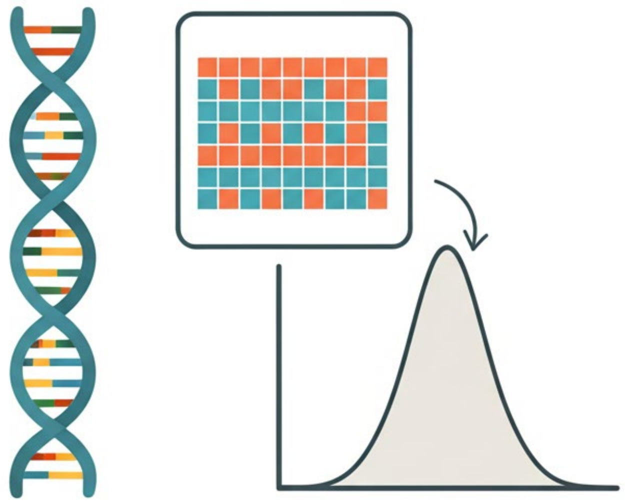
Estimating stationary mass, frequency by frequency

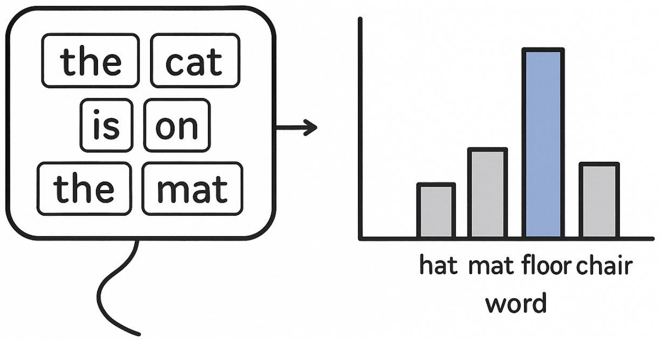
Milind Nakul   Vidya Muthukumar   Ashwin Pananjady

Georgia Institute of Technology

Distribution Estimation

- Discrete **distribution estimation** from **dependent** data.
- Focus: **Large alphabet regime.**
- Understand existing results in the IID setting.
- **Goal:** Design a **universally consistent** estimator when the samples are dependent.


DNA Sequencing



Natural Language Modelling

Kalai and Vempala `24, Lijoi et al `07, Church and Gale `91, Chen and Goodman `99

Dependent data and mixing

- **Data**: $X^n = \{X_1, \dots, X_n\}$, **single** trajectory from stochastic ergodic process
- π: Unique stationary distribution, $\mathcal{X}$: sample space.

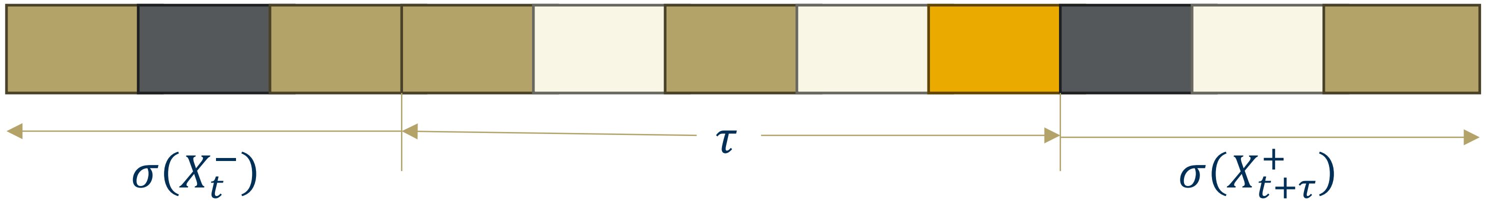
Assumption (exponentially α −mixing)

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

$\alpha(\tau) \leq \mu\rho^\tau$ for all $\tau \geq 1$, $\mu > 0$ and $\rho \in (0,1)$.

Mixing time

$$T_{mix} \coloneqq \min\{\tau : \alpha(\tau) \leq 1/4\}.$$



- **Practical modelling choice**: Subsumes finite state Markov chains, hidden Markov models, duplicated sequences, etc.

Count probabilities

- **Estimate:** $M^{\pi}(X^n) := \left(M_{\zeta}^{\pi}(X^n)\right)_{\zeta=0}^{n}$.

- For each $\zeta = 0,1,\ldots n$, define

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

where $N_x(X^n)$: number of times symbol x appears in X^n.

- $M_{\zeta}^{\pi}(X^n)$: Aggregates stationary masses of symbols with count ζ in X^n.

- $M_{\zeta}^{\pi}(X^n)$: Random functional which depends on X^n and π.

Example

- **Alphabet** $\mathcal{X} = \{a, b, c, d\}$ with $\pi_a = 0.4, \pi_b = 0.2, \pi_c = 0.3, \pi_d = 0.1$.

- **Data** $X^n = \{a, b, a, a, c, c, b\}$.
- ϕ_ζ: # of symbols appearing ζ times.
  $\phi_0 = 1\ (\{d\}), \phi_1 = 0,\ \phi_2 = 2\ (\{b, c\})\ , \phi_3 = 1(\{a\})$.

- Vector of count probabilities:

$$M^\pi(X^n) = (\underset{\substack{\downarrow \\ \boxed{\pi_d}}}{0.1},\ \ 0,\ \ \underset{\substack{\downarrow \\ \boxed{\pi_b + \pi_c}}}{0.5},\ \ \underset{\substack{\downarrow \\ \boxed{\pi_a}}}{0.4},\ \ 0, \ldots\ ,\ \ 0).$$

- **Goal:** Design an estimator $\widehat{\boldsymbol{M}}$**:** $\boldsymbol{\mathcal{X}^n} \to \boldsymbol{\Delta}(\{\boldsymbol{n}\})$ such that we minimize the risk

$$\mathbb{E}\left[\mathrm{d}_{TV}\left(M^\pi(X^n), \widehat{M}(X^n)\right)\right].$$

$\widehat{M}$ can be used to estimate π

- $\hat{q}$: Divide $\widehat{M}$ **equally** among all symbols that appear equal times.

$$\widehat{q_x} = \frac{\widehat{M}_\zeta}{\zeta} \text{ for all } x \text{ such that } N_x(X^n) = \zeta.$$

- $\hat{q}$ is a **natural estimator** (Orlitsky and Suresh `15).

Lemma (Nakul, Muthukumar and Pananjady `25)

$$\mathrm{d}_{TV}(\pi, \hat{q}) \le 2 \inf_{q \in Q^{nat}} \mathrm{d}_{TV}(\pi, q) + \mathrm{d}_{TV}\left(M^\pi, \widehat{M}\right).$$

- $\hat{q}$: **Competitive** with respect to the class of natural estimators.

Orlitsky and Suresh `15

Special case: IID data

- **Data**: $X^n = \{X_1, \ldots, X_n\}$, i.i.d. samples from π.
- $\alpha(\tau) = 0$ and $T_{mix} = 1$.
- One approach: Empirical or **Plug-in (PI)** Estimator.

$$\widehat{M}_{PI,\zeta} = \frac{\zeta \phi_\zeta(X^n)}{n}.$$

- Failure: **Large alphabet** regimes (most symbols are unseen).
- **Smoothing** techniques: Add constant estimators.
- **Hybrid** estimation: **Good Turing** (small ζ) + Plug-In (large ζ).

Laplace 1814, Krichevsky and Trofimov `81, Braess and Sauer `04, Acharya et al `13

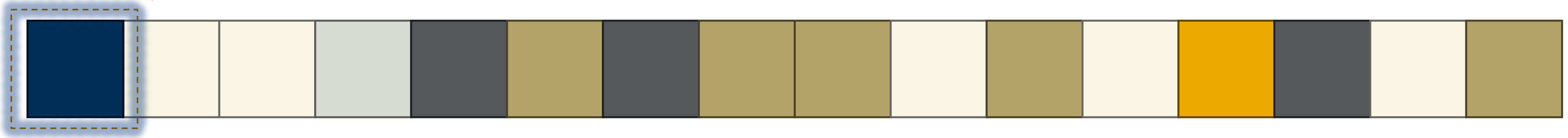
Good-Turing Estimator

- **Good `53**: Proposed the Good-Turing (GT) estimator:

$$\widehat{M}_{GT,\zeta} = \frac{(\zeta+1)\phi_{\zeta+1}(X^n)}{n}.$$

- **Observation**: $M_\zeta^\pi \approx \frac{(\zeta+1)\mathbb{E}[\phi_{\zeta+1}(X^n)]}{n}$.

$\mathbb{I}\left(X_i \notin \left(X^{-i}\right)\right) = 1$



$i = 1$

Example: $n = 16, \zeta = 0, \phi_1(X^n) = 3$

- **Leave-one-out interpretation:** $\widehat{M}_{GT,0} = \frac{1}{n}\sum_i \mathbb{I}\left(X_i \notin \left(X^{-i}\right)\right)$.

Good and Turing `53, McAllester and Schapire `00

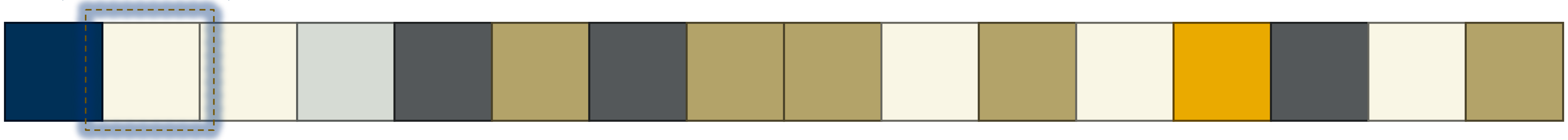
Good-Turing Estimator

- **Good `53**: Proposed the Good-Turing (GT) estimator:

$$\widehat{M}_{GT,\zeta} = \frac{(\zeta+1)\phi_{\zeta+1}(X^n)}{n}.$$

- **Observation**: $M_\zeta^\pi \approx \frac{(\zeta+1)\mathbb{E}[\phi_{\zeta+1}(X^n)]}{n}.$

$\mathbb{I}\left(X_i \notin \left(X^{-i}\right)\right) = 0$



$i = 2$

Example: $n = 16, \zeta = 0, \phi_1(X^n) = 3$

- **Leave-one-out interpretation:** $\widehat{M}_{GT,0} = \frac{1}{n}\sum_i \mathbb{I}\left(X_i \notin \left(X^{-i}\right)\right).$

Good and Turing `53, McAllester and Schapire `00

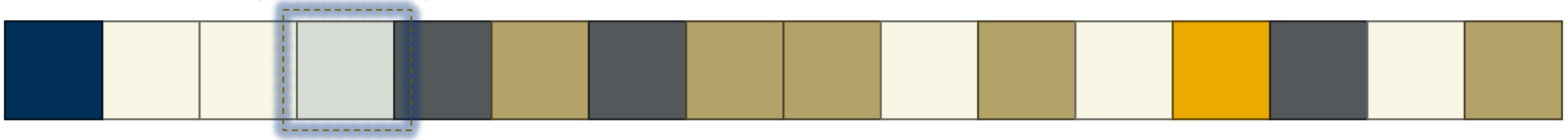
Good-Turing Estimator

- **Good `53**: Proposed the Good-Turing (GT) estimator:

$$\widehat{M}_{GT,\zeta} = \frac{(\zeta+1)\phi_{\zeta+1}(X^n)}{n}.$$

- **Observation**: $M_\zeta^\pi \approx \frac{(\zeta+1)\mathbb{E}[\phi_{\zeta+1}(X^n)]}{n}$.

$\mathbb{I}\left(X_i \notin (X^{-i})\right) = 1$



$i = 4$

Example: $n = 16, \zeta = 0, \phi_1(X^n) = 3$

- **Leave-one-out interpretation:** $\widehat{M}_{GT,0} = \frac{1}{n}\sum_i \mathbb{I}\left(X_i \notin (X^{-i})\right)$.

Good and Turing `53, McAllester and Schapire `00

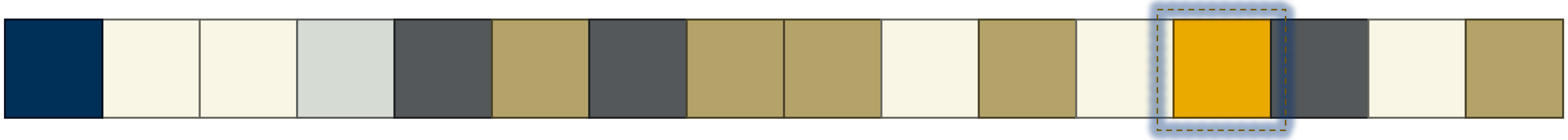
Good-Turing Estimator

- **Good `53**: Proposed the Good-Turing (GT) estimator:

$$\widehat{M}_{GT,\zeta} = \frac{(\zeta + 1)\phi_{\zeta+1}(X^n)}{n}.$$

- **Observation**: $M_\zeta^\pi \approx \frac{(\zeta+1)\mathbb{E}[\phi_{\zeta+1}(X^n)]}{n}$.

$\mathbb{I}\left(X_i \notin (X^{-i})\right) = 1$



$i = 13$

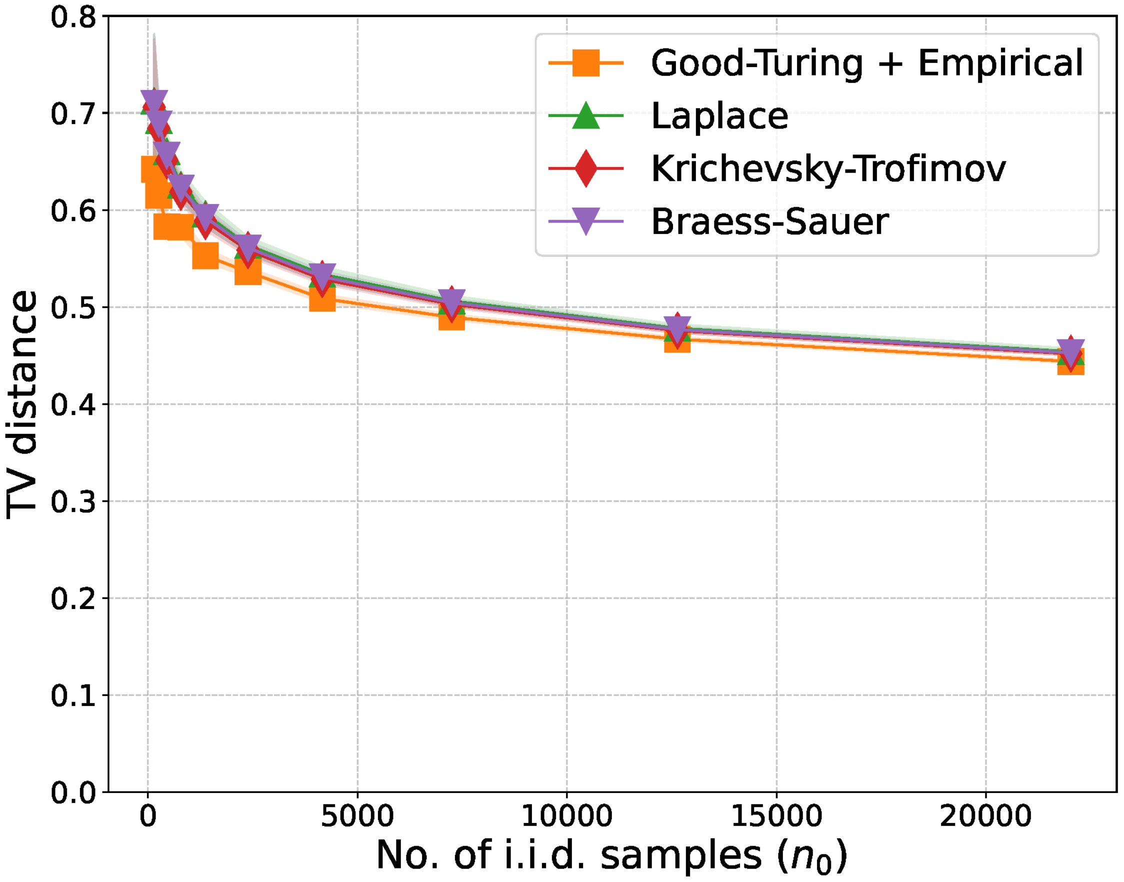
Example: $n = 16, \zeta = 0, \phi_1(X^n) = 3$

- **Leave-one-out interpretation:** $\widehat{M}_{GT,0} = \frac{1}{n}\sum_i \mathbb{I}\left(X_i \notin (X^{-i})\right)$.

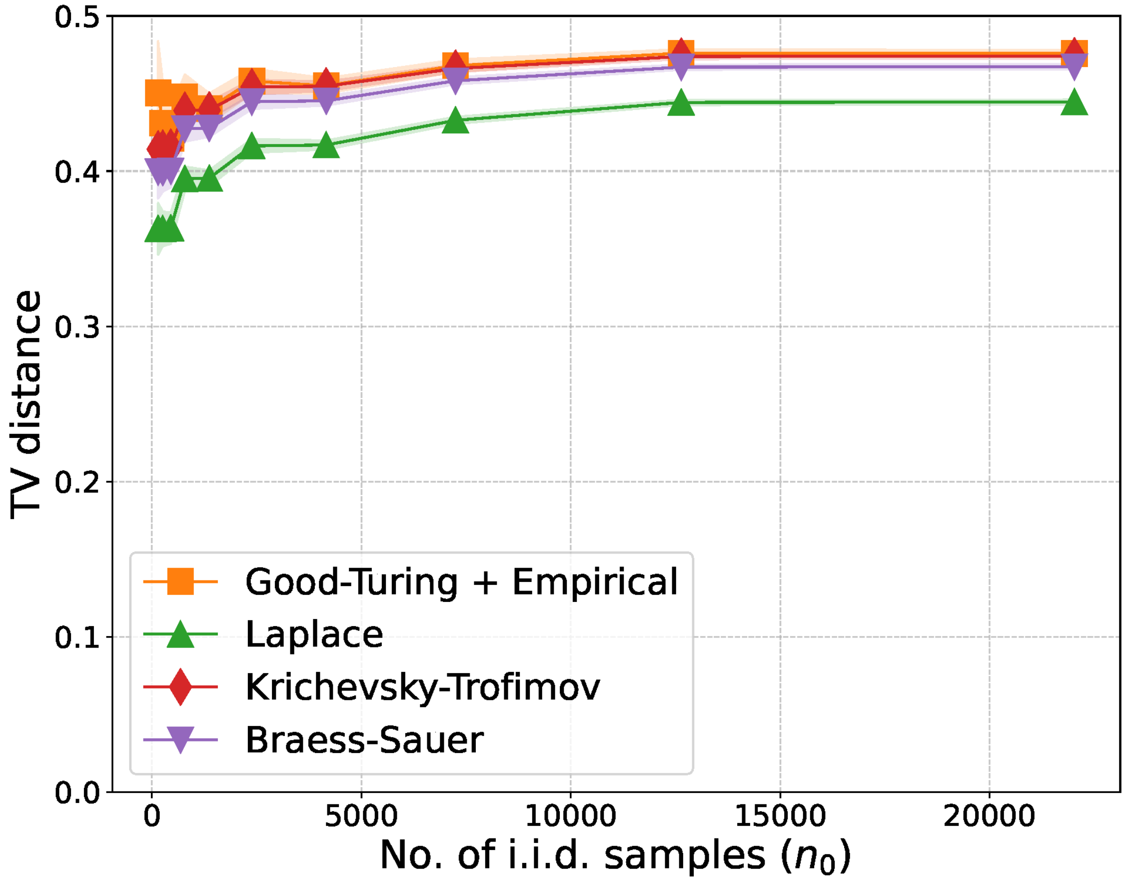
Good and Turing `53, McAllester and Schapire `00

Failure of IID estimators on sticky Markov chains

- Generate **base** i.i.d. sequence of length = n_0.
- Duplicate each symbol in the sequence by i.i.d. factor **Geometric**$(1/n_0^{0.2})$.



Power law distribution on base sequence



Uniform distribution on base sequence

Difficulties

- **Temporal dependence**.
- **Chandra et al `22** : GT on **Markovian** data has a **constant bias**.
- **Pananjady et al `24 :** Estimator for **missing mass** for **Markovian** sequences.

Question: Can we utilize the recent progress for Markovian sequences to perform **consistent distribution estimation** for **general mixing sequences**?

Chandra, Thangaraj, and Rajaraman `22, Pananjady, Muthukumar, and Thangaraj `24

Windowed-Good-Turing (Wing-It)

- **Recall:** leave-one-out GT interpretation:

$$\widehat{M}_{GT,\zeta} = \frac{1}{n}\sum_{i} \mathbb{I}\left(N_{X_i}\left(X^{-i}\right) = \zeta\right).$$

- Define the **“Independent”** set $\mathcal{J}_i = [1, i-\tau] \cup [i+\tau, n]$.

$$\widehat{M}_{WingIt,\zeta} = \frac{1}{n}\sum_{i} \mathbb{I}\left(N_{X_i}\left(X_{\mathcal{J}_i}\right) = \zeta\right).$$

- **Wing-It**: GT estimator with windowing (leave-a-window-out).

$\mathbb{I}\left(N_{X_i}(X_{\mathcal{J}_i}) = 0\right) = 1$

$i = 2$

Example: $n = 16, \zeta = 0, \tau = 2.$

Windowed-Good-Turing (Wing-It)

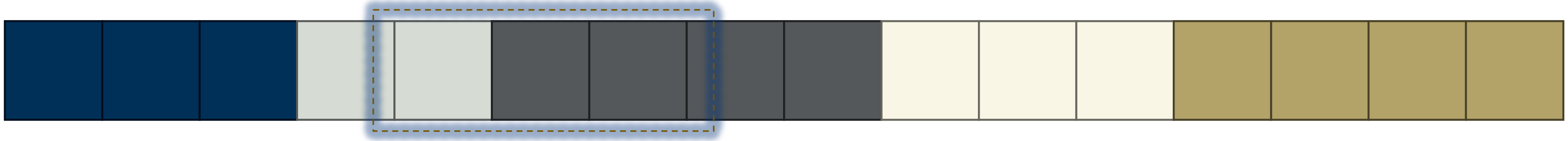
- **Recall:** leave-one-out GT interpretation:

$$\widehat{\boldsymbol{M}}_{\boldsymbol{GT},\boldsymbol{\zeta}} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i}} \mathbb{I}\left(\boldsymbol{N}_{\boldsymbol{X}_{\boldsymbol{i}}}\left(\boldsymbol{X}^{-\boldsymbol{i}}\right) = \boldsymbol{\zeta}\right).$$

- Define the **"Independent"** set $\mathcal{J}_i = [1, i - \tau] \cup [i + \tau, n]$.

$$\widehat{\boldsymbol{M}}_{\boldsymbol{WingIt},\boldsymbol{\zeta}} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i}} \mathbb{I}\left(\boldsymbol{N}_{\boldsymbol{X}_{\boldsymbol{i}}}\left(\boldsymbol{X}_{\boldsymbol{\mathcal{J}}_{\boldsymbol{i}}}\right) = \boldsymbol{\zeta}\right).$$

- **Wing-It**: GT estimator with windowing (leave-a-window-out).



Example: $n = 16, \zeta = 0, \tau = 2.$

Windowed-Good-Turing (Wing-It)

- **Recall:** leave-one-out GT interpretation:

$$\boldsymbol{\widehat{M}_{GT,\zeta} = \frac{1}{n}\sum_i \mathbb{I}\left(N_{X_i}\left(X^{-i}\right) = \zeta\right).}$$

- Define the “**Independent**” set $\mathcal{J}_i = [1, i-\tau] \cup [i+\tau, n]$.

$$\boldsymbol{\widehat{M}_{WingIt,\zeta} = \frac{1}{n}\sum_i \mathbb{I}\left(N_{X_i}\left(X_{\mathcal{J}_i}\right) = \zeta\right).}$$

- **Wing-It**: GT estimator with windowing (leave-a-window-out).

$\mathbb{I}\left(N_{X_i}(X_{\mathcal{J}_i}) = 0\right) = 1$

$i = 11$

Example: $n = 16, \zeta = 0, \tau = 2.$

Estimator for mixing sequences

Hybrid Estimator ($\widehat{M}$)

$0 \le \zeta \le \bar{\zeta}$

$\bar{\zeta} < \zeta \le n$

Wing-It for small frequencies

$$\widehat{M}_\zeta = \widehat{M}_{WingIt,\zeta}.$$

Implementable in $\boldsymbol{n\tau}$ **time**.

PI for large frequencies

$$\widehat{M}_\zeta = \widehat{M}_{PI,\zeta} = \frac{\zeta \phi_\zeta(X^n)}{n}.$$

Implementable in **linear time**.

Theorem (Nakul, Muthukumar and Pananjady `25)

For mixing sequences, the hybrid estimator with $\bar{\zeta} \asymp n^{1/3}$, achieves

$$\mathrm{d}_{TV}\left(M^\pi, \widehat{M}\right) \lesssim \sqrt{T_{mix} \log n} \cdot n^{-1/6}.$$

Theorem (Nakul, Muthukumar and Pananjady `25)
For mixing sequences, the hybrid estimator with $\bar{\zeta} \asymp n^{1/3}$, achieves

$$\mathrm{d}_{TV}(M^{\pi}, \widehat{M}) \lesssim \sqrt{T_{mix} \log n} \cdot n^{-1/6}.$$

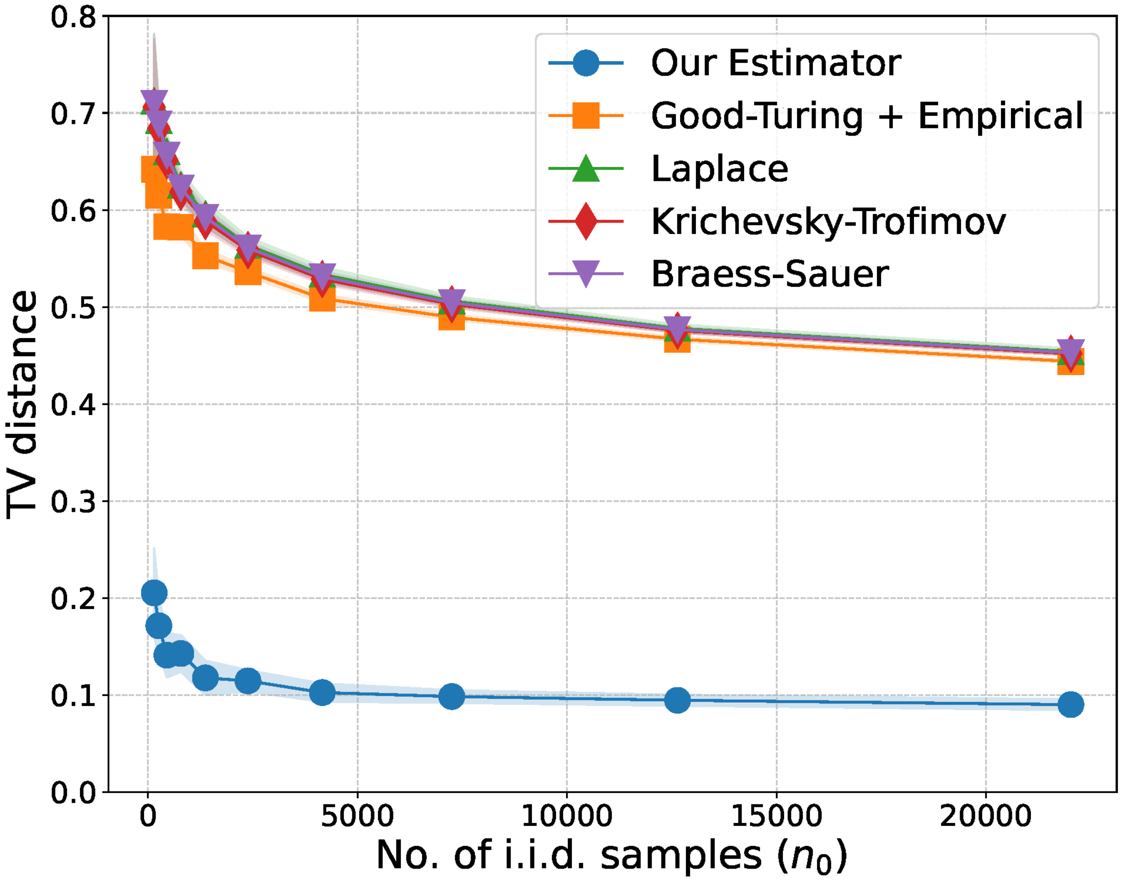
- **Universal** estimation in n for any alphabet size $|\mathcal{X}|$.
- Recovers i.i.d. results as special case $\tau = 1$ without **Poissonization**.
- New analysis for **small-frequency** and **large frequency** errors.

Adaptive analysis which adjusts to the properties of the stationary distribution.

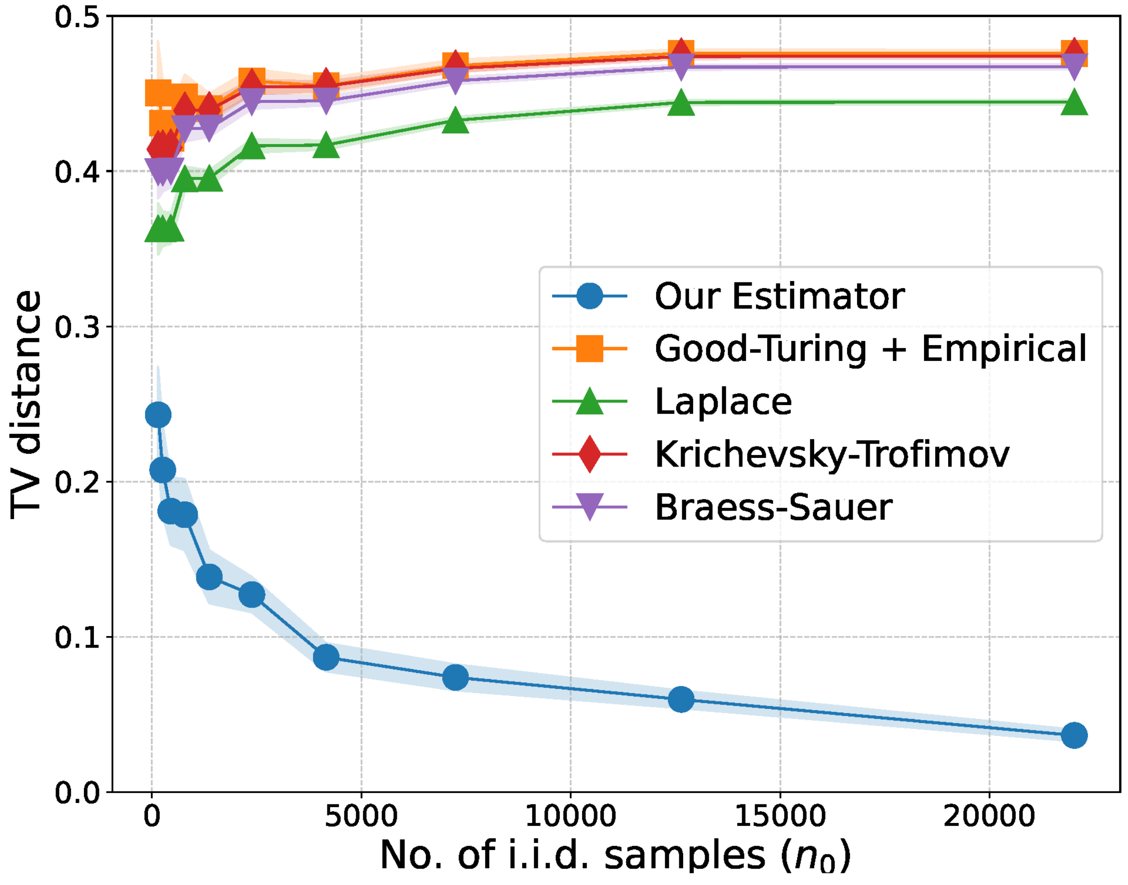
Self-normalized concentration inequalities for α-mixing sequences.

Success of Proposed Estimator

- Generate **base** i.i.d. sequence of length = n_0.
- Duplicate each symbol in the sequence by i.i.d. factor **Geometric**$(1/n_0^{0.2})$.



Power law distribution on base sequence



Uniform distribution on base sequence

Key takeaways:

- Discrete distribution estimation from **dependent** samples.
- Proposed estimator: Complementary strengths of **Wing-It** and **Plug-In**.
- **Consistent** estimation for all alphabet sizes.

Future direction:

- **Minimax-optimal** rates for distribution estimation on mixing sequences?
- Beyond **TV distance.**

Thank You