the label complexity bound of CAL (Cohn, Atlas, Ladner'94; Hanneke'14)

- Key drawback of DP: computationally intractable
  - Approximating $\mathrm{Cost}(\mathcal{H})$ within $\mathrm{o}(\log|\mathcal{H}|)$ is NP-Hard

Main results

- Separation between active fairness auditing and active learning
  - Two examples: choosing between iid sampling and active learning is information-theoretically optimal
- Algorithms for general $(\mathcal{H}, D)$:
  - Optimal deterministic algorithm
  - Oracle-efficient algorithm with competitive guarantees
  - Manipulation-proof auditing and empirical evaluation

Oracle-efficient algorithms with competitive guarantees

- Oracle 1: mistake-bounded online learning oracle for $\mathcal{H}$

Example	x_1	x_2	x_3	x_4	x_5	…
Prediction	$-$	$+$	$+$	$-$	$-$	…
Actual label by h^*	$+$	$-$	$+$	$-$	$-$	…

#Mistakes $\leq M$

  - Efficient implementation: Perceptron, Sampling-based Halving (Bertsimas & Vempala ’04) for linear $\mathcal{H}$

- Oracle 2: constrained classification oracle for $\mathcal{H}$
  - Input: labeled dataset S, T
  - Output: $\text{argmin}_{h \in \mathcal{H}} \Pr_S(h(x) \neq y) \quad \text{s.t. } \Pr_T(h(x) \neq y) = 0$
  - Used for efficient active learning, e.g. (Dasgupta et al’07, Huang et al’15)

Oracle-efficient algorithms with competitive guarantees

- Main idea (inspired by Hegedus'95):
  - Reducing active fairness auditing to online learning and *teaching* $\mu(h)$
  - Use the recent online set cover-based teaching algorithm (Dasgupta et al, 2019) to efficiently teach $\mu(h)$ with the classification oracle

- Theorem: our algorithm oracle-efficiently estimates $\mu(h^*)$ with error ϵ, and queries h^* at most $\mathrm{O}(M \cdot \mathrm{Cost}(\mathcal{H}) \cdot \ln|\mathcal{H}|)$ times

Main results

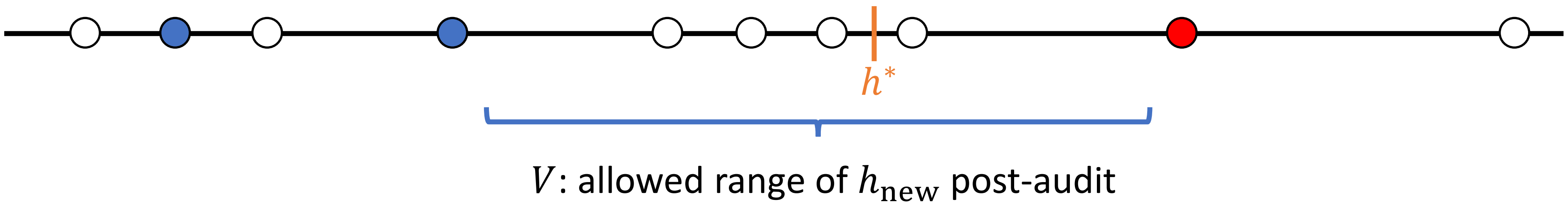
- Separation between active fairness auditing and active learning
  - Two examples: choosing between iid sampling and active learning is information-theoretically optimal

- Algorithms for general $(\mathcal{H}, D)$:
  - Optimal deterministic algorithm
  - Oracle-efficient algorithm with competitive guarantees
  - Manipulation-proof auditing and empirical evaluation

Manipulation-proof auditing

- Motivation: companies may change the model post-audit from h^* to some other $h_{\text{new}} \in \mathcal{H}$ to improve profit

- Constraint: h_{new} in the version space induced by the examples collected in the auditing process

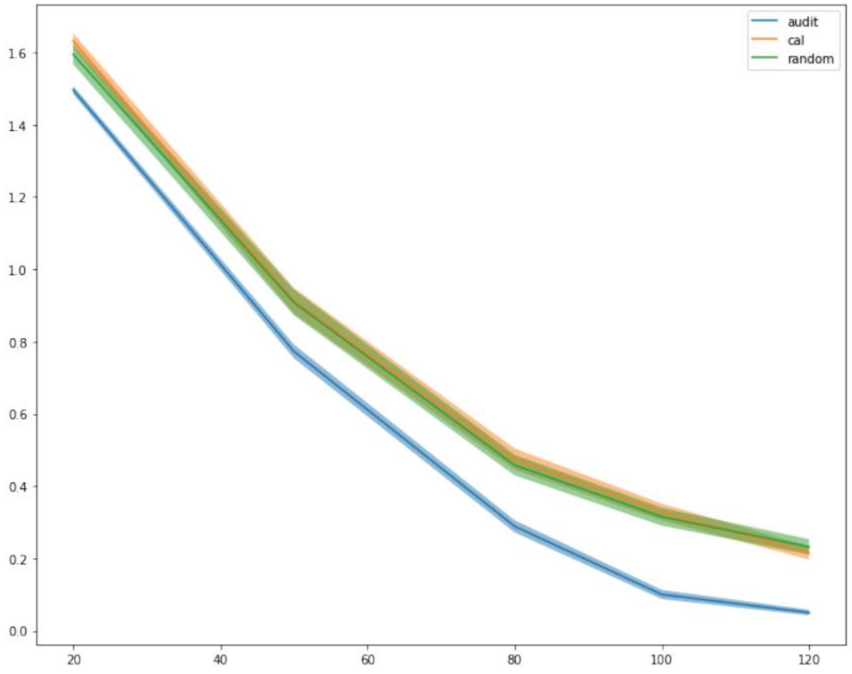


- A set of queries is ϵ-manipulation-proof (MP) if its induced version space V has $\text{diam}_\mu(V) \leq 2\epsilon$
- Observation: our two algorithms & active learning are MP, while iid sampling may not

Empirical evaluation

- Query algorithms: i.i.d. sampling, CAL (active learning), ours

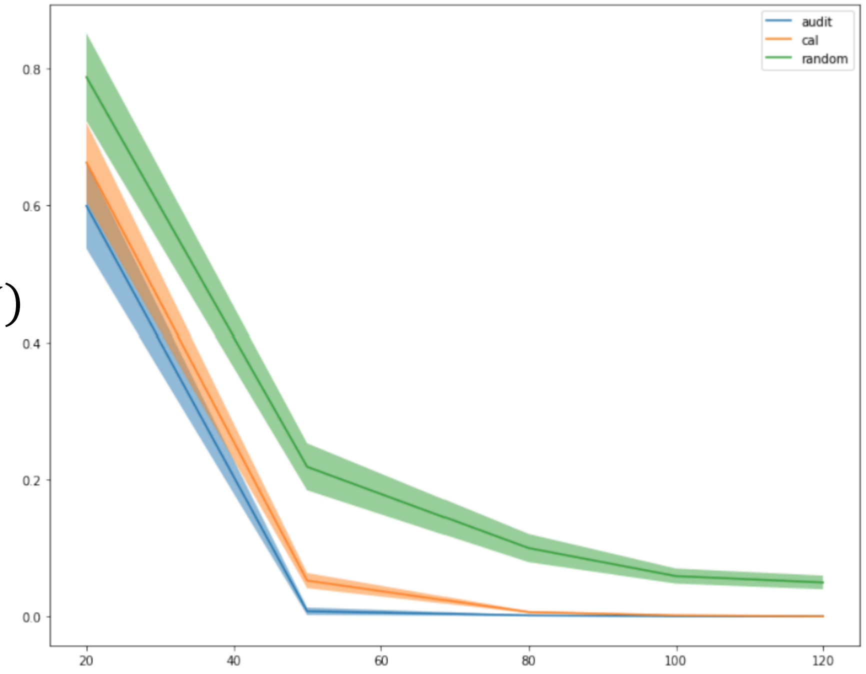
Student Performance



$\mathrm{diam}_{\mu}(V)$

#queries

COMPAS



$\mathrm{diam}_{\mu}(V)$

#queries

Conclusions

- We formulate active fairness auditing, putting responsible machine learning onto a firmer foundation

- We present general and efficient algorithms with query complexity guarantees

- Follow-up work (arXiv update soon):
  - Example when active fairness auditing strategies strictly improve over both baselines
  - Fundamental limitations of manipulation-proof and deterministic auditing

Thank you

arXiv:2206.08450