Enhancing Group Fairness in Online Settings Using Oblique Decision Forests

Somnath Basu Roy Chowdhury[1,3] Nicholas Monath[2] Ahmad Beirami[1] Rahul Kidambi[1]
Kumar Avinava Dubey[1] Amr Ahmed[1] Snigdha Chaturvedi[3]

[1] Google Research [2] Google DeepMind [3] UNC NLP

Motivation

- ML systems often produce unfair decisions against certain groups
- We study the challenging problem of achieving fairness in online settings

Group Fairness

Group Fairness techniques focus on enhancing the fairness of ML algorithms by ensuring that different groups receive equal treatment.

Batch-wise Group Fairness

- In batch-wise settings, a learning function f can be optimized as shown:

$$\min_{f} L(f(x), y), \text{ subject to } |\mathbb{E}[f(x \,|\, a = 0)] - \mathbb{E}[f(x \,|\, a = 1)]| < \epsilon\,.$$

a is the sensitive attribute (e.g., gender)

Batch-wise Group Fairness

- In batch-wise settings, a learning function f can be optimized as shown:

$$\min_f L(f(x), y), \text{ subject to } |\mathbb{E}[f(x|a=0)] - \mathbb{E}[f(x|a=1)]| < \epsilon \,.$$

Prediction for group 0

Batch-wise Group Fairness

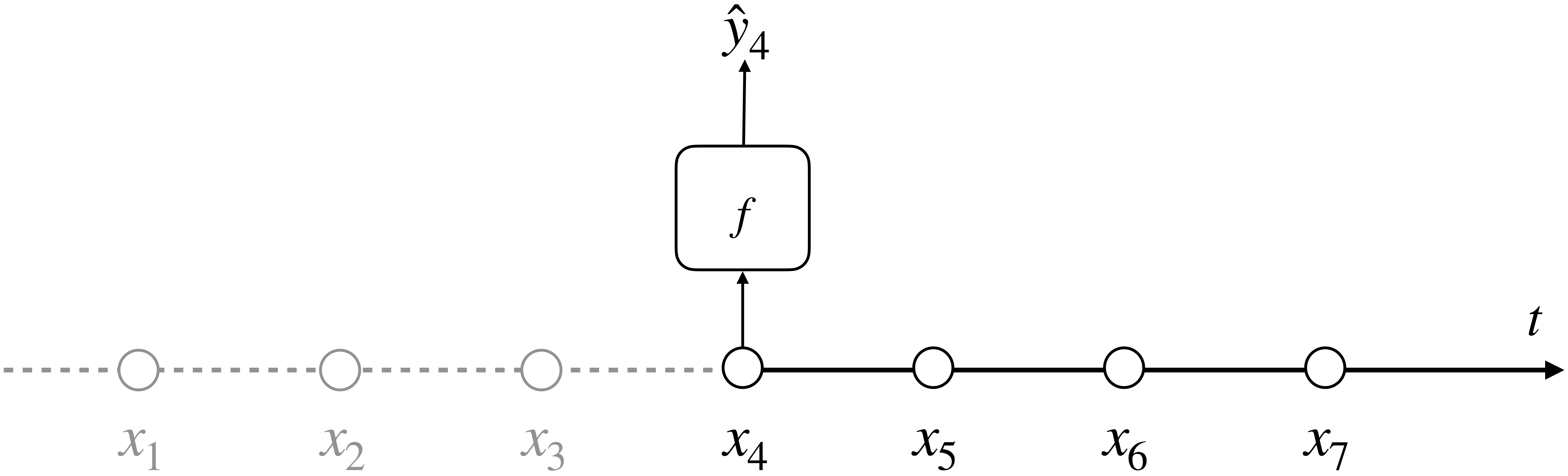
- In batch-wise settings, a learning function f can be optimized as shown:

$$\min_f L(f(x), y), \text{ subject to } |\mathbb{E}[f(x \mid a = 0)] - \mathbb{E}[f(x \mid a = 1)]| < \epsilon .$$

Difference between predictions of two groups

Online Setting

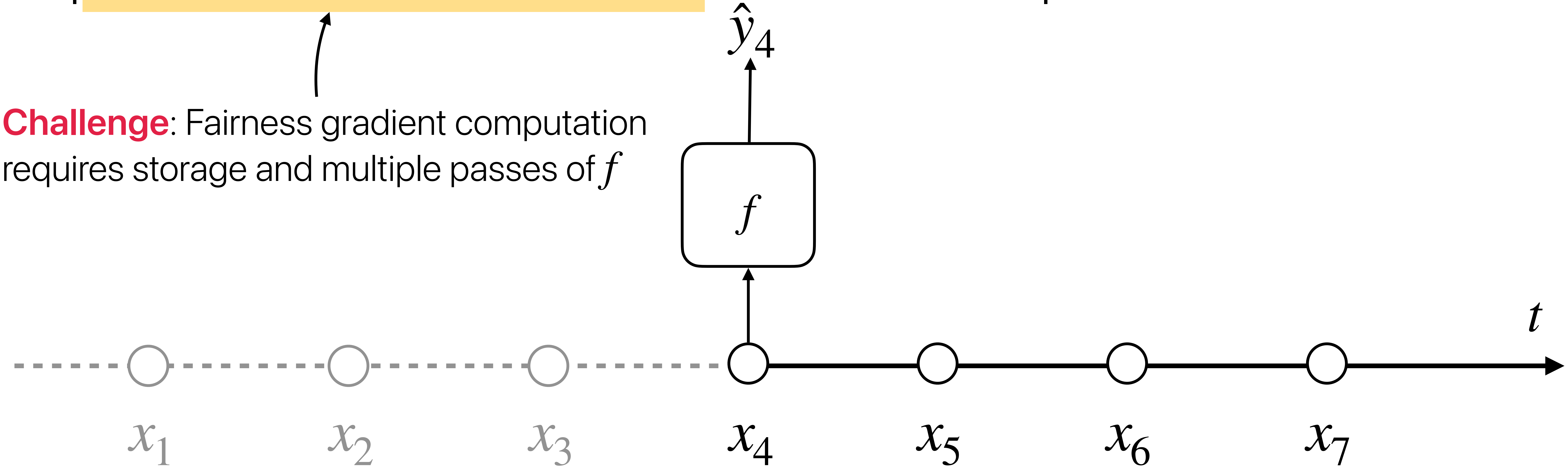
- In online setup, input points $x_1, x_2, \ldots$ arrive one at a time



Online Setting

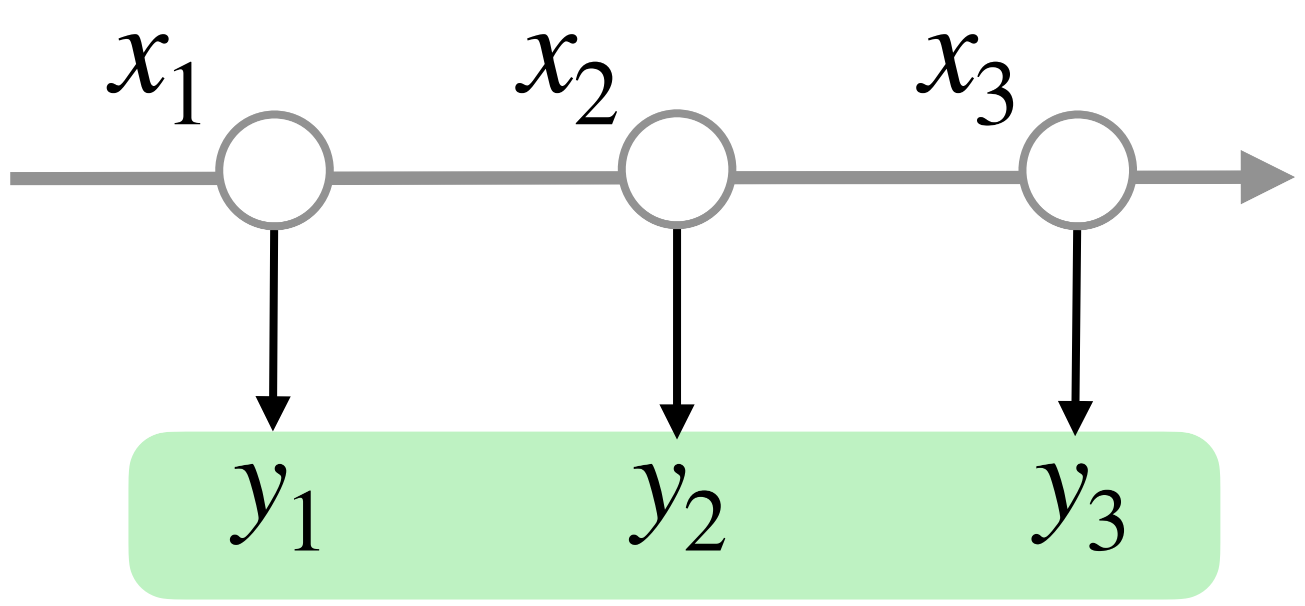
$$\left| \frac{f(x_1 \mid a = 0) + \ldots + f(x_n \mid a = 0)}{n} - \mathbb{E}[f(x \mid a = 1)] \right| < \epsilon \, .$$

Challenge: Fairness gradient computation requires storage and multiple passes of f



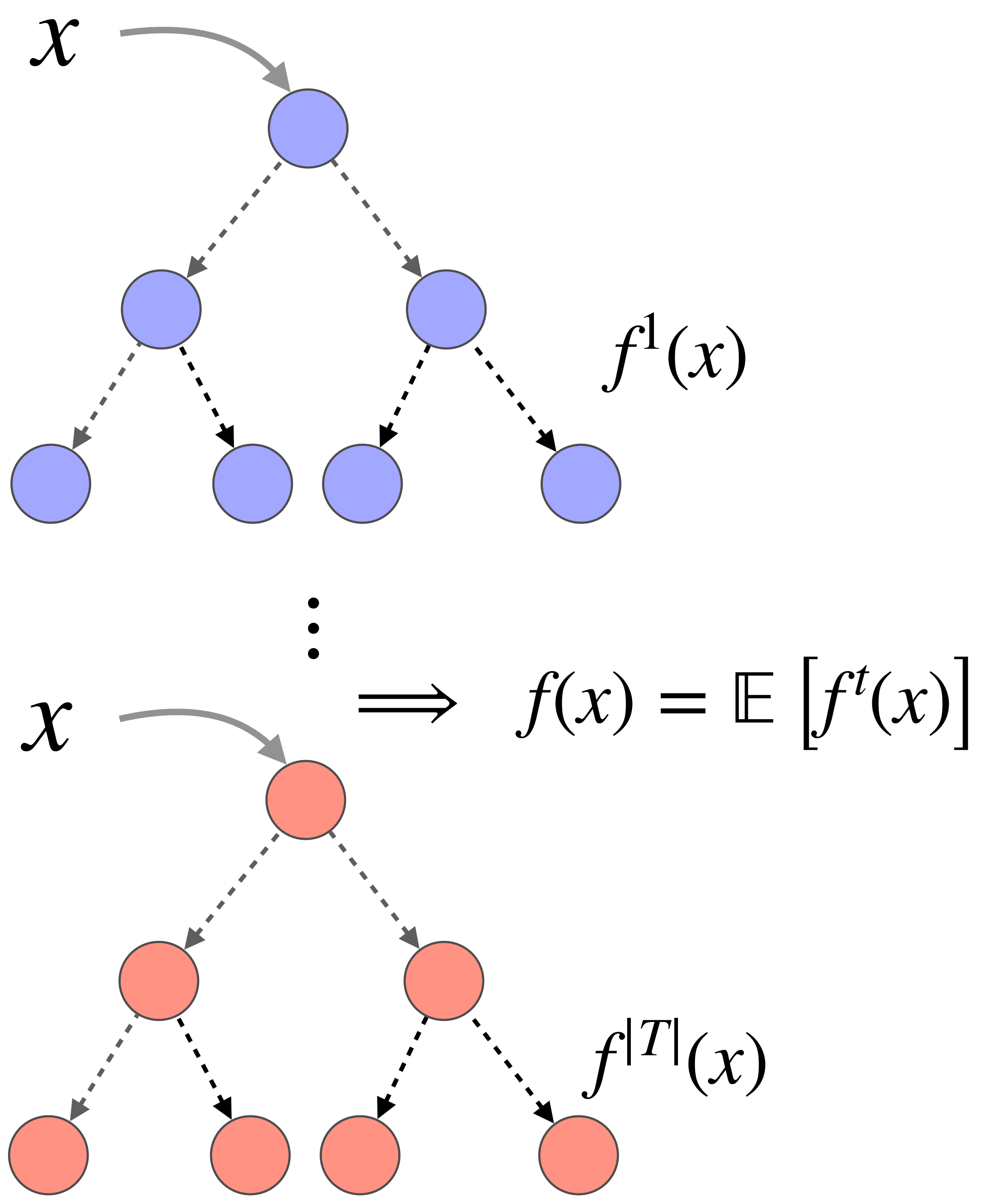
Overview of Aranyani

Online Learning For Group Fairness

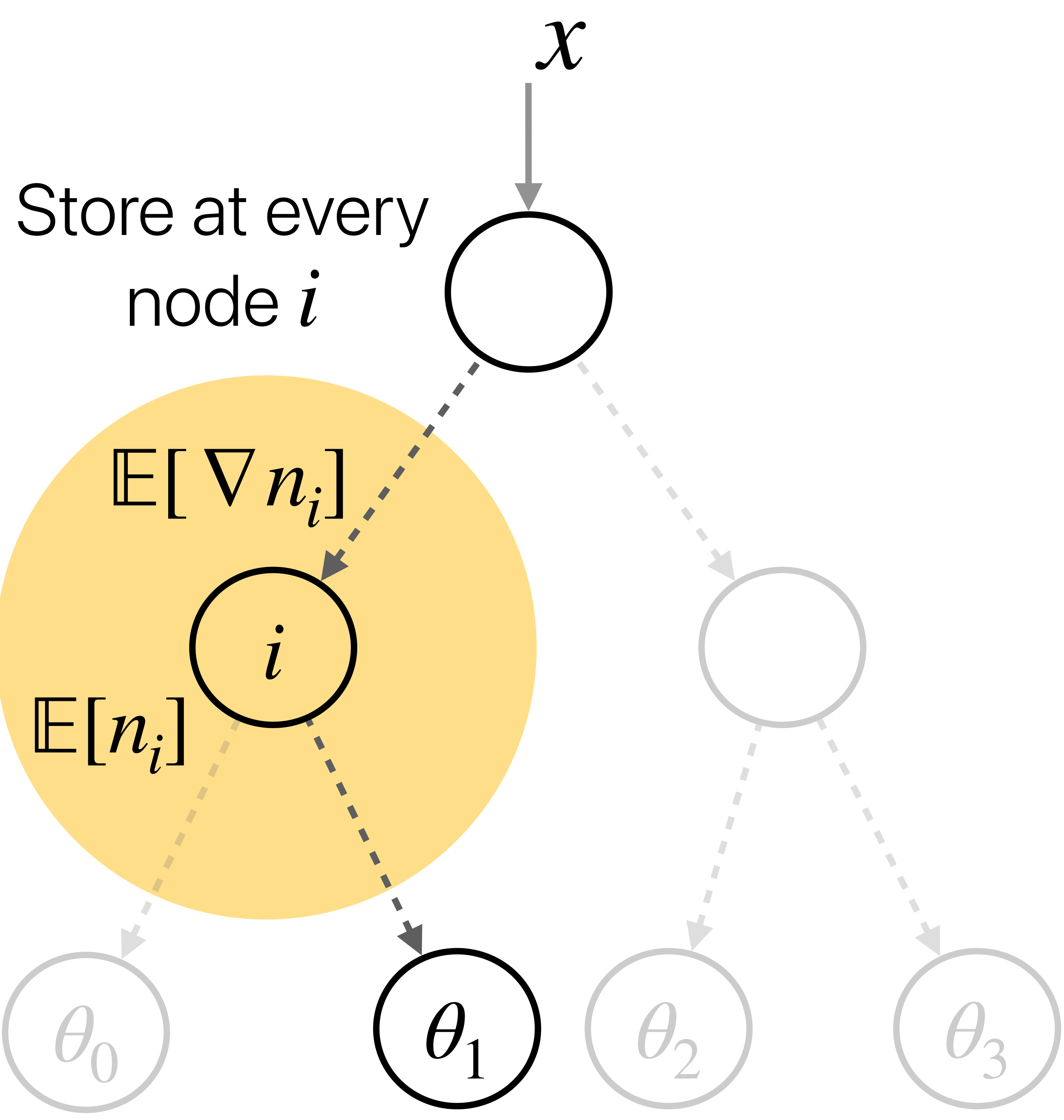


Discrimination $< \epsilon$

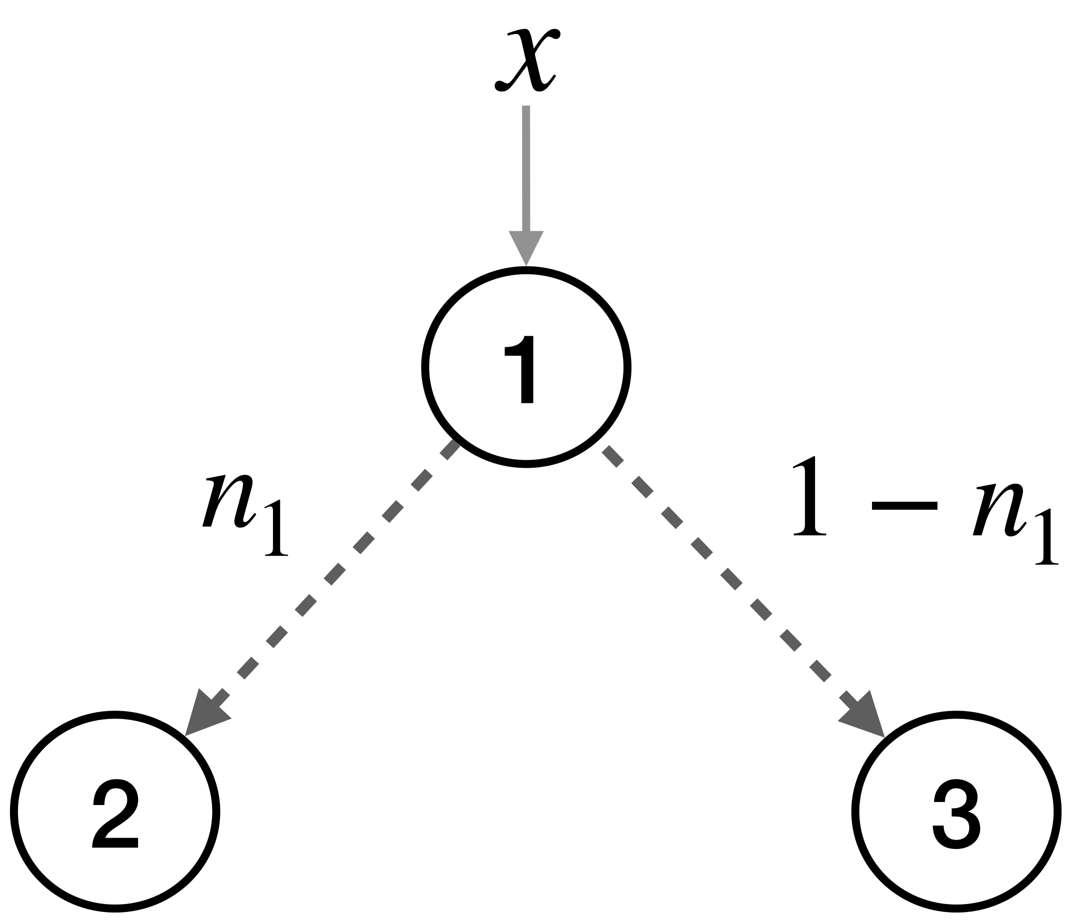
Prediction Using Oblique Decision Forests



Gradient Estimation Using Aggregate Statistics

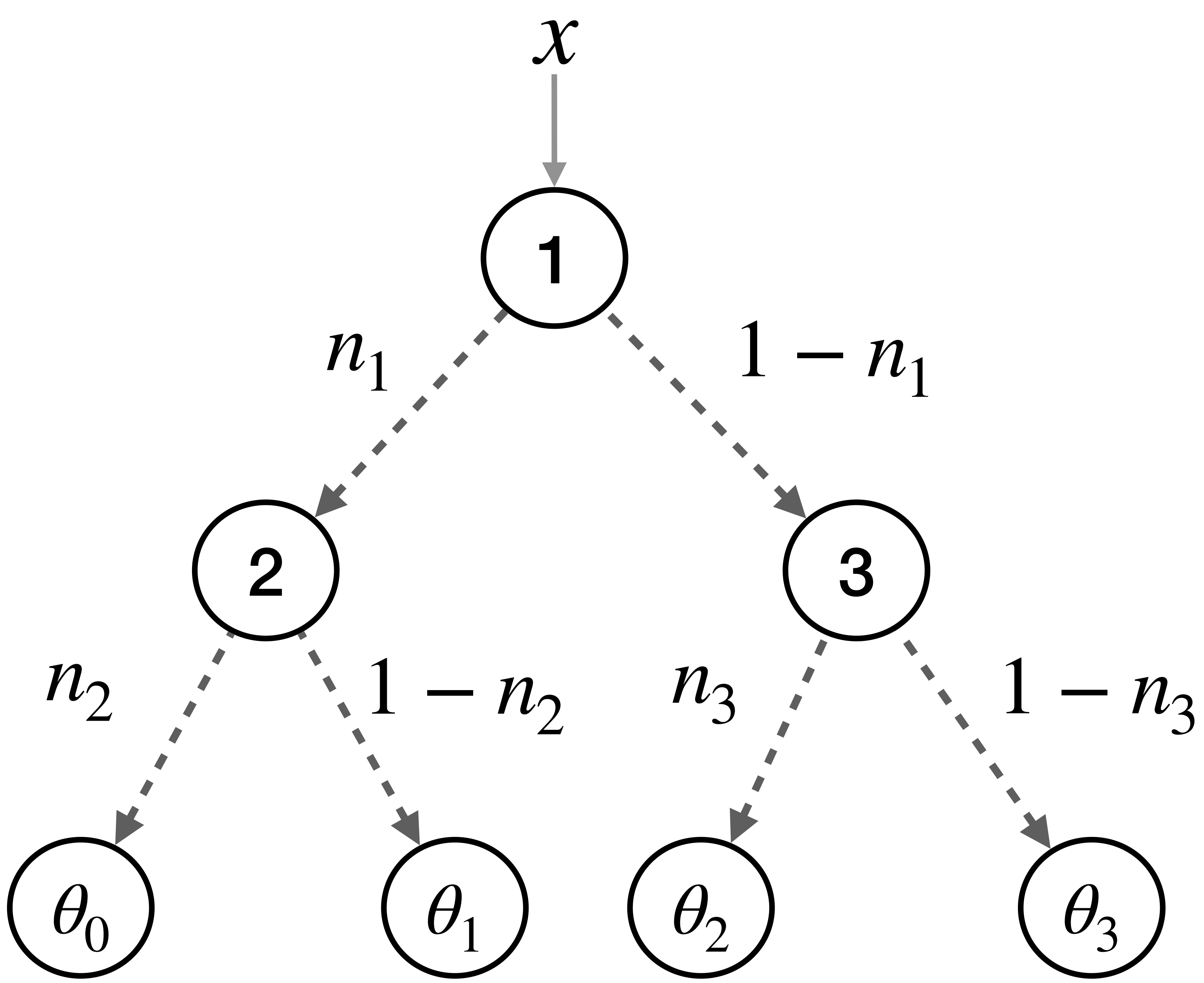


Aranyani

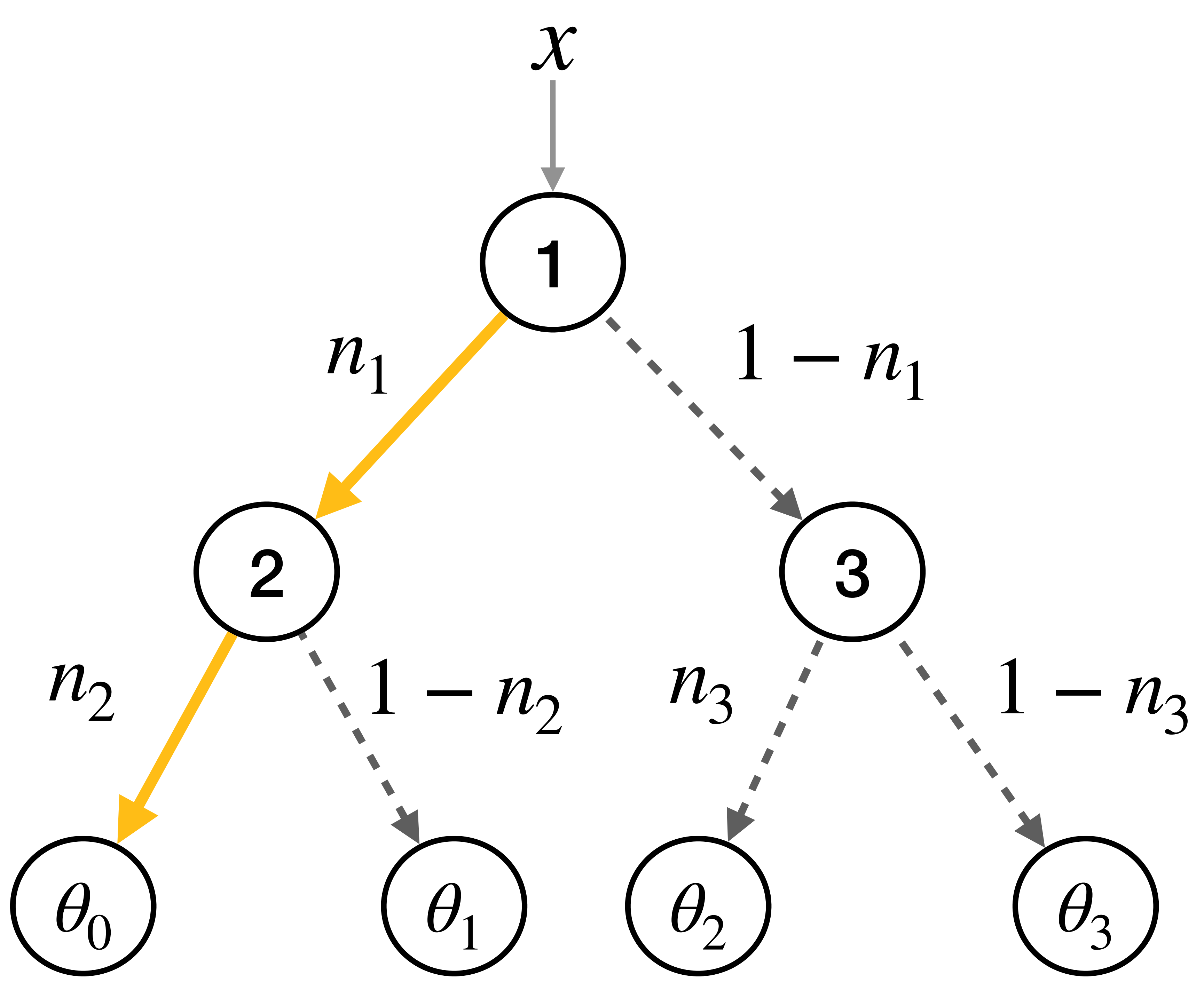


$$n_i = g(\mathbf{w}_i^T x + \mathbf{b}_i)$$

Aranyani

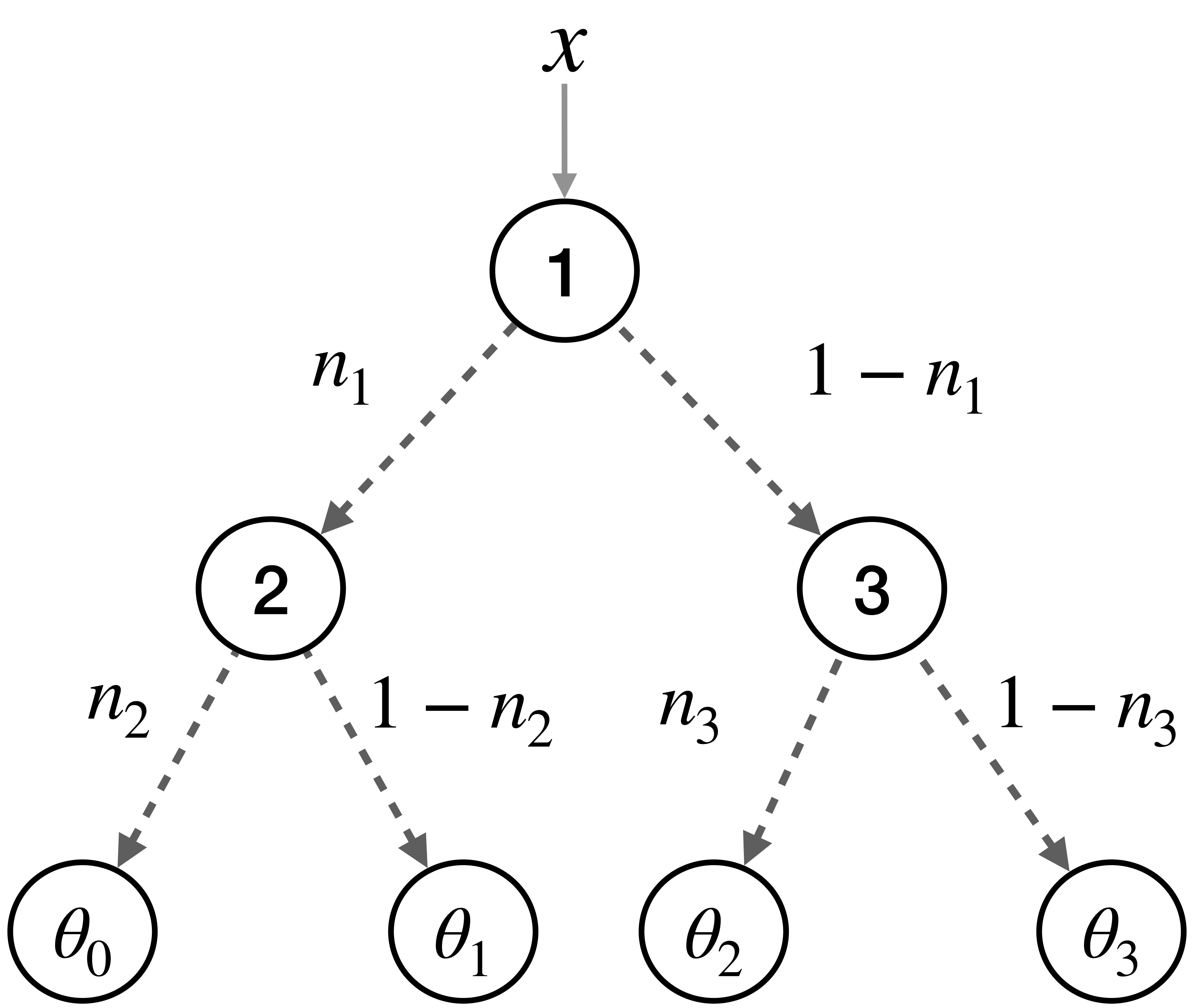


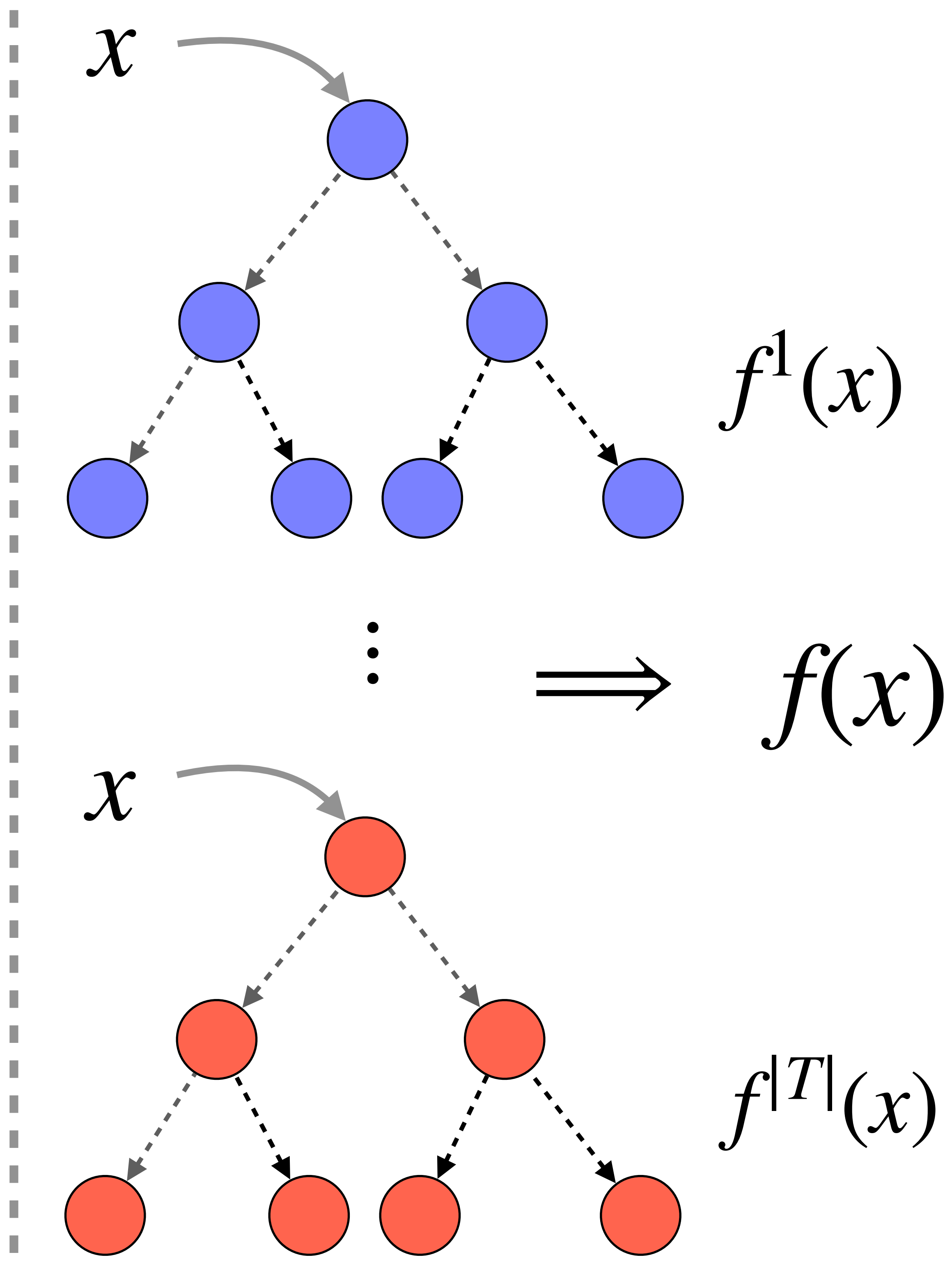
Aranyani



$$f(x) = \boxed{n_1 n_2 \theta_0} + n_1(1 - n_2)\theta_1 + (1 - n_1)n_3\theta_2 + (1 - n_1)(1 - n_3)\theta_3$$

Aranyani





$$\Longrightarrow \quad f(x) = \mathbb{E}\left[f^t(x)\right]$$

Fairness Gradient Estimation

- The fairness gradient estimation process is shown below:

$$G(\Theta) = \nabla_{\Theta} L(f(x), y) + \lambda \sum_{i,j} \nabla_{\Theta} H_{\delta}(F_{ij})$$

Differentiable Huber loss for node-level decisions

Fairness Gradient Estimation

- The fairness gradient estimation process is shown below:

$$G(\Theta) = \nabla_\Theta L(f(x), y) + \lambda \sum_{i,j} \nabla_\Theta H_\delta(F_{ij})$$

$$\nabla_\Theta H_\delta(F_{ij}) = \begin{cases} F_{ij} \nabla_\Theta F_{ij}, & \text{if } |F_{ij}| < \delta \\ \delta \,.\, \text{sgn}(F_{ij} - \delta/2) \nabla_\Theta F_{ij}, & \text{otherwise} \end{cases}$$

Fairness Gradient Estimation

- The fairness gradient estimation process is shown below:

$$G(\Theta) = \nabla_\Theta L(f(x), y) + \lambda \sum_{i,j} \nabla_\Theta H_\delta(F_{ij})$$

$$\nabla_\Theta H_\delta(F_{ij}) = \begin{cases} F_{ij} \nabla_\Theta F_{ij}, & \text{if } |F_{ij}| < \delta \\ \delta \,.\, \mathrm{sgn}(F_{ij} - \delta/2) \nabla_\Theta F_{ij}, & \text{otherwise} \end{cases}$$

$$\nabla_\Theta F_{ij} = \mathbb{E}[\underline{\nabla_\Theta n_{ij}(x \mid a = 0)}] - \mathbb{E}[\underline{\nabla_\Theta n_{ij}(x \mid a = 1)}]$$

These can be estimated using *aggregate statistics* of node gradients alleviating the need for storing samples.

Theoretical Results

- Estimation error of fairness gradients is bounded: $\delta B/2$

δ: Huber constant, B: input bound

Theoretical Results

- Estimation error of fairness gradients is bounded: $\delta B/2$
- The gradient norm Φ_T is bounded by

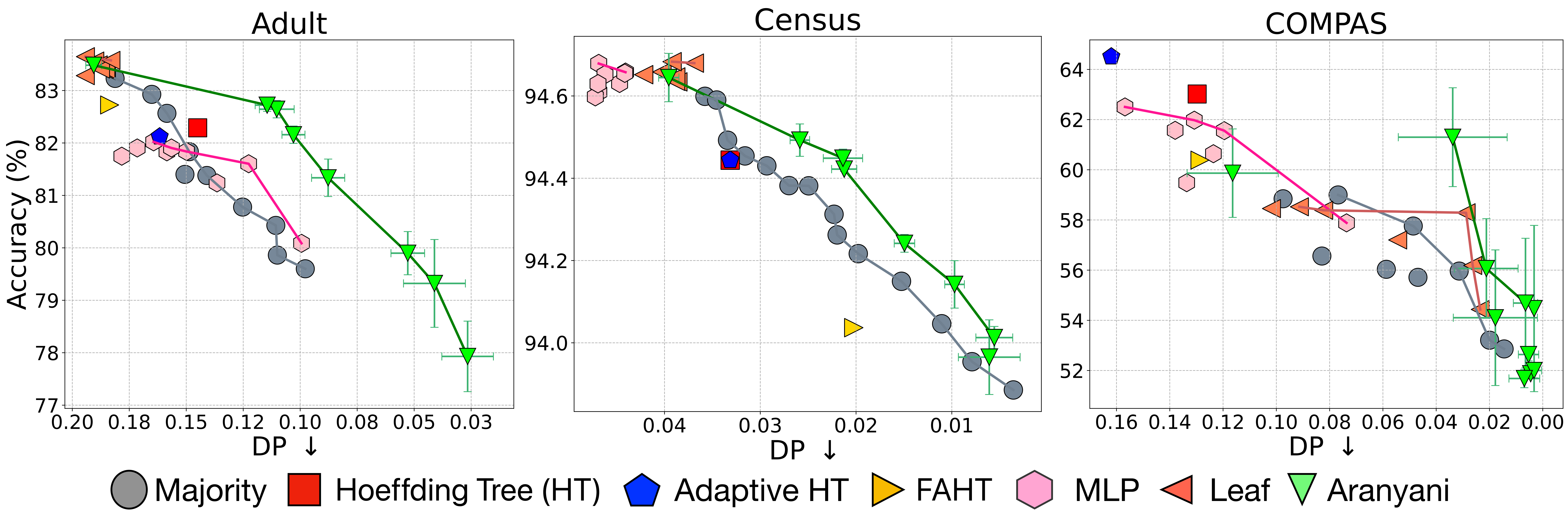
$$\Phi_T \leq \left(\epsilon + 2^{h-2}\lambda^2\delta^2B^2\right)$$

h: tree height, λ: loss hyperparamater

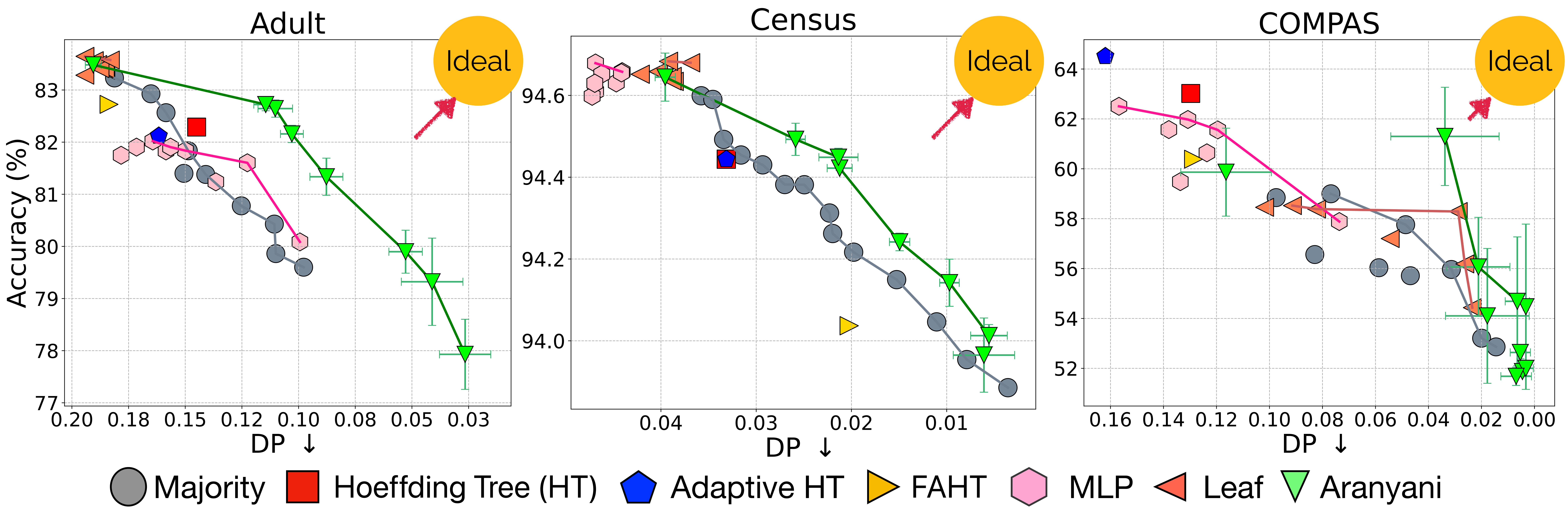
Experiments

- Experiments show effectiveness in *Tabular, Vision, and Language* datasets
- During online learning, at each step we measure the task performance and fairness
- We report the average performances at the final step, T

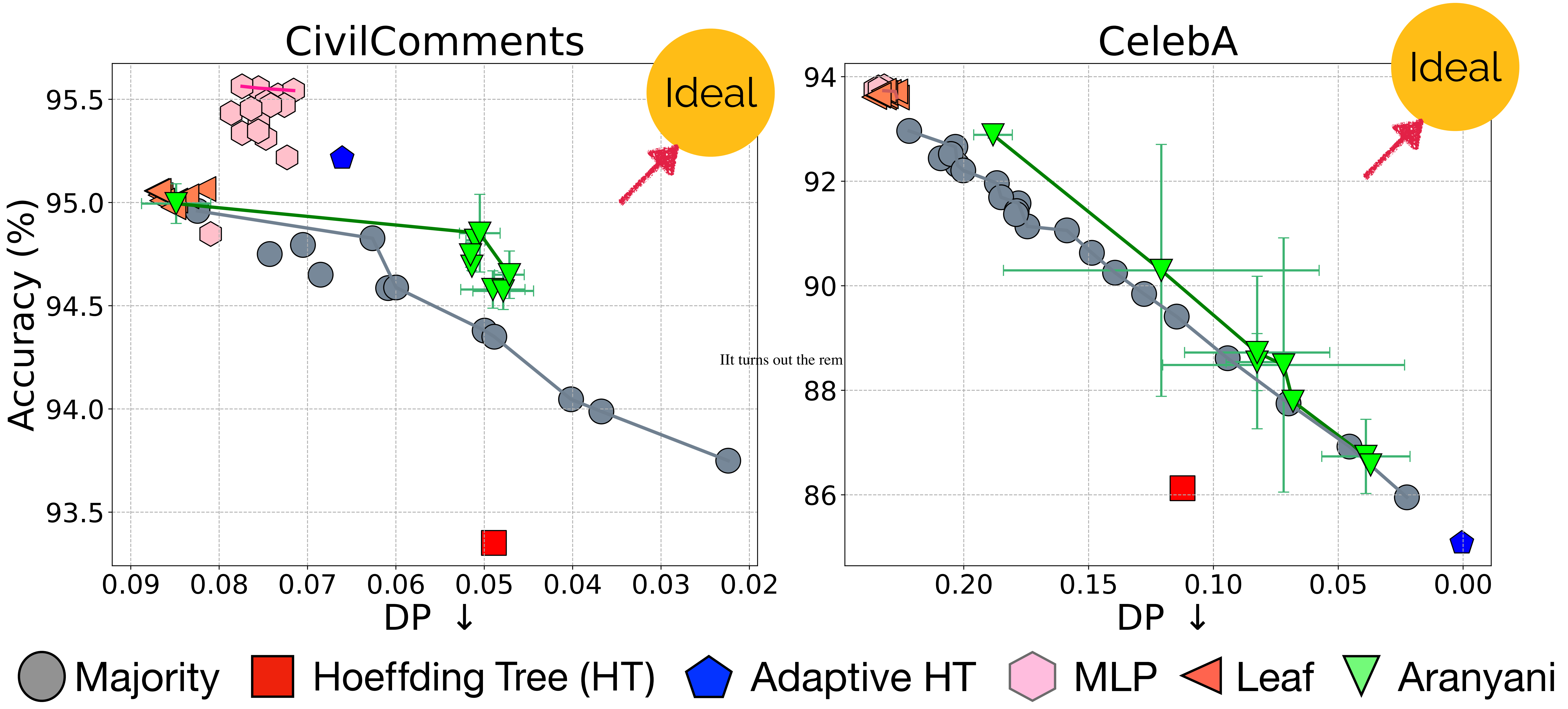
Tabular Datasets



Tabular Datasets



Vision & Language Datasets

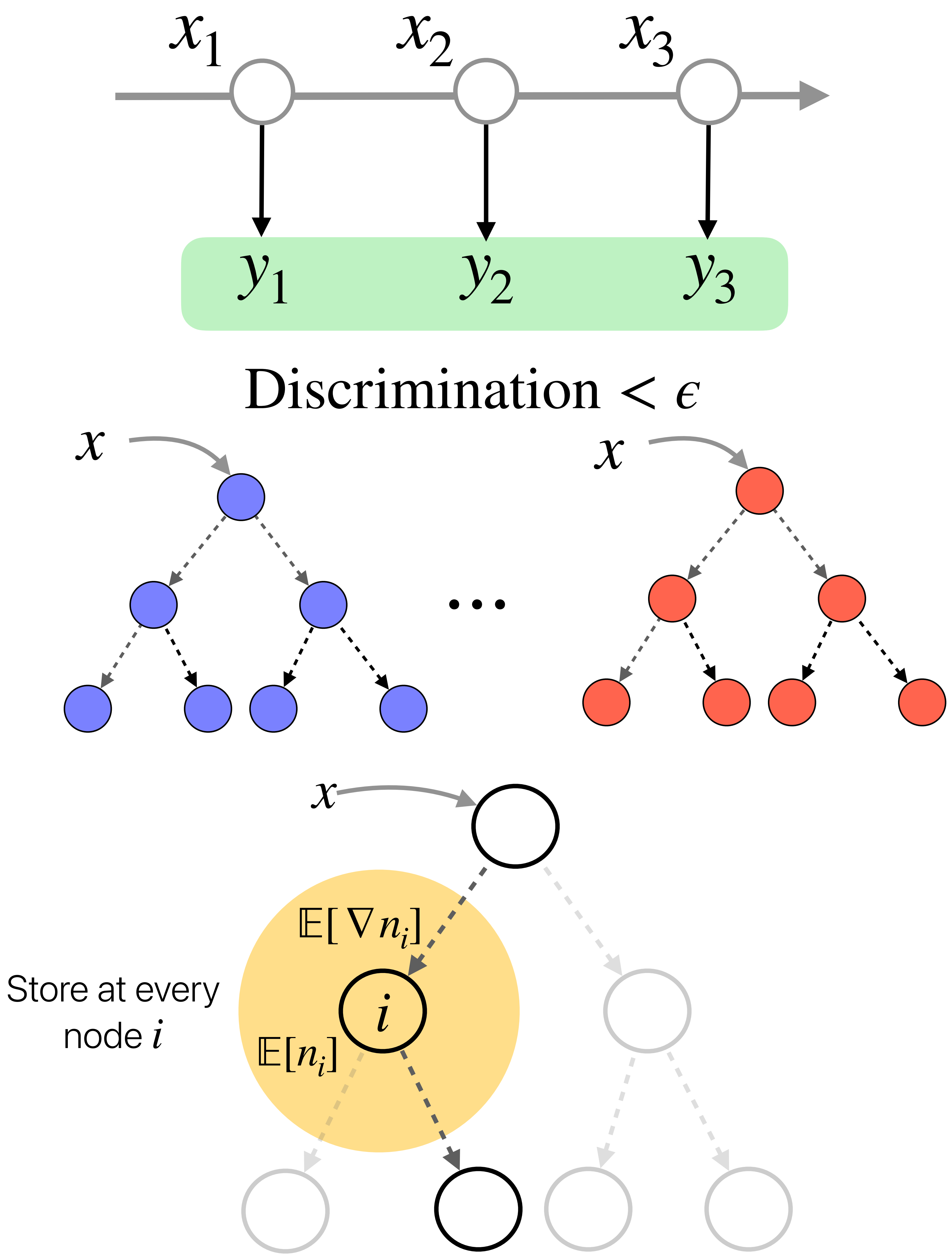


Summary

We propose **Aranyani** to achieve group fairness in online environments

Aranyani leverages oblique decision forests for efficient online gradient computation

Fairness gradient estimation using aggregate statistics achieves impressive performance in real-world scenarios



Thank You!

Contact Info:

Somnath Basu Roy Chowdhury

UNC Chapel Hill

somnath@cs.unc.edu

Paper

Code