Using Noise to Infer Aspects of Simplicity Without Learning

Zachary Boner*, Harry Chen*, Lesia Semenova*, Ronald Parr, Cynthia Rudin

Duke DEPARTMENT *of* COMPUTER SCIENCE

Tabular dataset in high stakes domains

sparse decision trees

Recidivism datasets

Amsterdam	7 leaves
Broward	9 leaves
COMPAS	5 leaves
NIJ	10 leaves

Credit default datasets

Australian	5 leaves
German	4 leaves
GMSC	5 leaves
FICO	5 leaves

These are the number of leaves for trees that achieve black-box accuracy

Noise in data generation increases regularization

Semenova et al. The path to simpler models starts with noise. NeurIPS, 2023

Noise in data generation

increases regularization

How much?

KEY QUESTION:

Given **the noise level ρ,** how much **simpler can ML models** get

as compared to the **non-noisy case**

while still maintaining similar generalization performance?

Noise increases regularization

Optimizing

0-1 loss with regularization penalty λ and random label noise ρ

is equivalent to optimizing

0-1 loss with regularization penalty $\frac{\lambda}{1-2\rho}$ over clean data

Theorems 1, 2

Noise increases regularization

Optimizing for linear models

Exponential loss with additive attribute noise ρ

is equivalent to optimizing for linear models

Exponential loss with l_2 regularization penalty $\frac{1}{2}\boldsymbol{\rho}^2$ over clean data

Theorem 9

Feedback for policymakers and ML practitioners

Our results are **the initial steps**

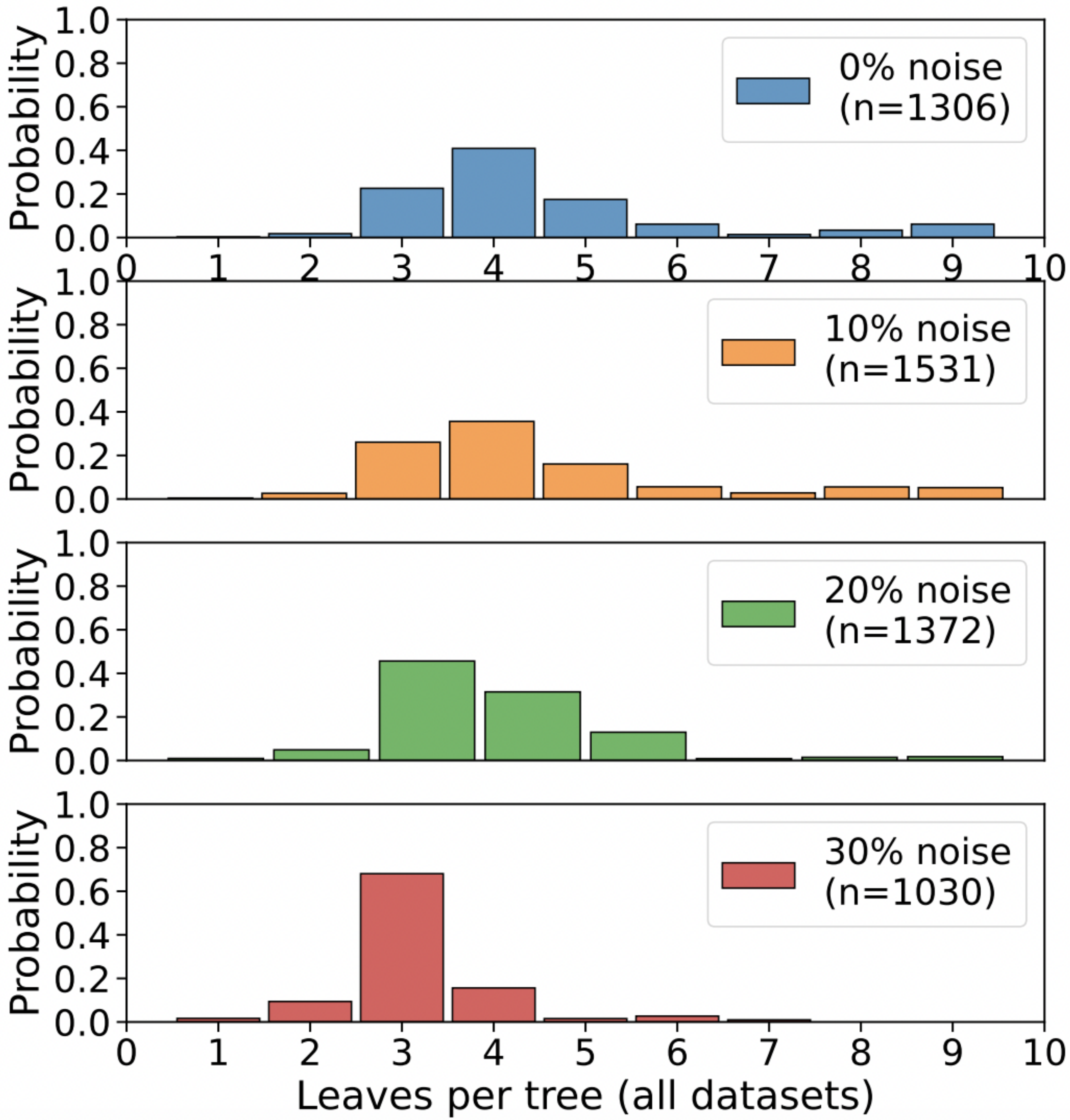
that will help to **reason about the simplicity of models**

encountered for many **high-stakes** decision domains.

Noise in data generation

increases regularization,

simplifies the set of near-optimal models,



Not only optimal,

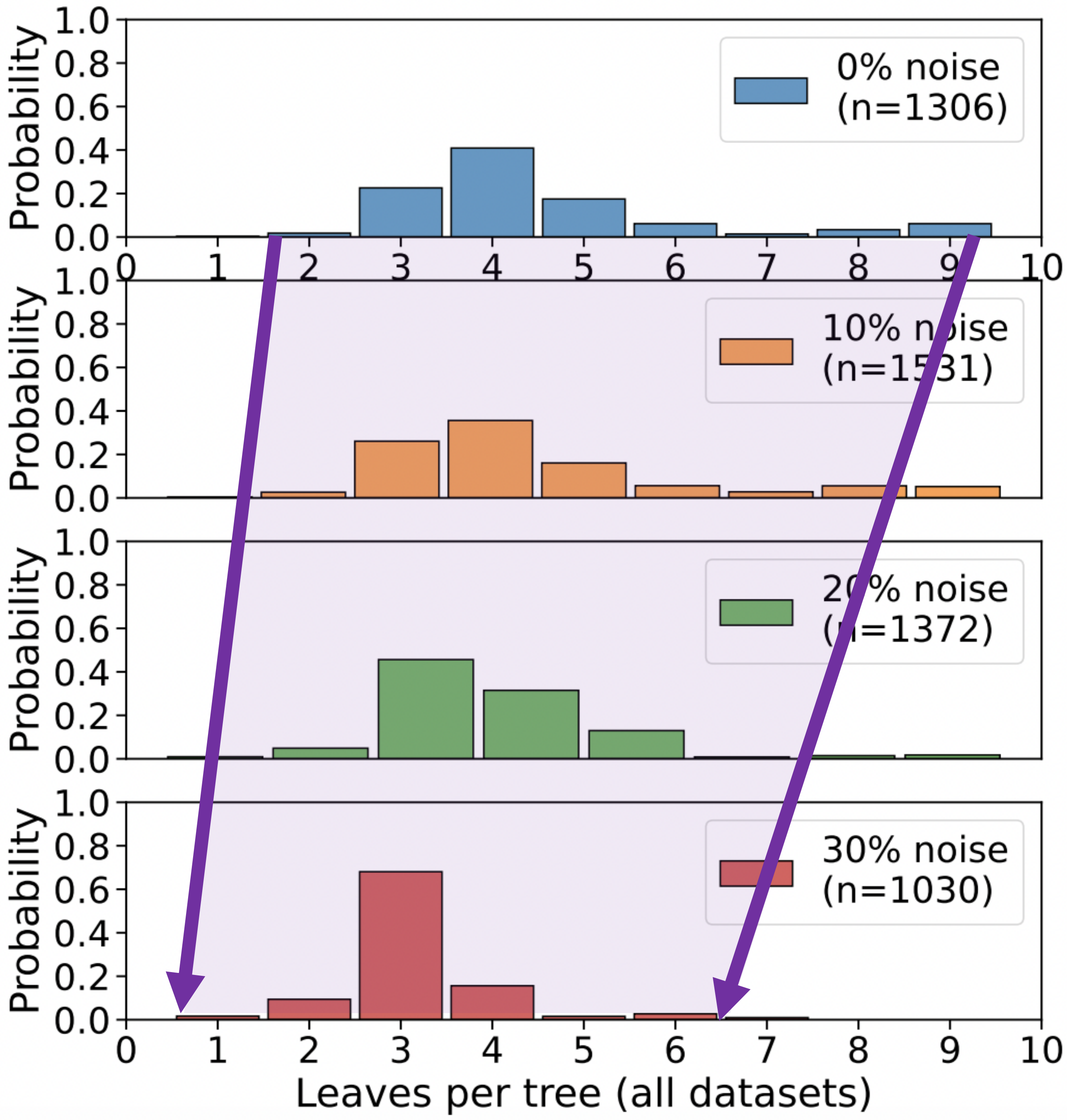
but **all near-optimal models**

(as know as the Rashomon set)

become simpler

under reasonable conditions

Theorems 3, 11



Not only optimal,

but **all near-optimal models**

(as know as the Rashomon set)

become simpler

under reasonable conditions

Theorems 3, 11

Noise in data generation

increases regularization,

simplifies the set of near-optimal models,

increases the set of (relatively) good features

Under the same amount of uniform random label noise ρ,

the expected AUC of features with **higher value**

decreases faster than

the expected AUC of features with **lower value**

Theorem 5, Corollary 6, 7

Noise in data generation

- ❖ increases regularization,
- ❖ simplifies the set of near-optimal models,
- ❖ increases the set of (relatively) good features

Using Noise to Infer Aspects of Simplicity Without Learning

Zachery Boner* **Harry Chen*** **Lesia Semenova*** **Ronald Parr** **Cynthia Rudin**
Department of Computer Science, Duke University
{zachery.boner,harry.chen084,lesia.semenova,ronald.parr,cynthia.rudin}@duke.edu

https://openreview.net/pdf?id=b172ac0R4L