Estimation of Smoothed Entropy

Paul Cuff, Peter Park, Yucel Altug, Lanqing Yu
(Princeton University)

Estimation of Smoothed Support

Paul Cuff, Peter Park, Yucel Altug, Lanqing Yu
(Princeton University)

Problem

- Take n samples from an unknown distribution (i.id.)
- Estimate the entropy
- Estimate the support

Many Incarnations

- Shakespeare's vocabulary

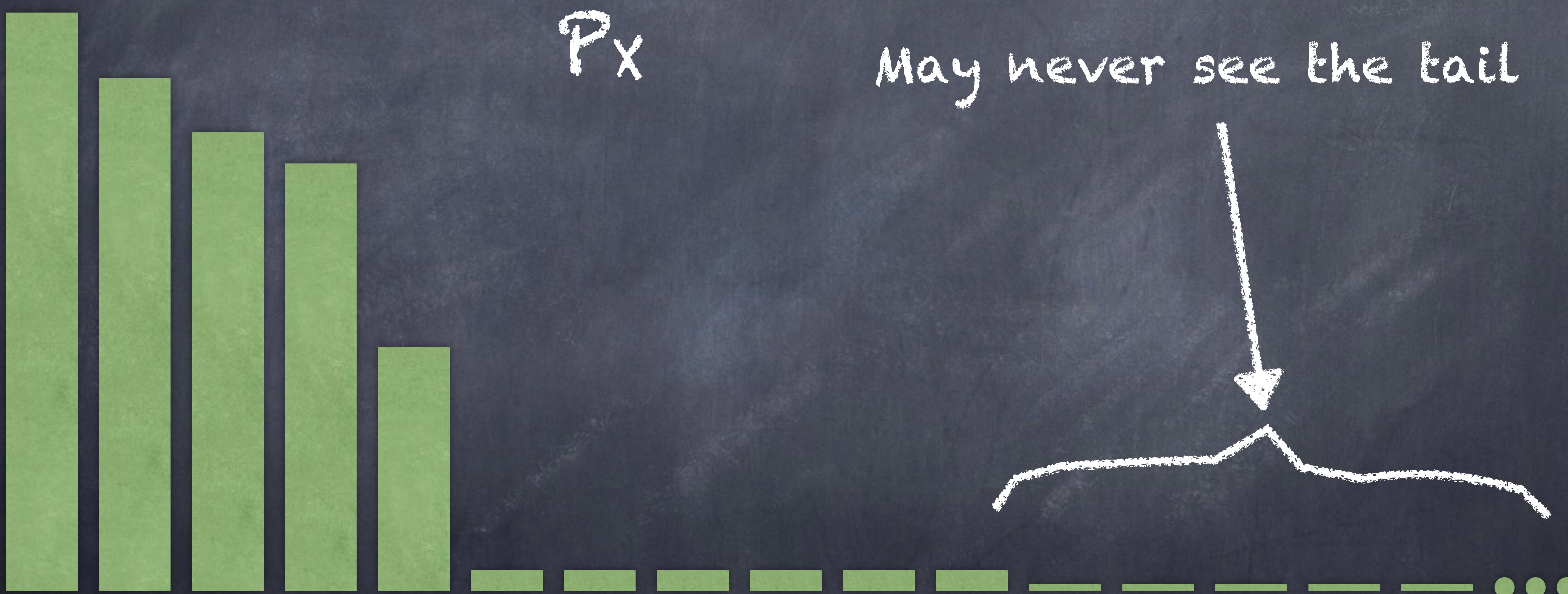
- How many species?

- Good-Turing estimator

Long History

- Recent:
  - [Valiant-Valiant 10]
  - [Acharya-Jafarpour-Orlitsky-Suresh-Wu 13, 15]
  - [Jiao-Venkat-Han-Weissman 15]

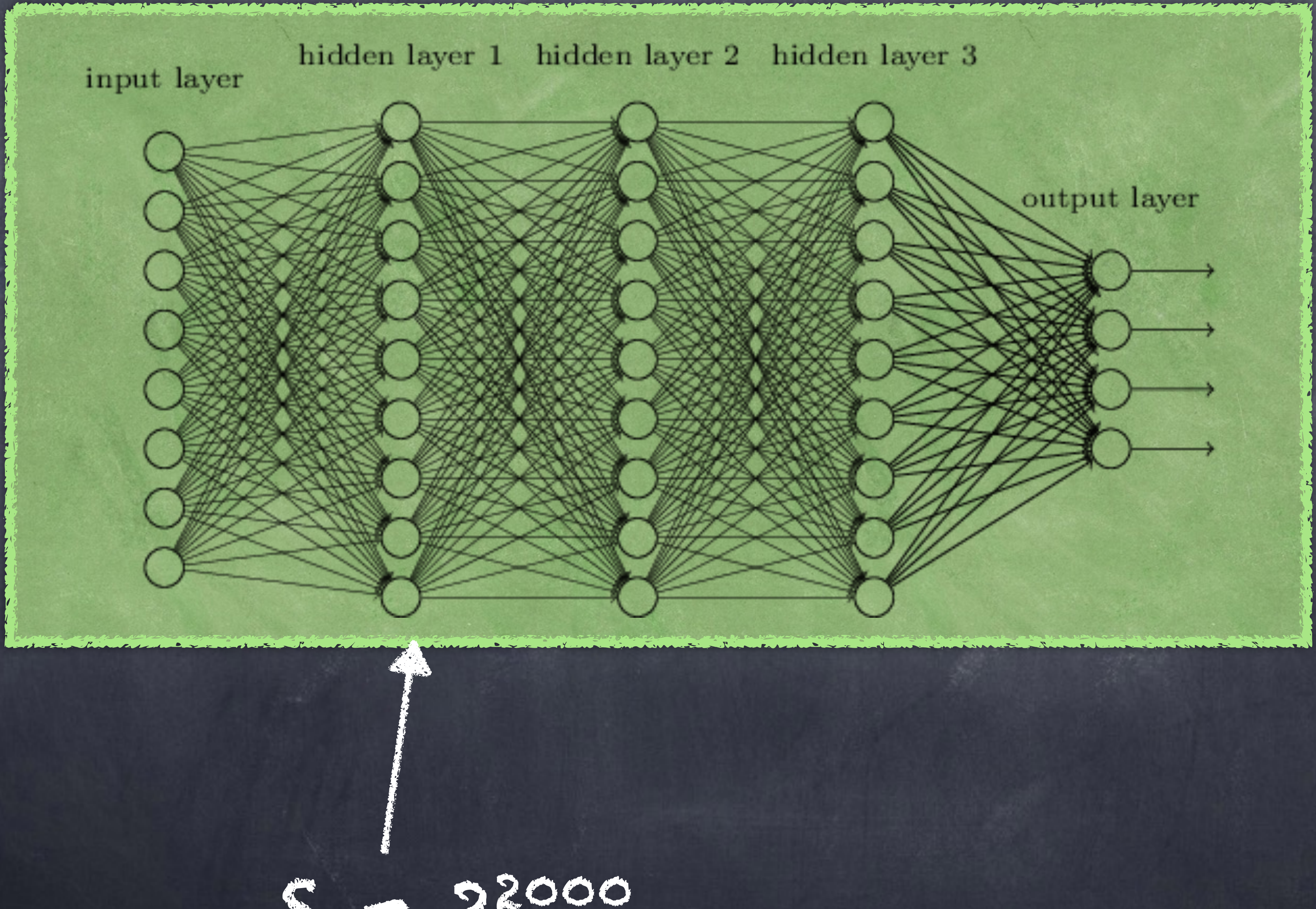
The problem



The usual assumption

- Recent work
  - Entropy: assume a bound on the support size (S)
  - Support: assume a minimum probability mass (1/S)
  - Sample complexity: $n \sim \dfrac{S}{\log S}$

Death by S



$S = 2^{2000}$

What can we do with no assumption?

Perhaps nothing

- Cannot reliably decide that entropy or support is finite.
- Reason: Every distribution has an $H=\infty$ neighbor (in total variation)

Yikes

- After one million samples of seeing only one outcome, can we not say anything?

Two Changes

1. Estimate smoothed entropy/support

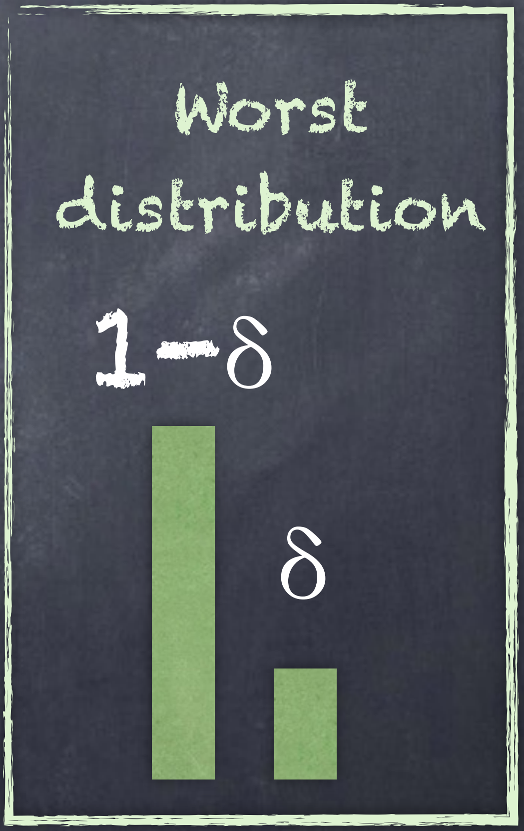
$$S_\delta(P_X) = \min_{Q : \|P_X - Q\|_{TV} \leq \delta} |\text{Support}(Q)|$$

2. Confidence bounds: Estimator can fail as long as it knows when it fails

$$\underline{S}_\delta(X^n) \quad (\quad S_\delta \quad) \quad \overline{S}_\delta(X^n)$$

All Samples the Same

- Conclude: H=0, Support=1
- Error prob. $< \varepsilon$ if $n \geq \frac{\log \frac{2}{\epsilon}}{\log \frac{1}{1-\delta}}$
- 459 samples (for $\delta=\varepsilon=0.01$)



All Samples Different

- No upper bound possible
- Lower bound: Support = $\Omega(n^2)$

ε-Achieving

$$\sup_P \mathbb{P}\left(S_\delta(P) \notin \left[\underline{S}_\delta(X^n), \overline{S}_\delta(X^n)\right]\right) \leq \epsilon$$

Simple estimator

- Build estimator based on a simple statistic:
  - R = fraction of unique samples

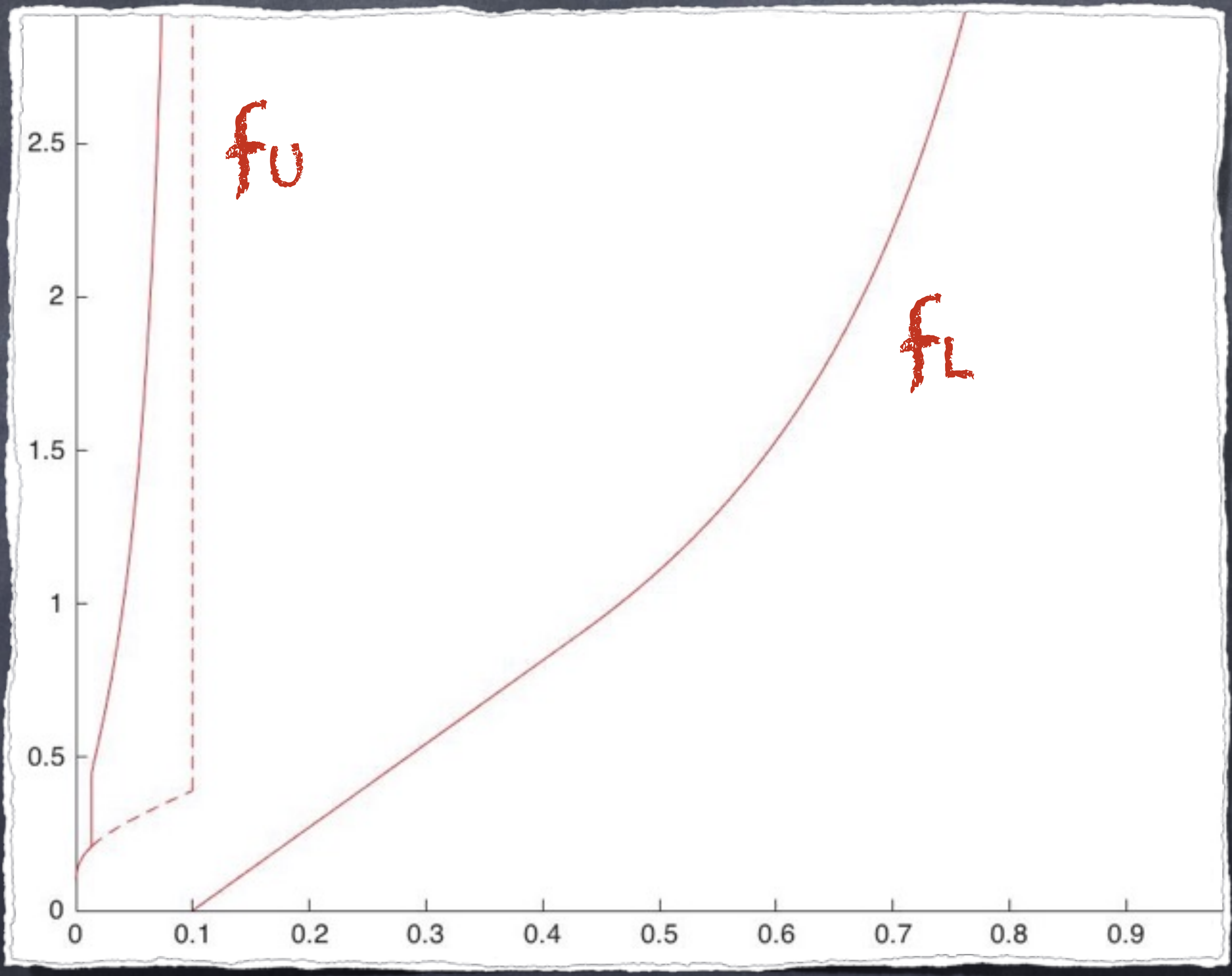
Claim

- Choose c>3:
- ε-achieving (for large enough n):

$$\underline{S}_\delta(R) = n f_L\left(R + c\sqrt{\frac{\log n}{n}}\right) \qquad \overline{S}_\delta(R) = n f_U\left(R - c\sqrt{\frac{\log n}{n}}\right)$$

$$f_L(r) = \begin{cases} 0 & r \le \delta \\ e(r-\delta) & \delta < r < \delta + e^{-1}(1-\delta) \\ \frac{1-\delta}{\log \frac{1-\delta}{r-\delta}} & r \ge \delta + e^{-1}(1-\delta) \end{cases} \qquad f_U(r) = \begin{cases} \frac{1-\delta}{\log \frac{\delta}{r}} & r < \delta \\ \infty & r \ge \delta \end{cases}$$

$\delta=0.1$



$\frac{S_\delta}{n}$

R

Best bounds

- Small R: $$\overline{S}_\delta = O\left(\frac{n}{\log n}\right)$$
- Large R: $$\overline{S}_\delta = \Omega\left(n^{3/2}\right)$$

↑ Maybe n^2

Proof - 2 Steps

1. Connect to Poisson Approximation
2. Analyze Poisson Approximation

Step 1

Poisson approximation — Non-discrete part — Bernstein — Discrete Tail

$$\mathbb{P}\left(|R - \mathbb{E}_{X^N} R| > 3\Delta\right) < e\sqrt{n}\left(\exp\left(-\frac{n\Delta^2}{2(1+\Delta)}\right) + \exp\left(-\frac{n\Delta^2}{2}\right) + \exp\left(-\frac{n\Delta^2}{2(1+\Delta/3)}\right) + \frac{1}{n}\right)$$

Plug in $\Delta = \frac{c}{3}\sqrt{\frac{\log n}{n}}$

Step 2

- Define fingerprint: $X \sim P_X$

  $Y = P_X(X) = e^{-\imath_X(X)}$

- P_Y is fingerprint of P_X

$$S_\delta(P_X) = \mathbb{E}\,\frac{1}{Y}1\{Y > \mathbb{F}_Y(\delta)\}$$

$$\mathbb{E}_{X^N} R = \mathbb{E}\,e^{-nY}$$

Sample Complexity

- Choose c > e
- If n > cS/δ and n large enough:

$$\overline{S}_\delta \leq n(1-\delta)/2$$

Bottom Line

- With 11 million samples, start to have guarantees
- With 100 million samples, guarantee for S < 1,797,000