Estimating stationary mass, frequency by frequency

Milind Nakul [*] Vidya Muthukumar [†,*] Ashwin Pananjady [*,†]

[*] Industrial & Systems Engineering, Georgia Institute of Technology
[†] Electrical & Computer Engineering, Georgia Institute of Technology

Problem of interest

Distribution Estimation:

- **Input:** $X^n := (X_1, \ldots, X_n)$, a stochastic ergodic process defined over a finite state space $\mathcal{X}$.
- Estimate the unique stationary distribution denoted by π in a frequency by frequency sense.
- For each $\zeta = 0, 1, \ldots, n$, we define

$$M_\zeta^\pi(X^n) := \sum_{x \in \mathcal{X}} \pi_x \cdot \mathbb{I}\{N_x = \zeta\}.$$

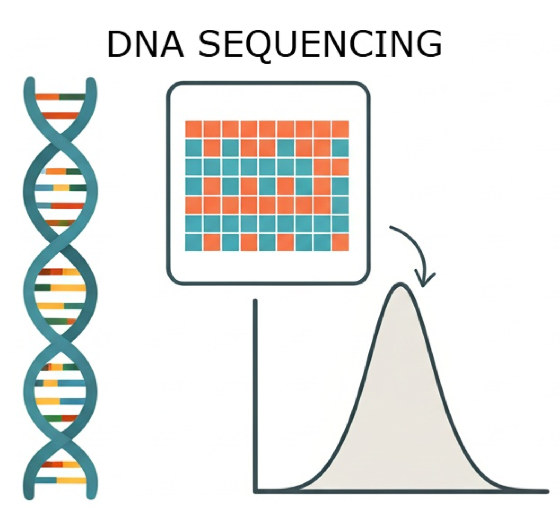
Example: $\mathcal{X} = \{a, b, c, d\}$ with $\pi_a = 0.2, \pi_b = 0.5, \pi_c = 0.2, \pi_d = 0.1$. Observed sequence $X^n = \{a, a, b, b, b, c, c\}$. Thus, vector of interest

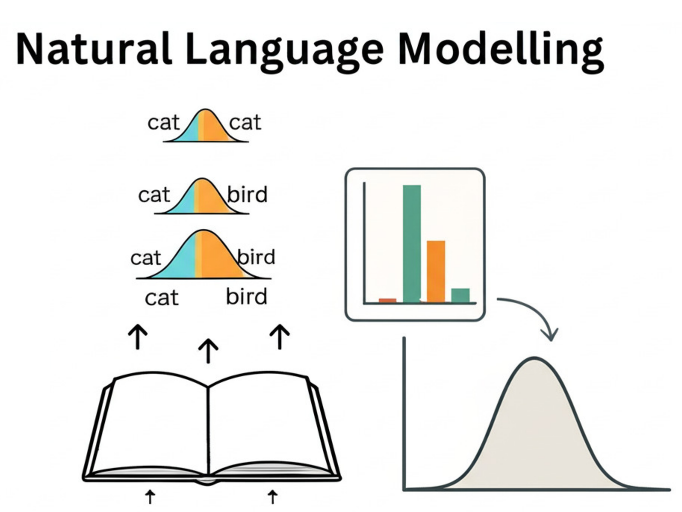
$$M^\pi(X^n) = (0.1, 0.0, 0.4, 0.5, 0.0, 0.0, 0.0).$$

- **Goal**: Design $\widehat{M} : \mathcal{X}^n \to \Delta(\{0, 1, \ldots, n\})$ such that we minimize the risk:

$$\mathcal{R}_n(M^\pi, \widehat{M}) := \mathbb{E}\left[d_{\mathsf{TV}}(M^\pi(X^n), \widehat{M}(X^n))\right]. \tag{1}$$

- **Applications:** Genomics and natural language modelling where data has temporal dependencies.





IID Estimators

- When the samples X^n are i.i.d., there is a consistent estimator for π in the existing literature.
- **Plug-in** (PI) estimator, assigns the following mass:

$$\widehat{M}_{\mathsf{PI},\zeta} = \frac{\zeta \varphi_\zeta}{n},$$

$\varphi_\zeta := \#$ of symbols appearing ζ times. However, it underestimates missing mass.
- [2] proposed the **Good–Turing** (GT) estimator for missing mass,

$$\widehat{M}_{\mathsf{GT},0} = \frac{\varphi_1(X^n)}{n}.$$

- GT estimates the missing mass by the mass of samples appearing once in X^n.
- **Consistent estimation:** GT estimator for small frequencies and PI for large frequencies.
- **Convergence:** GT+PI converges at $\mathcal{O}(n^{-1/6})$ in the i.i.d. setting [3].

Challenge with correlated samples

- **Bias:** [4] established that using GT estimator on Markovian samples suffers from a non-vanishing bias.
- **Poissonization** is no longer valid in the correlated setting.
- Different algorithmic tools are required for addressing temporal dependence.

Mixing assumption

- For an ergodic stochastic process $\{X_t\}_{t \geq 1}$, the α-mixing coefficient is given by

$$\alpha(\tau) := \sup_{t \in \mathbb{N}} \sup \left\{ |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| : A \in \sigma(X_t^-), B \in \sigma(X_{t+\tau}^+) \right\}.$$

- We assume that X^n satisfies *exponential α-mixing*, i.e., there exist constants $\mu > 0$ and $\rho \in (0, 1)$ such that

$$\alpha(\tau) \leq \mu \rho^\tau \quad \text{for all } \tau \geq 1. \tag{2}$$

- Define the mixing time of such a process to level $\epsilon \in (0, 1)$ as

$$\mathsf{t}_{\mathsf{mix}}(\epsilon) := \min\{\tau \in \mathbb{N} : \alpha(\tau) \leq \epsilon\}.$$

Goal of the work

Design a universally consistent estimator for the stationary distribution of any exponentially α-mixing sequence for any alphabet size $|\mathcal{X}|$.

The Estimator

- We use the **windowed** version of the GT estimator [1].
- For each index $i \in [n]$ in the sequence, define the following index sets:

$$\mathcal{D}_i = \{k \in [n] : |k - i| < \tau\} \quad \text{and} \quad \mathcal{I}_i = [n] \setminus \mathcal{D}_i.$$

- **WingIt** estimator:

$$\widehat{M}_{\tau,\zeta}^{(i)} := \mathbb{I}\{N_{X_i}(X_{\mathcal{I}_i}) = \zeta\}, \quad \text{and} \quad \widehat{M}_{\mathsf{WingIt},\zeta}(\tau) = \frac{1}{n} \sum_{i=1}^{n} \widehat{M}_{\tau,\zeta}^{(i)}.$$

- **Plug-in** estimator:

$$\widehat{M}_{\mathsf{PI},\zeta} = \varphi_\zeta \cdot \frac{\zeta}{n}.$$

- Both the **WingIt** and **PI** estimators can be computed in $n\tau$ time (**linear-time**).
- **Combined** estimator:

$$\widehat{M}_\zeta(\tau; \bar{\zeta}) = \begin{cases} \nu^{-1} \cdot \widehat{M}_{\mathsf{WingIt},\zeta}(\tau) & \text{if } \zeta \leq \bar{\zeta} \\ \nu^{-1} \cdot \widehat{M}_{\mathsf{PI},\zeta} & \text{if } \zeta > \bar{\zeta}, \end{cases}$$

where $\bar{\zeta}$: transition point, and ν: normalizing constant.

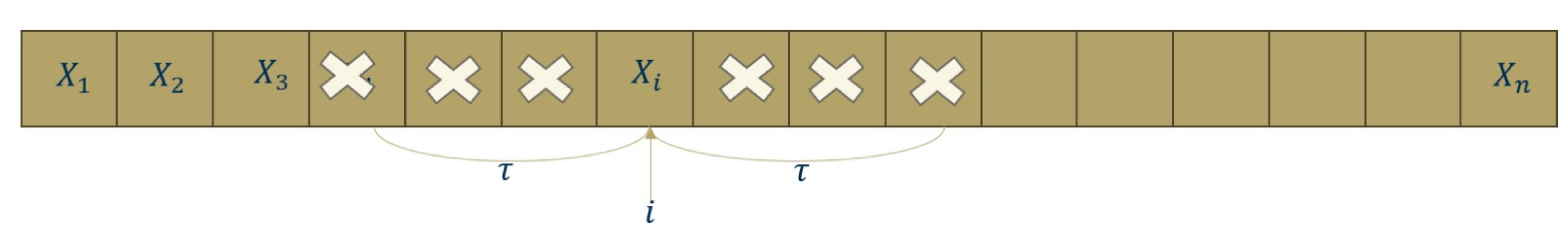


Figure 1. WingIt implementation

Theorem

There exists a universal positive constant C such that if we choose the window size $\tau \geq \mathsf{t}_{\mathsf{mix}}(n^{-5})$ and transition point $\bar{\zeta} = \lfloor n^{1/3} \rfloor - 1$, then

$$\mathcal{R}_n(M^\pi, \widehat{M}(\tau; \bar{\zeta})) \leq C \cdot \left(\frac{\sqrt{\tau \log(Cn)}}{n^{1/6}} \right).$$

- For the i.i.d. case, Theorem 1 recovers the TV guarantee of [3], but without requiring Poissonization.
- A universal result independent of the alphabet size.
- Applying the reverse Pinsker inequality to Theorem 1 implies consistent estimation in KL-divergence.

Oracle inequality

- An estimator $\widehat{M}$ for M^π "naturally" leads to an estimator $\widehat{q}$ for π.
- Divide $\widehat{M}_\zeta$ equally among all elements of $\mathcal{X}$ that occur ζ times in X^n.
- **Oracle inequality** w.r.t. the class of **natural** estimators ($\mathcal{Q}^{\mathsf{nat}}$):

$$d_{\mathsf{TV}}(\pi, \widehat{q}(X^n)) \leq 2 \cdot \inf_{q \in \mathcal{Q}^{\mathsf{nat}}} d_{\mathsf{TV}}(\pi, q) + d_{\mathsf{TV}}(\widehat{M}(X^n), M^\pi(X^n)).$$

The infimum on the RHS is taken over all natural estimators, including those that have perfect knowledge of π but are constrained to be natural.

Frequency by frequency error

(**PI-error**) Fix $\delta, \epsilon > 0$. Let $\tau_0 := \mathsf{t}_{\mathsf{mix}}(\epsilon/n^2)$. If $n \geq 24\tau_0$, then for each large enough ζ we have

$$\left| M_\zeta^\pi - \widehat{M}_{\mathsf{PI},\zeta} \right| \leq \widetilde{\mathcal{O}} \left(\frac{\sqrt{\zeta \mathsf{t}_{\mathsf{mix}}(\epsilon/n^2)} \cdot \varphi_\zeta(X^n)}{n} \right) \text{ with probability at least } 1 - \delta - 3\epsilon.$$

(**Adaptive WingIt -error**) If the window size τ is large enough, then we have

$$\mathbb{E}\left[\left| M_\zeta^\pi(X^n) - \widehat{M}_{\mathsf{WingIt},\zeta} \right| \right] \leq \mathcal{O} \left(\sqrt{\frac{\tau}{n}} \left(\sqrt{\mathbb{E}[M_\zeta^\pi]} + \sqrt{\zeta \log(2\tau) \mathbb{E}[M_\zeta^\pi]} + \sqrt{\sum_{u=1}^{4\tau-2} \frac{(\zeta+u)}{u} \mathbb{E}\left[M_{\zeta+u}^\pi\right]} \right) + \frac{(\zeta+1)\tau}{n} \right).$$

Numerical experiments

Comparison of our proposed estimator against popular i.i.d. estimators.

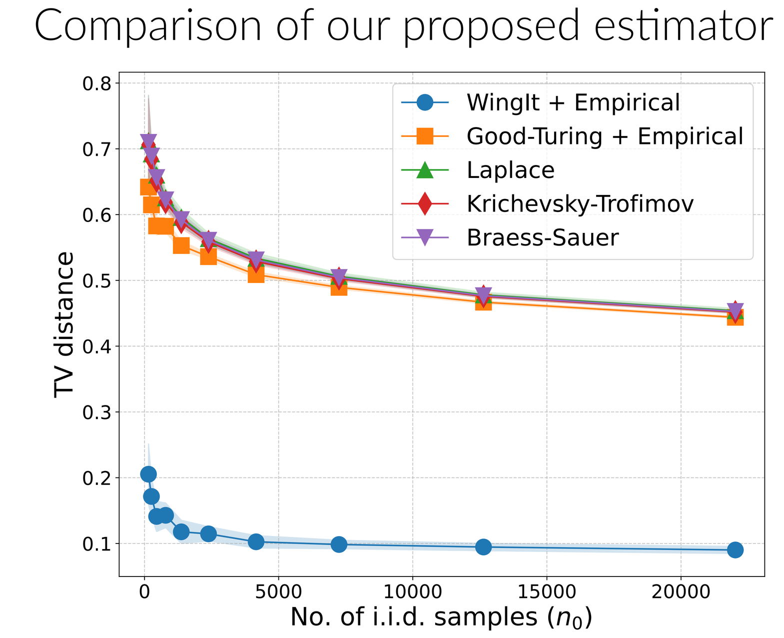


Figure 2. Power law distribution

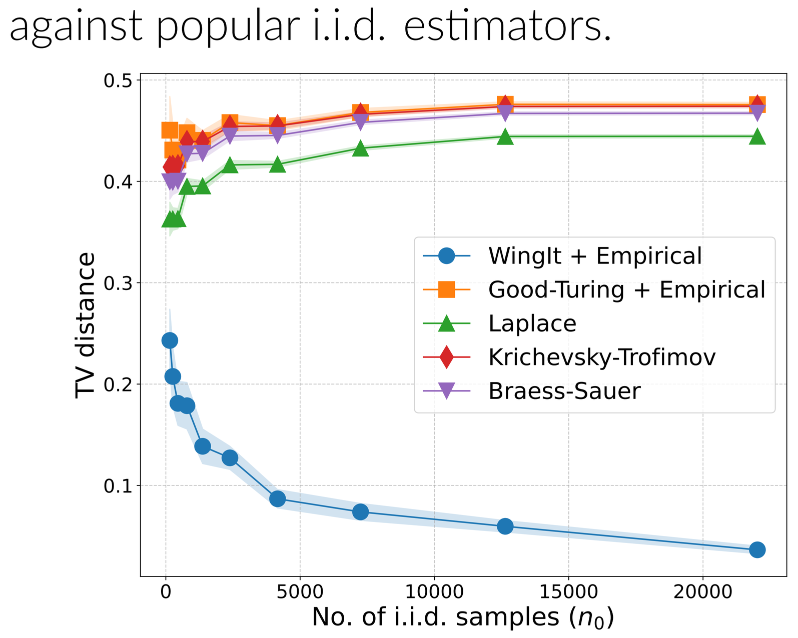


Figure 3. Uniform distribution

Discussions and Future direction

Conclusion:

- We proposed a flexible estimator and analysis of the vector of count probabilities $M^\pi(X^n)$ of any exponentially α-mixing stochastic process.
- An explicit construction for the IID case [3] reveals that our estimation error rate is sharp in its dependence on n.
- Obtaining a minimax optimal estimator for the correlated setting remains an interesting direction of future work.

References

[1] Ashwin Pananjady, Vidya Muthukumar and Andrew Thangaraj. Just Wing it: Near-optimal estimation of missing mass in a Markovian sequence.

[2] I. J. Good. The population frequencies of species and the estimation of population parameters.

[3] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky and Ananda Theertha Suresh. Optimal probability estimation with applications to prediction and classification.

[4] Prafulla Chandra and Andrew Thangaraj and Nived Rajaraman. How good is Good–Turing for Markov samples?