
# Beating Natural Distribution Estimation with Side Information*

EE Research Symposium

**Haricharan Balasundaram**, Prof. Andrew Thangaraj

*Partly presented at Allerton 2025 and partly submitted to ISIT 2026

# Dickens' simulator

- Machine that simulates Charles Dickens' writing
- Can be trained on previous Dickens books (data)

**How to get a 'Good' Simulator?**

# Next Word Prediction Problem

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of _____

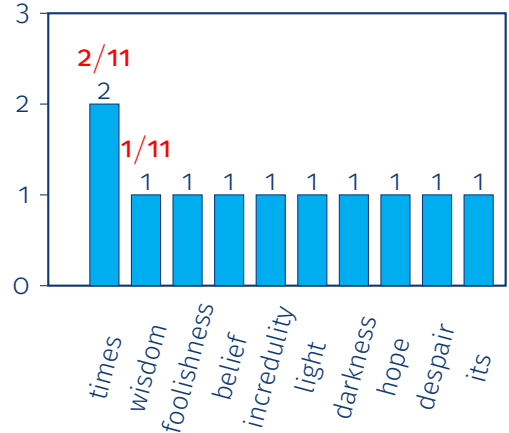
**Given the previous word, what is the Probability of the next word being something?**

| Words occurring after 'of' | Probabilities |
|---|---|
| Surprise | 0.3 |
| Incredulity | 0.2 |
| Amazement | 0.05 |

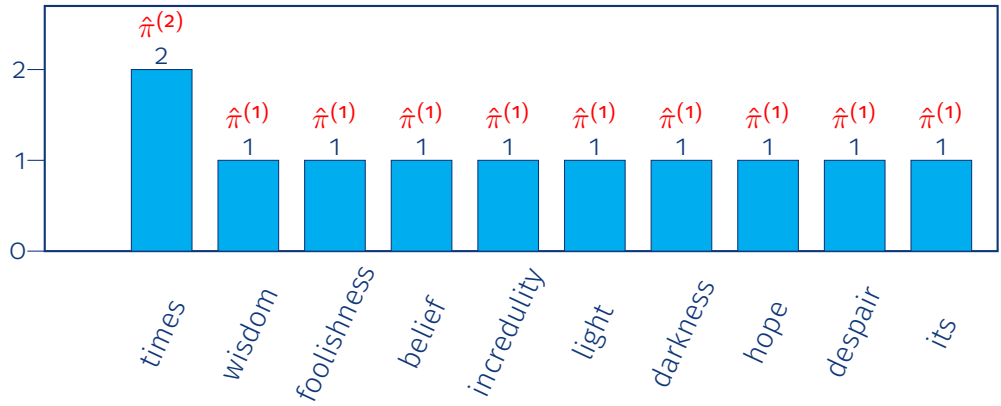
**Table 1:** Example estimate for the Word occurring after 'of': $\hat{\pi}(\cdot|\text{of})$

# The Empirical Estimator for $\pi(\cdot|\text{'of'})$

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair …that some of its noisiest authorities insisted on its being received, for good or for evil …



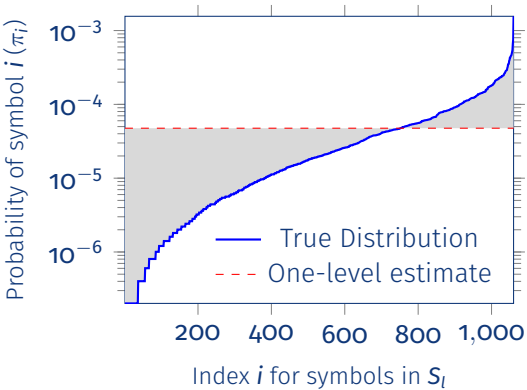
# What is a Natural Estimator?




**Natural Estimators (Empirical, smoothed, profile-based, etc.)**

Letters occurring $l$ times ($S_l$) assigned the same estimator probability mass $\hat{\pi}^{(l)}$

# Why do Natural Estimators Have Error?

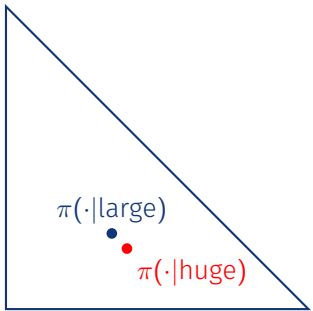



Error in estimation of unigram probabilities for 10000 contiguous words taken from TOTC (Total 100,000 words!)

- Over species occurring $l$ times, we assign a single probability mass $S_l$, while the real probabilities could be anything!
- This error is a random variable that depends on $S_l$
- We estimate this error!

# An example of side information

TOTC: 'huge' occurs once, 'large' occurs 44 times
**Huge and Large are synonyms!**




- Probability distribution of words after 'Huge' ($\pi(\cdot|\text{huge})$) and words after 'Large' ($\pi(\cdot|\text{large})$)
- The closeness can be used to estimate $\pi(\cdot|\text{huge})$ accurately
- Estimator:

$$\alpha \cdot \hat{\pi}(\cdot|\text{huge}) + (1 - \alpha) \cdot \hat{\pi}(\cdot|\text{large})$$

# Our Results

# The Setting

Discrete distribution $\pi = (\pi_1, \pi_2, \ldots, \pi_d)$ $X^n$: $n$ i.i.d. samples drawn from $\pi$

**Distribution Estimation**

Devise an estimate $\hat{\pi}(X^n)$ to minimize the squared $\ell_2$–norm error between $\hat{\pi}(X^n)$ and $\pi$.

$$R(\pi, \hat{\pi}) = E[(\hat{\pi}(X^n) - \pi)^2]$$

# Our Results (Informal)[1]

$$\underbrace{\sum_{i \in S_l} (\pi_i - \hat{\pi}^{(l)})^2}_{\text{Error over } S_l} = \underbrace{\sum_{i \in S_l} (\pi_i - \pi^{(\text{or})})^2}_{\text{Unavoidable (Oracle) Error}} + \underbrace{\frac{(M_l - \hat{M}_l)^2}{\phi_l}}_{\text{Estimator Error}}$$

$$\phi_l = |S_l| \qquad \pi^{(\text{or})} = \frac{M_l}{\phi_l}$$

We provide an estimator $T_l$ for the unavoidable oracle error $Q_l$ based on the **Good-Turing estimator**:

- Show low bias
- Show error probability at most $\delta$ with (i) $n \geq O\left(d^2/\delta\right)^{\frac{1}{3}}$ samples in case of $l = 0$, and (ii) $n \geq O\left(1/\delta\right)$ samples for $l > 0$

[1] Estimating Error in Natural Distribution Estimation. H. Balasundaram, A. Thangaraj. Allerton 2025.

# Error in the Estimator for various distributions

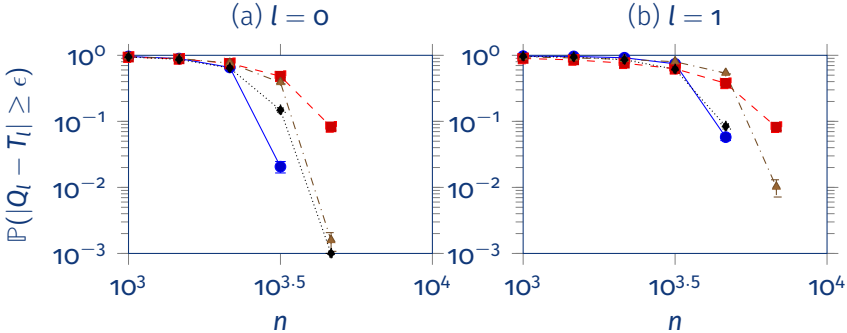
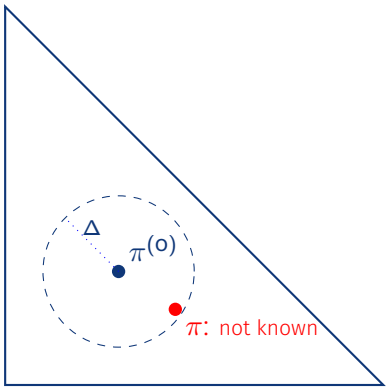



**Figure 1:** $\mathbb{P}(|Q_l - T_l| \geq \epsilon)$ vs $n$ for various distributions, with $d = 1000$ and $\epsilon = 0.001$. uniform, zipf, uniform mixture, and random distributions.

# Local Min-max Problem Formulation



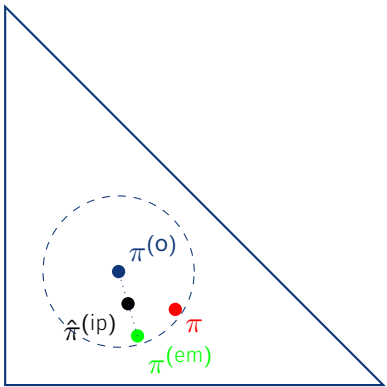

- We know $\pi^{(o)}$ such that

$$\|\pi - \pi^{(o)}\|_2 \leq \Delta.$$

- Estimator $\hat{\pi}$ is a function of $(X^n, \pi^{(o)}, \Delta)$

How to minimize the squared $\ell_2$–error?

# Interpolation Estimator




$$\hat{\pi}^{(ip)} = \alpha\pi^{(em)} + (1-\alpha)\pi^{(o)}$$

**Theorem: $\hat{\pi}^{(ip)}$ Error**

Worst-case Risk:

$$\leq \min\left(\Delta^2, \frac{1-(\|\pi^{(o)}\|-\Delta)^2}{n}\right).$$

# Summary of Risk Lower Bounds[2]

| Setting | Upper Bound | Lower Bound |
|---|---|---|
| General $\pi^{(0)}$ | $O\left(\min\left(\Delta^2, \frac{1-(\|\pi^{(0)}\|-\Delta)^2}{n}\right)\right)$ | $\Omega\left(\frac{(1-\|\pi^{(0)}\|^2)}{d} \cdot \min\left(\Delta^2, \frac{1}{n}\right)\right)$ LeCam's Method |
| Deterministic $\pi^{(0)} = (1, 0, \ldots, 0)$ | $O\left(\min(\Delta^2, \frac{\Delta}{n})\right)$ | $\Omega\left(\min(\Delta^2, \frac{\Delta}{n})\right)$ LeCam's Method |
| Uniform $\pi^{(0)} = (\frac{1}{d}, \frac{1}{d}, \ldots, \frac{1}{d})$ | $O\left(\min(\Delta^2, \frac{1}{n})\right)$ | $\Omega\left(\min(\Delta^2, \frac{1}{n})\right)$ Assouad's Lemma |

[2] Distribution Estimation with Side Information. H. Balasundaram, A. Thangaraj. Submitted to ISIT.

# Simulations

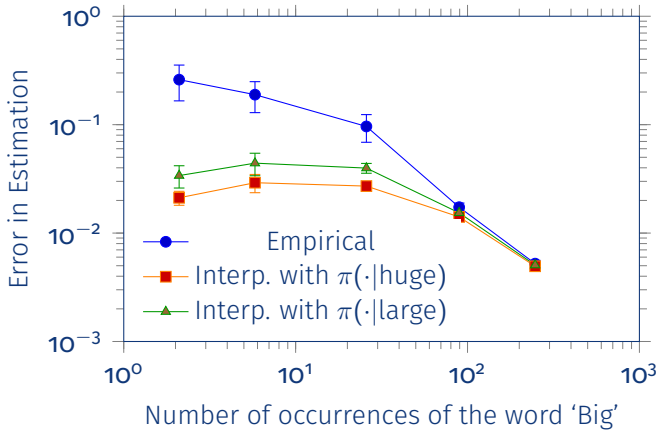



**Figure 2:** Estimation errors vs. number of samples for the Empirical and Interpolation Estimators for $\pi^{(0)}$ from the dataset.

# Questions?