Auditing Fairness By Betting

by Ben Chugg, Santiago Cortes-Gomez, Bryan Wilder, and Aaditya Ramdas

Dhruv Shah

April 9, 2025

Table of contents

Introduction

Main Idea

Develop sequential methods to track the fairness of real-world systems. Approach based on *Sequential Hypothesis Testing*. The main concepts used in this paper are the following:

- Anytime-valid inference
- Game-theoretic framework ("Testing by betting")

Hypothesis Testing: Brief Overview

Task: Make decisions about the system from the data/information available.

Key Components: - Null Hypothesis (H_0): Assumes no effect or no difference; the default position. - Alternative Hypothesis (H_1): Represents the effect or difference being tested.

Steps:

1. Formulate H_0 and H_1.
2. Choose Level of significance (α).
3. Select a test statistic and compute its value.
4. Decision Rule: Reject H_0 if p-value $\leq \alpha$. Else fail to reject H_0

Types of Error:

- Type - I error: Test reject H_0 when H_0 is true.
- Type - II error: Does not reject H_0 when H_0 is false.

Fairness Audit as Hypothesis Test

Task: Determine the model deployed is fair or not. Given the data concerning the system's decision are *gathered over time*.

Therefore, this can be easily framed as a hypothesis test.
Define H_0: Model is fair. and H_1: Model is unfair.

Here, we might not have fixed, i.i.d data points. So, traditional hypothesis testing isn't realistic. Hence, we need to perform *sequential hypothesis testing*. Therefore, we require the following:

- continuously monitor the data (peeking).
- focus on rejecting null as early as possible.

Sequential Hypothesis Testing

Idea: Continuously test H_0 as data arrives, without specifying a pre-defined sample size.

Method

Thought Experiment

- Imagine a skeptical better evaluating a machine learning system's fairness.
- The better plays an iterated game by betting on audit results over time.
- Betting Strategy:
  - If the system is unfair then Expected payoff is large (wealth increases).
  - If the system is fair then Expected payoff remains small.
- Decision Rule:
  - The null hypothesis of fairness is rejected if the better's wealth surpasses a predetermined threshold.
  - Wealth growth signals potential unfairness in the system.

Definition of Fairness

Main focus on "group" fairness. So, here we ask *"Which groups of individuals are at risk for experiencing harms?"* (Source: Fairlearn)

Definition: Let $\{\xi_j(A, X, Y)\}_{j\in\{0,1\}}$ denote conditions on sensitive attribute A, covariates X, and outcomes Y. A predictive model $\varphi : \mathcal{X} \to [0, 1]$ is fair with respect to $\{\xi_j\}$ if:

$$\mathbb{E}_{X\sim\rho}[\varphi(X) \mid \xi_0(A, X, Y)] = \mathbb{E}_{X\sim\rho}[\varphi(X) \mid \xi_1(A, X, Y)].$$

Fairness Notions:

1. *Equality of Opportunity*: $\xi_0 = \{A = 0, Y = 1\}$, $\xi_1 = \{A = 1, Y = 1\}$.
2. *Predictive Equality*: Similar to above, but for $Y = 0$.
3. *Statistical Parity*: $\xi_0 = \{A = 0\}$, $\xi_1 = \{A = 1\}$.
4. Other fairness notions arise for appropriate choices of conditions ξ_j.

What is a fairness audit?

- **Objective**: Test fairness of a model by comparing predictions for two groups ($b = 0, 1$) over time.
- **Predictions**:

$$Z^0 = \{\varphi(X_t^0)\}_{t \in T_0}, \quad Z^1 = \{\varphi(X_t^1)\}_{t \in T_1},$$

  where T_0 and T_1 are time indices for predictions from groups $b = 0$ and $b = 1$.
- **Goal**: Construct a **sequential hypothesis test**:
  - **Null Hypothesis** (H_0): The model is fair.
  - **Alternative Hypothesis** (H_1): The model is unfair.

- **Time Indices**:

$$T_b[t] = T_b \cap [t],$$

where $T_b[t]$ is the set of times predictions from group b are received up to time t.

- **Test Function**:

$$\phi_t = \phi_t \left(\bigcup_{t \in T_0[t]} Z_t^0, \bigcup_{t \in T_1[t]} Z_t^1 \right),$$

where $\phi_t = 1$ means "reject H_0" and $\phi_t = 0$ means "fail to reject H_0."

- **Stopping Time**:

$$\tau = \inf\{t : \phi_t = 1\}.$$

- **Sequential level-α test**

$$\sup_{P \in H_0} P(\exists t \geq 1 : \phi_t = 1) \leq \alpha \quad \text{or equivalently} \quad \sup_{P \in H_0} P(\tau < \infty) \leq \alpha.$$

To ensure a small false postive rate (Type-I error)

Testing By Betting

Objective: Use a "betting framework" to detect unfairness in a model's predictions between two groups.

Key Idea: A fictitious skeptic places bets on model predictions $(\hat{Y}_t^0, \hat{Y}_t^1)$ under the assumption that the model is fair $(H_0 : \mu_0 = \mu_1)$.

Skeptic's Wealth Process:

$$K_t = \prod_{i=1}^{t} S_i(\hat{Y}_i^0, \hat{Y}_i^1),$$

where S_i is a *payoff function* chosen to maximize wealth growth if H_0 is false.

Betting Mechanism: - At time t, the skeptic bets on the difference in predictions ($\hat{Y}_t^0 - \hat{Y}_t^1$) with a payoff function:

$$S_t = 1 + \lambda_t(\hat{Y}_t^0 - \hat{Y}_t^1),$$

where $\lambda_t \in [-1, 1]$ is adaptively chosen to maximize wealth growth.

Null Hypothesis (H_0): If the model is fair, the wealth K_t is a supermartingale, i.e., it should not grow on average.

Properties and Stopping Rule

Ville's Inequality: Guarantees that under H_0, the probability of $K_t > 1/\alpha$ is at most α.

Stopping Rule: - Reject H_0 (detect unfairness) when:

$$K_t > 1/\alpha.$$

Adaptive Betting Strategy: - Use Online Newton Step (ONS) to choose λ_t, ensuring optimal growth under the alternative hypothesis $(H_1 : \mu_0 \neq \mu_1)$:

$$\lambda_t = \left(\frac{g_t}{2 - \ln(3) + \sum_{i=1}^{t-1} z_i^2} \right) \wedge 1 \vee -1,$$

where $g_t = \hat{Y}_t^0 - \hat{Y}_t^1$, and $(z_i = g_i/(1 - \lambda_i g_i)$.

Interpretation: - If the model is unfair (H_1), K_t grows exponentially, leading to early rejection of H_0.

- The method guarantees a *controlled false positive rate* (α) and high power.

THANK YOU!