Estimating stationary mass, frequency by frequency

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

H. Milton Stewart School of Industrial and Systems Engineering*
School of Electrical and Computer Engineering†
Georgia Institute of Technology

June 9, 2025

Abstract

Suppose we observe a trajectory of length n from an exponentially α-mixing stochastic process over a finite but potentially large state space. We consider the problem of estimating the probability mass placed by the stationary distribution of any such process on elements that occur with a certain frequency in the observed sequence. We estimate this vector of probabilities in total variation distance, showing universal consistency in n and recovering known results for i.i.d. sequences as special cases. Our proposed methodology—implementable in linear time—carefully combines the plug-in (or empirical) estimator with a recently-proposed modification of the Good–Turing estimator called WINGIT, which was originally developed for Markovian sequences. En route to controlling the error of our estimator, we develop new performance bounds on WINGIT and the plug-in estimator for exponentially α-mixing stochastic processes. Importantly, the extensively used method of Poissonization can no longer be applied in our non i.i.d. setting, and so we develop complementary tools—including concentration inequalities for a natural self-normalized statistic of mixing sequences—that may prove independently useful in the design and analysis of estimators for related problems. Simulation studies corroborate our theoretical findings.

1 Introduction

Estimating the stationary distribution of a stochastic process from an n-length trajectory is a foundational problem in statistics and machine learning, with broad implications for fields such as ecology (Fisher et al., 1943), genomics (Favaro et al., 2012; Lijoi et al., 2007) and natural language processing (Church and Gale, 1991; Chen and Goodman, 1999; Ney et al., 1994). Historically, most research has focused on settings where the data is either i.i.d. or exchangeable, where classical techniques were based on so-called "add-constant" estimators (Laplace, 1814; Krichevsky and Trofimov, 1981). Other estimators that see appealing empirical performance are absolute discounting (Ney et al., 1994), Jelinek–Mercer smoothing (Jelinek, 1985), and the Good–Turing estimator (Good, 1953). These estimators proceed by first estimating the probability mass[1] M_ζ^π placed by the distribution π on symbols occurring with each frequency $\zeta = 0, 1, \ldots, n$ in the sequence, and this estimation problem will form the focus of the current paper. One can then transform the estimate of the vector $(M_\zeta^\pi)_{\zeta=0}^n$ to an estimator of π itself; see Orlitsky and Suresh (2015). Among the aforementioned estimators, the Good–Turing estimator has been thoroughly analyzed in the i.i.d. setting, starting from the pioneering work of McAllester and Schapire (2000). In particular, these authors showed a *universal* result of the following form: For any i.i.d. sequence from a

---

[1]Note that M_ζ^π is a random functional of the stochastic process, since the set of symbols occurring ζ times in the observed sequence is random.

distribution π defined on a finite alphabet, the Good–Turing estimator consistently estimates the probability mass M_ζ^π placed by the distribution on symbols occurring with each fixed frequency ζ in the sequence. Subsequent work (Drukh and Mansour, 2005; Acharya et al., 2013; Orlitsky and Suresh, 2015) refined these bounds and proposed a *hybrid* estimator, i.e. the Good–Turing estimator to approximate M_ζ^π for small ζ (i.e., to estimate the probability mass placed on small-frequency symbols) and the plug-in or empirical estimator to approximate M_ζ^π for large ζ (i.e., to estimate the probability mass placed on large-frequency symbols). In addition, the problem of estimating the missing mass, M_0^π, has garnered significant interest in the i.i.d. setting and has been analyzed thoroughly (McAllester and Ortiz, 2003; Berend and Kontorovich, 2012, 2013). In particular, the problem of estimating the probability vector $(M_\zeta^\pi)_{\zeta=0}^n$ has been studied in various divergences including total variation, Kullback–Leibler, and chi-square. In addition to upper bounds, we also have an intricate understanding of the sharpness properties of various estimators as well as minimax lower bounds (Acharya et al., 2013; Orlitsky and Suresh, 2015). There has also been a significant body of follow-up work in the i.i.d. setting (see, e.g., Ohannessian and Dahleh, 2012; Painsky, 2023; Hao and Orlitsky, 2019).

While the aforementioned papers have provided a detailed understanding of the i.i.d. setting, many real-world stochastic processes exhibit temporal dependencies. Examples abound in natural language processing, where sequences of words are sometimes better captured through hidden Markov modeling, and genomics, where DNA sequences are formed from interdependent base pairs. These dependencies in the stochastic process pose significant challenges for estimation of the stationary distribution. Indeed, the naive application of estimators such as Good–Turing is no longer valid even when the stochastic process is Markovian (Chandra et al., 2024) and different algorithmic tools are required for addressing temporal dependence (Hao et al., 2018; Skorski, 2020; Chandra et al., 2024). Motivated by this issue, recent work by Pananjady et al. (2024) studied the Markovian setting and introduced a "leave-a-window-out" variant of the Good–Turing estimator for estimating the stationary *missing mass*, i.e., the probability mass placed by the stationary distribution on elements that do not occur in the observed sequence. They showed that this variant of Good–Turing — which they termed *windowed Good–Turing* or WINGIT — estimates the stationary missing mass with minimax optimal mean-squared error. Some extensions were also provided for estimation of the "small-count" probability, i.e., the mass placed by the distribution on elements occurring *at most* a certain number of times in the observed sequence. However, these results do not imply a satisfactory bound on distribution estimation. In particular, we still do not have an estimator for the stationary distribution in a frequency-by-frequency sense that estimates the vector of probabilities $(M_\zeta^\pi)_{\zeta=0}^n$ consistently in any natural divergence measure. Finally, we mention that besides frequency-by-frequency stationary distribution estimation, researchers have studied various learning problems involving other parameters of Markov chains (e.g., Hao et al., 2018; Han et al., 2023; Wolfer and Kontorovich, 2019).

In this paper, we propose and theoretically analyze an estimator for the stationary distribution in a frequency-by-frequency sense. Specifically, we consider a stochastic process $X^n = (X_1, X_2, \ldots, X_n)$ defined over a finite but unknown state space $\mathcal{X}$ in which we accommodate the regime $|\mathcal{X}| \gg n$. We assume that the process is exponentially α-mixing —which subsumes mixing Markov chains and some hidden Markov models; see Section 2.1 to follow — and aim to estimate its stationary distribution π in the aforementioned sense. Specifically, letting M_ζ^π denote the probability mass placed by the stationary distribution on elements occurring ζ times in X^n, our goal is to estimate the vector $(M_\zeta^\pi)_{\zeta=0}^n$ in total variation distance. Importantly, we would like to develop a

universal result like in the i.i.d. case, which shows that estimation can always be performed consistently in n. We show that such an estimator then naturally yields an estimator for the stationary distribution π defined over the sample space $\mathcal{X}$ (see Lemma 1).

Contributions and organization

Our main contributions are summarized below:

- We propose a simple and efficient estimator for the stationary distribution of any exponentially α-mixing stochastic process. Our estimator combines the WINGIT estimator (Pananjady et al. (2024)) for low frequency and the plug-in estimator for high frequency symbols.

- In Theorem 1, we provide a risk bound on the proposed hybrid estimator, showing that with the appropriate parameter settings it attains total variation risk $\widetilde{\mathcal{O}}\left(n^{-1/6}\sqrt{\mathsf{t_{mix}}}\right)$ for any exponentially α-mixing process with mixing time $\mathsf{t_{mix}}$ (see Eqs. (1a) and (2)). This rate recovers known guarantees in the i.i.d. setting as a special case.

Our main result, Theorem 1, is proved through two key technical pieces of possible independent interest. First, Proposition 1, presented in Section 4, provides a high-probability bound on the ℓ_1 error of the plug-in estimator. Notably, these results recover the known guarantees for the i.i.d. setting as special cases, but our analysis does not rely on Poissonization, a common technique tailored to the i.i.d. setting used to handle dependent multiplicities of symbols. Instead, we develop an alternative approach based on blocking arguments, which decomposes the sequence into approximately independent blocks. In the process, we prove concentration inequalities for a natural self-normalized statistic for mixing stochastic processes.

Second, we establish a bound on the ℓ_1-risk of the WINGIT estimator for a general exponentially α-mixing process in Proposition 2, presented in Section 4. Importantly, simply applying the worst-case bounds provided in Pananjady et al. (2024) would not yield Theorem 1, and so we provide a sharper bound on the ℓ_1-risk that *adapts* to the behavior of the stationary distribution π. Specifically, if π assigns low probability to low-frequency symbols, the bound naturally adjusts to reflect this, leading to a smaller error for the WINGIT estimator.

The rest of this paper is organized as follows. Section 2 describes basic background and formalizes our problem statement. Section 3 describes the methodology for our estimator, and our main results for the analysis of this estimator are described in Section 4. In Section 5, we simulate our estimator and compare its performance to alternatives that are available in the literature. In Section 6, we provide proofs of our main results, deferring statements and proofs of auxiliary lemmas to the appendix.

Notation

Let $[n]$ denote the set of natural numbers less than or equal to n. For an index set $P \subseteq [n]$, let $X_P = (X_i)_{i \in P}$ denote the sequence of random variables with indices in P, ordered canonically. For a sequence $Z \in \mathcal{X}^{\mathbb{N}}$ and $x \in \mathcal{X}$, we let $N_x(Z) = \#\{i : Z_i = x\}$ denote the number of occurrences of x in Z. We frequently use the shorthand $N_x = N_x(X^n)$. We also let $\varphi_\zeta(Z) = \#\{x : N_x(Z) = \zeta\}$ denote the number of symbols occurring with frequency ζ in Z, once again adopting the shorthand $\varphi_\zeta = \varphi_\zeta(X^n)$. We use $\Delta(S)$ to denote the set of all probability mass functions on a finite set S. We use $\oplus$ to denote the concatenation of sequences, e.g., $(X_1, X_2) \oplus (X_4, X_5) = (X_1, X_2, X_4, X_5)$.

We use the notation $f(u) \lesssim g(u)$ to mean that there exists some absolute positive constant C that is independent of all problem parameters, such that $f(u) \leq C \cdot g(u)$ for all u in the domain of f and g. We use the notation $f(u) \gtrsim g(u)$ when $g(u) \lesssim f(u)$. We write $f(u) \asymp g(u)$ if both relations $f(u) \gtrsim g(u)$ and $g(u) \lesssim f(u)$ hold. Logarithms are taken to the base e. We use (c, C) to denote universal positive constants that could be different in each instantiation. We use $\mathsf{var}(X)$ to denote the variance of a random variable X and $\mathrm{cov}(X, Y)$ to define the covariance between two random variables X and Y. For a sequence of (deterministic or random) scalars $a^n = (a_1, \ldots, a_n)$, let $\mathsf{sp}(a^n) := \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} (a_i - a_j)^2$ denote its spread.

2 Background and problem formulation

As mentioned before, we are interested in estimating the stationary distribution of a mixing stochastic process X^n. In this section, we formally set up the assumptions on our stochastic process as well as the metric in which we wish to perform estimation.

2.1 Model for the stochastic process

We assume that the sequence $X^n := (X_1, \ldots, X_n)$ is an ergodic stochastic process defined over a finite (and possibly unknown) state space $\mathcal{X}$. The unique stationary distribution of this ergodic process is denoted by π and we assume that the process is initialized at stationarity[2] with $X_1 \sim \pi$.

We also assume that the stochastic process is exponentially α-mixing, which ensures that the dependencies between lagged past and future observations in the process decay exponentially with the time lag. Formally, for an ergodic stochastic process $\{X_t\}_{t \geq 1}$, define its α-mixing coefficient as

$$\alpha(\tau) := \sup_{t \in \mathbb{N}} \sup \left\{ |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| : A \in \sigma(X_t^-), B \in \sigma(X_{t+\tau}^+) \right\}, \tag{1a}$$

where at time t, we use $\sigma(X_t^-)$ to denote the σ-algebra generated by the past RVs $(X_1, X_2, \ldots, X_t)$, and $\sigma(X_{t+\tau}^+)$ to denote the σ-algebra generated by the future RVs $(X_{t+\tau}, X_{t+\tau+1}, \ldots)$. Then, the stochastic process is said to satisfy *exponential α-mixing* if there exist constants $\mu > 0$ and $\rho \in (0, 1)$ such that

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

We also define the mixing time of such a process to level $\epsilon \in (0, 1)$ as

$$\mathsf{t}_{\mathsf{mix}}(\epsilon) := \min\{\tau \in \mathbb{N} : \alpha(\tau) \leq \epsilon\}. \tag{2}$$

Note that for an exponentially mixing process satisfying Eq. (1b), we have

$$\mathsf{t}_{\mathsf{mix}}(\epsilon) \leq \frac{\log(\mu/\epsilon)}{\log(1/\rho)}. \tag{3}$$

The assumption of exponential α-mixing is satisfied by a wide range of stochastic processes, making it a versatile and practical modeling choice. Below, we discuss some examples:

---

[2]This assumption can easily be relaxed by allowing for a burn-in period.

1. **I.I.D. sequences:** IID sequences are special cases of α-mixing sequences. Since all samples are i.i.d., we have from Eq. (1a) that $\alpha(\tau) = 0$ for all $\tau \in \mathbb{N}$. Then, from the definition of mixing time in Eq. (2) we obtain $\mathsf{t}_{\mathsf{mix}}(\epsilon) = 1$ for all $\epsilon > 0$.

2. **Finite-State Ergodic Markov Chains:** Any ergodic Markov chain on a finite state space satisfies the exponential α-mixing condition (1b). In particular, the parameters (μ, ρ) depend on the size of the state space and the spectral gap of the Markov transition matrix (Levin and Peres, 2017, Theorem 4.9).

3. **Hidden Markov Models (HMMs):** In HMMs, the observations depend on a latent Markov chain. When the latent Markov chain $(H_t)_{t \geq 1}$ is exponentially α-mixing and the observed process $(X_t)_{t \geq 1}$ is some measurable function f_t of H_t and another exponentially mixing Markov process W_t, then $X_t = f_t(H_t, W_t)$ is also exponentially α-mixing (Doukhan, 2012).

4. **Random duplication model:** Suppose i.i.d. samples from π are input to a random duplication model (Chandra and Thangaraj, 2024), which for each input x, outputs the duplicated sequence $\underbrace{(x, \ldots, x)}_{k \text{ times}}$ with probability α, and x otherwise. The output process is ergodic and satisfies exponentially α-mixing, and the mixing time satisfies $\mathsf{t}_{\mathsf{mix}}(\epsilon) = k$ for all $\epsilon > 0$.

2.2 Frequency-by-frequency estimation of the stationary distribution

The total variation (TV) distance between two probability mass functions p and q defined on a common probability space $(\mathcal{X}, \mathcal{F})$ is given by

$$d_{\mathsf{TV}}(p, q) = \sup_{A \in \mathcal{F}} |p(A) - q(A)| = \frac{1}{2} \sum_{x \in \mathcal{X}} |p_x - q_x|,$$

where we have used the shorthand $p_x = p(\{x\})$ and $q_x = q(\{x\})$.

Recall that our goal is to estimate the vector of probabilities $(M_\zeta^\pi)_{\zeta=0}^n$ consistently in total variation, and thereby the stationary distribution π. Let us set up this problem formally. For each $\zeta = 0, 1, \ldots, n$, define

$$M_\zeta^\pi(X^n) := \sum_{x \in \mathcal{X}} \pi_x \cdot \mathbb{I}\{N_x = \zeta\} \tag{4}$$

as the mass placed by π on elements occurring exactly ζ times in the observed sequence X^n. Note that each scalar $M_\zeta^\pi(X^n)$ is a random functional, since it depends both on the underlying distribution and on the realized sequence X^n. Denote the vector of count probabilities by

$$M^\pi(X^n) = (M_0^\pi(X^n), M_1^\pi(X^n), \ldots, M_n^\pi(X^n)), \tag{5}$$

which is a random vector taking values in the probability simplex $\Delta(\{0, 1, \ldots, n\})$.

Our goal is to develop an estimator of the random vector $M^\pi(X^n)$ as a functional of only the observed sequence X^n. In particular, we seek to design an estimator $\widehat{M} : \mathcal{X}^n \to \Delta(\{0, 1, \ldots, n\})$ and measure its error using total variation distance. Specifically, define the risk

$$\mathcal{R}_n(M^\pi, \widehat{M}) := \mathbb{E}\left[d_{\mathsf{TV}}(M^\pi(X^n), \widehat{M}(X^n))\right], \tag{6}$$

where the expectation is taken over X^n and any additional randomness in the estimator.

While the problem of estimating the vector M^π is interesting in itself, it is worth pausing to ask in what sense this yields an estimator for the stationary distribution π. To describe this, we recall the notion of a *natural* estimator due to Orlitsky and Suresh (2015), which is an estimator of π that assigns the same probability to all symbols appearing with the same frequency in X^n. Formally, define the set of all natural estimators

$$\mathcal{Q}^{\mathsf{nat}} := \{q : \mathcal{X}^n \to \Delta(\mathcal{X}) \mid q_x = q_y \text{ if } N_x(X^n) = N_y(X^n)\}.$$

Note that natural estimators are indeed intuitive, since if we know only that a sequence is exponentially α-mixing and therefore reaches stationarity, then it is *natural* to assign the same probability to symbols that occur an equal number of times. In particular, any estimator $\widehat{M} : \mathcal{X}^n \to \Delta(\{0, 1, \ldots, n\})$ of the count probability vector can be used to generate an estimator $\widehat{q}$ for the stationary distribution by dividing the mass $\widehat{M}_\zeta$ equally among all elements of $\mathcal{X}$ that occur ζ times in X^n. The estimator $\widehat{q}$ is natural by definition. The following lemma shows that this estimator is competitive with respect to the class of natural estimators $\mathcal{Q}^{\mathsf{nat}}$.

Lemma 1. *Suppose $\widehat{q} : \mathcal{X}^n \to \Delta(\mathcal{X})$ is generated from an estimator $\widehat{M} : \mathcal{X}^n \to \Delta(\{0, 1, \ldots, n\})$ by splitting the mass $\widehat{M}_\zeta$ equally among all elements that occur exactly ζ times in the sequence X^n. Then, we have*

$$d_{\mathsf{TV}}(\pi, \widehat{q}(X^n)) \leq 2 \cdot \inf_{q \in \mathcal{Q}^{\mathsf{nat}}} d_{\mathsf{TV}}(\pi, q) + d_{\mathsf{TV}}(\widehat{M}(X^n), M^\pi(X^n)). \tag{7}$$

Note that the infimum on the RHS is taken over all natural estimators, including those that have perfect knowledge of π but are constrained to assign the same probability to symbols appearing the same number of times in X^n. We thus have an *oracle inequality* with approximation factor 2, and estimating the vector of count probabilities automatically yields a good estimator of the stationary distribution. As an aside, we note that such a competitive relation was proved for the KL-divergence by Orlitsky and Suresh (2015), who showed an approximation constant of 1 instead of as above. A corresponding result for the TV distance has not appeared before (to our knowledge).

3   Methodology

We now turn our attention to the estimation problem set up above. Our distribution estimation algorithm is based on a careful combination of the WINGIT estimator (Pananjady et al., 2024) for low-frequency symbols and the plug-in estimator for high-frequency symbols. Let us begin by describing these two estimators.

3.1   WingIt Estimator

The WINGIT estimator generalizes the Good-Turing estimator's leave-one-out technique to provide a "leave-a-window-out" framework for estimation with dependent data. To describe this framework, we define some additional notation. For each index $i \in [n]$ in the sequence, define the following index sets:

$$\mathcal{D}_i = \{k \in [n] : |k - i| < \tau\} \quad \text{and} \quad \mathcal{I}_i = [n] \setminus \mathcal{D}_i. \tag{8}$$

For a suitable choice of τ, the set of indices $\mathcal{D}_i$ represents the "dependent set", in that the random variables $\{X_j\}_{j \in \mathcal{D}_i}$ may depend strongly on the random variable X_i. Conversely, $\mathcal{I}_i$ represents the "independent set": If τ is chosen large enough and the stochastic process is mixing, then we expect the random variables indexed by the set $\mathcal{I}_i$ to be approximately independent of X_i. With this notation in hand, we define the *exact-count* estimator

$$\widehat{M}_{\tau,\zeta}^{(i)} := \mathbb{I}\{N_{X_i}(X_{\mathcal{I}_i}) = \zeta\}, \quad \text{and the averaged estimator} \quad \widehat{M}_{\text{WINGIT},\zeta}(\tau) = \frac{1}{n} \sum_{i=1}^{n} \widehat{M}_{\tau,\zeta}^{(i)}, \quad (9)$$

where $\widehat{M}_{\text{WINGIT},\zeta}(\tau)$ is the total probability mass assigned to symbols appearing ζ times in the sample by the WINGIT estimator. Note that substituting $\tau = 1$ yields the Good–Turing estimator. We note that the definitions of the exact-count WINGIT estimators in our work differ slightly from those in Pananjady et al. (2024). Specifically, while Pananjady et al. (2024) were concerned with estimating the *small-count* probability given by $M_{\leq \zeta}^{\pi} = \sum_{s=0}^{\zeta} M_s^{\pi}$, which corresponds to the total mass of elements appearing up to ζ times, our focus is on estimating the exact count probabilities M_{ζ}^{π}, which are objects that are more directly useful for distribution estimation.

We now present an explicit implementation of the WINGIT estimator defined in Eq. (9). We implement the WINGIT estimator in parallel for all values of $\zeta \in \{0, \ldots, n\}$ and return the $n+1$-dimensional vector $\widehat{M}_{\text{WINGIT}}$. This implementation runs in $\mathcal{O}(n\tau)$ time, where τ is the window size used. Given a sequence X^n and a natural number τ, our implementation proceeds via two passes through the data as shown in Algorithm 1.

Algorithm 1 Linear time implementation of the WINGIT estimator.

Require: Sequence $X^n = (X_1, \ldots, X_n)$, Natural number τ (window size)
 1: Initialize `locations` as a dictionary
 2: **for** $i = 1, \ldots, n$ **do**
 3:    **if** $X_i \notin$ `locations` **then**
 4:       Initialize `locations`$[X_i]$ as a list
 5:    **end if**
 6:    Append i to `locations`$[X_i]$
 7: **end for**
 8: $\widehat{M}_{\text{WINGIT}} \leftarrow [\underbrace{0, \ldots, 0}_{n+1}]$
 9: **for** $i = 1, \ldots, n$ **do**
10:    count $\leftarrow 0$
11:    **for** $j = \max(0, i - \tau), \ldots, \min(i + \tau, n)$ **do**
12:       **if** $X_j == X_i$ **then**
13:          count $\leftarrow$ count $+1$
14:       **end if**
15:    **end for**
16:    $\zeta \leftarrow$ `len(locations`$[X_i])-$ count
17:    $\widehat{M}_{\text{WINGIT}}[\zeta] \leftarrow \widehat{M}_{\text{WINGIT}}[\zeta] + 1/n$
18: **end for**
19: **return** $\widehat{M}_{\text{WINGIT}}$

The first loop in Algorithm 1 (Steps 2-7) requires a single pass through the data and thus runs in

$\mathcal{O}(n)$ time. This loop stores the locations of each symbol in the dictionary `locations`. In Step 8 we initialize the vector of count probabilities $\widehat{M}_{\text{WINGIT}}$ as a vector of zeros of size $n + 1$ corresponding to each frequency. The ζ-th element of this vector, denoted by $\widehat{M}_{\text{WINGIT}}[\zeta]$, corresponds to the quantity $\widehat{M}_{\text{WINGIT},\zeta}(\tau)$ defined in Eq. (9). The second loop in this algorithm (Steps 9-18) runs in $\mathcal{O}(n\tau)$ time. The outer loop performs a single pass through the data, which takes $\mathcal{O}(n)$ time, and the inner loop (Steps 11-15) counts the number of occurrences of a symbol in a window around its current occurrence, which takes $\mathcal{O}(\tau)$ time. The variable ζ counts the number of occurrences of the symbol outside the window in Step 16, which can be implemented in $\mathcal{O}(1)$ time on Python. Finally, the algorithm returns the vector of count probabilities of the WINGIT estimator and runs in $\mathcal{O}(n\tau)$ time.

3.2 Plug-in estimator

Recall that we denote by φ_ζ the number of symbols appearing ζ times in X^n. Then, the plug-in (or empirical) estimator assigns the following total mass to all symbols with multiplicity ζ:

$$\widehat{M}_{\text{PI},\zeta} = \varphi_\zeta \cdot \frac{\zeta}{n}. \tag{10}$$

Clearly, the plug-in estimator (simultaneously for all values of ζ) is linear-time implementable.

When a particular symbol occurs sufficiently many times, the law of large numbers guarantees that the relative frequency of the symbol in the sample X^n provides a reliable estimate of its stationary probability. Therefore, we expect this estimator to be accurate for sufficiently large frequencies ζ. Conversely, when ζ is small, the plug-in estimator $\widehat{M}_{\text{PI},\zeta}$ will typically underestimate M_ζ^π.

3.3 Combined estimator

To leverage the complementary strengths of the WINGIT estimator (which we expect to be inaccurate for large ζ) and the plug-in estimator (which we expect to be inaccurate when ζ is small), we define a hybrid estimator that switches between the WINGIT estimator and the plug-in estimator based on the symbol multiplicity ζ. Specifically, our estimator takes the following form:

$$\widehat{M}_\zeta(\tau; \overline{\zeta}) = \begin{cases} \nu^{-1} \cdot \widehat{M}_{\text{WINGIT},\zeta}(\tau) & \text{if } \zeta \leq \overline{\zeta} \\ \nu^{-1} \cdot \widehat{M}_{\text{PI},\zeta} & \text{if } \zeta > \overline{\zeta}, \end{cases} \tag{11}$$

where $\overline{\zeta} \in \mathbb{N}$ is some parameter to be chosen, and ν is chosen to be a normalizing constant to ensure that $\sum_{\zeta=0}^n \widehat{M}_\zeta(\tau) = 1$. Note that $\overline{\zeta}$ determines the *transition point*, or the threshold for transitioning from the WINGIT to the plug-in estimator, and we make a specific choice of $\overline{\zeta}$ in stating our theorem to follow. When $\overline{\zeta}$ and τ are clear from context, we often use the shorthand $\widehat{M}_\zeta := \widehat{M}_\zeta(\tau; \overline{\zeta})$ to denote the scalar estimate and $\widehat{M} := (\widehat{M}_\zeta(\tau; \overline{\zeta}))_{\zeta=0}^n$ to denote its vector counterpart.

The combined estimator can be implemented in linear time using the linear time implementation of the plug-in estimator and that of the WINGIT estimator in Algorithm 1.

4 Main results

Our main result is a bound on the TV error attained by our hybrid estimator $\widehat{M}(\tau; \overline{\zeta})$.

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

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

A few remarks are in order. First, in the special case of an i.i.d. sequence we have $t_{\text{mix}}(n^{-5}) = 1$, so we can set $\tau = 1$ and our proposed estimator reduces to the combination of the Good–Turing and plug-in estimators studied in Acharya et al. (2013). In this case, Theorem 1 recovers the TV guarantee of Acharya et al. (2013), but without requiring Poissonization (see the discussion below).

Second, note that by the discussion in Section 2.1, we can always choose the window size $\tau = \log(n^5 \mu)/\log(1/\rho)$ in a general mixing stochastic process satisfying Eq. (1b). This parameter depends logarithmically on n and also depends on the mixing parameters μ and ρ, and yields the bound (ignoring $\log \log$ factors)

$$\mathbb{E}_{X^n}\left[d_{\text{TV}}(M^\pi(X^n), \widehat{M}(\tau; \overline{\zeta})) \right] \lesssim \left(\frac{\log^3(n^5\mu)/\log^3(1/\rho)}{n} \right)^{1/6}$$

uniformly over all exponentially α-mixing stochastic processes satisfying Eq. (1b).

Third, we highlight that our proof technique necessarily departs from the ones in Acharya et al. (2013) and Orlitsky and Suresh (2015), which rely on Poissonization. Poissonization essentially translates a sequence of fixed length into a sequence where the total number of samples itself is an independent Poisson random variable, which in turn renders the multiplicity of each individual symbol independent. While powerful, Poissonization requires independent random variables in the original sequence, and can no longer be used in our setting. Instead, our techniques to bound the error of the WINGIT estimator are based on the leave-a-window-out interpretation of the estimator, and careful control of its error in terms of appropriately defined indicator random variables. To handle the error of the plug-in estimator, we use a careful blocking technique and prove a self-normalized concentration inequality for mixing sequences — see Lemma 2. Our control on the error of both estimators is in terms of fine-grained quantities — see Propositions 1 and 2.

As a side remark, we note that Acharya et al. (2013) and Orlitsky and Suresh (2015) also analyze the error of their estimator in KL-divergence, and show that it achieves a rate of $\mathcal{O}(n^{-1/3})$ for IID sequences. By applying the reverse Pinsker inequality (Sason, 2015, Theorem 1) to Theorem 1 above, one can show that a minor modification of our estimator[3] attains KL risk $\widetilde{\mathcal{O}}((\tau^3/n)^{1/6})$ for general mixing sequences. While consistent, this rate is suboptimal, and proving a faster rate in KL error for our estimator remains an open problem.

As mentioned above, the proof of our main result relies on a concentration inequality for a self-normalized statistic of exponentially α-mixing sequences, which we state below along with a more standard Bernstein-type inequality. We prove the lemma in Section 6.3.

Lemma 2. *Fix $\delta, \epsilon > 0$ and let $U^n := (U_1, U_2, \ldots, U_n)$ denote a sequence of random variables from an ergodic, stochastic process that is exponentially α-mixing with parameters μ and ρ, with each $U_j \in [0, B]$ almost surely. Recall the definition of mixing time from Eq. (2).*

---

[3]To apply the reverse Pinsker inequality, we need to ensure that our estimator does not assign 0 probability to any ζ. This can be done by adding 1 to the total count of each ζ, ensuring that $\widehat{M}_\zeta(\tau; \overline{\zeta}) > 1/n$ for all ζ.

(a) For each fixed $\tau \geq \mathsf{t}_{\mathsf{mix}}(\epsilon/n)$, the following Bernstein-type inequality holds:

$$\left| \mathbb{E}[U_1] - \frac{1}{n} \sum_{j=1}^{n} U_j \right| \leq \sqrt{\frac{4\tau v^2 \log(1/\delta)}{n}} + \frac{4B\tau \log(1/\delta)}{3n}$$

with probability at least $1 - 4\delta - \epsilon$. Here $v^2 := \frac{1}{\tau^2} \mathsf{var}\left(\sum_{j=1}^{\tau} U_j \right)$ is the normalized block variance.

(b) In addition, if $\sum_{j=1}^{n} U_j \geq 36B \log(2/\delta) \cdot \mathsf{t}_{\mathsf{mix}}(\epsilon/n)$ and $n \geq 24\mathsf{t}_{\mathsf{mix}}(\epsilon/n)$, then the following holds:

$$\left| \mathbb{E}[U_1] - \frac{1}{n} \sum_{j=1}^{n} U_j \right| \leq 82 \cdot \frac{\sqrt{B \log(2/\delta) \cdot \left(\sum_{j=1}^{n} U_j \right) \cdot \mathsf{t}_{\mathsf{mix}}(\epsilon/n)}}{n}$$

with probability at least $1 - 10\delta - 2\epsilon$.

On the one hand, it is worth comparing the Bernstein bound in Lemma 2(a) to the Bernstein bound for exponentially α-mixing sequences given in Merlevède et al. (2009, Theorem 2). The authors in that paper show that if U^n is α-mixing and satisfies $\alpha(\tau) \leq \exp(-2c\tau)$, then defining

$$r^2 := \sup_i \left(\mathsf{var}(U_i) + 2 \sum_{j>i} |\mathsf{cov}(U_i, U_j)| \right),$$

we have the tail bound

$$\mathbb{P}\left(\left| \mathbb{E}[U_1] - \frac{1}{n} \sum_{j=1}^{n} U_j \right| > t \right) \leq \exp\left(-\frac{Cnt^2}{r^2 + \frac{B^2}{n} + B \log(n)^2 t} \right), \tag{12}$$

where C is an unspecified constant that depends on the parameter c defined in assuming a bound on $\alpha(\tau)$. By comparison, Lemma 2(a) makes the weaker α-mixing assumption (1a) and implies the following tail bound in terms of the mixing time:

$$\mathbb{P}\left(\left| \mathbb{E}[U_1] - \frac{1}{n} \sum_{j=1}^{n} U_j \right| > t \right) \leq 4 \exp\left(-\frac{nt^2}{4r^2 + \frac{4\tau B}{3} t} \right) + \epsilon. \tag{13}$$

In stating Eq. (13), we have used the fact that $\tau v^2 \leq r^2$, which is evident from the following sequence of calculations:

$$v^2 = \frac{1}{\tau^2} \mathsf{var}\left(\sum_{i=1}^{\tau} U_i \right) = \frac{1}{\tau^2} \left(\sum_{i=1}^{\tau} \mathsf{var}(U_i) + 2 \sum_{i=1}^{\tau} \sum_{j>i}^{\tau} \mathsf{cov}(U_i, U_j) \right) \leq \frac{\tau}{\tau^2} \sup_i \left(\mathsf{var}(U_i) + 2 \sum_{j>i} |\mathsf{cov}(U_i, U_j)| \right).$$

Comparing Eqs. (13) and (12), we see that our bound is stated explicitly in terms of mixing times and may be more user-friendly under the mixing assumption (1a).

On the other hand, Lemma 2(b) should be viewed as a concentration inequality on a self-normalized statistic constructed from a mixing sequence — note that the constants $(24, 36, 82)$ appearing in the theorem are not sharp and can likely be improved. A consequence of this result

10

— stated in a form that is typical in this literature (de la Peña et al., 2004) — is that for an exponentially α-mixing sequence U^n satisfying Eq. (1a) with $U_j \in [0, B]$ a.s. for all $j \in [n]$, we have

$$\mathbb{P}\left(\frac{\left| \mathbb{E}[U_1] - \frac{1}{n}\sum_{j=1}^n U_j \right|}{\sqrt{\frac{B\sum_{j=1}^n U_j}{n}}} > 82\sqrt{\frac{\log(2/\delta) \cdot \mathsf{t}_{\mathsf{mix}}(\epsilon/n)}{n}} \text{ and } \frac{B\sum_{j=1}^n U_j}{n} > \frac{36B^2 \log(2/\delta) \cdot \mathsf{t}_{\mathsf{mix}}(\epsilon/n)}{n} \right)$$

$$\leq 10\delta + 2\epsilon.$$

Such a self-normalized concentration inequality for exponentially α-mixing sequences may be of independent interest.

We conclude this section with the two key propositions that we prove, which provide bounds—for each ζ—on the error made by both the plug-in and WINGIT estimators in approximating M_ζ^π.

Proposition 1. *Consider any fixed $\delta, \epsilon > 0$ and let $\tau_0 := \mathsf{t}_{\mathsf{mix}}(\epsilon/n^2)$ for convenience. There is a universal positive constant C such that if $n \geq 24\tau_0$, then for each*

$$\zeta \geq \max\left\{ 36\tau_0 \log(22n/\delta), 1 + \sqrt{(4 + 8\tau_0 \log(11n/\delta)} + \frac{4\tau_0 \log(11n/\delta)}{3} \right\},$$

we have

$$\left| M_\zeta^\pi - \widehat{M}_{\mathrm{PI},\zeta} \right| \leq C\left(\frac{\sqrt{\zeta \mathsf{t}_{\mathsf{mix}}(\epsilon/n^2)} \cdot \varphi_\zeta(X^n)}{n} \sqrt{\log\left(\frac{Cn}{\delta}\right)} \right)$$

with probability at least $1 - \delta - 3\epsilon$.

Proposition 2. *There exists a universal positive constant C such that the following holds for all ζ. If the window size is chosen such that $\tau \geq \mathsf{t}_{\mathsf{mix}}(n^{-2})$, then we have*

$$\mathbb{E}\left[\left| M_\zeta^\pi(X^n) - \widehat{M}_{\mathrm{WINGIT},\zeta} \right| \right]$$

$$\leq C\sqrt{\frac{\tau}{n}}\left(\sqrt{\mathbb{E}[M_\zeta^\pi]} + \sqrt{\zeta \log(2\tau)\mathbb{E}[M_\zeta^\pi]} + \sqrt{\sum_{u=1}^{4\tau-2} \frac{(\zeta+u)}{u}\mathbb{E}\left[M_{\zeta+u}^\pi\right]} \right) + C\frac{(\zeta+1)\tau}{n}.$$

It is worth comparing Proposition 2 to the analogous bound for missing mass with $\zeta = 0$ (Pananjady et al., 2024). Substituting $\zeta = 0$ in the above bound and ignoring the final (lower order) term, we have

$$\mathbb{E}\left[\left| M_0^\pi(X^n) - \widehat{M}_{\mathrm{WINGIT},0} \right| \right] \lesssim \sqrt{\frac{\tau}{n}}\left(\sqrt{\mathbb{E}[M_0^\pi]} + \sqrt{\sum_{u=1}^{4\tau-2} \mathbb{E}\left[M_u^\pi\right]} \right).$$

Because $\sum_{u=1}^{4\tau-2} \mathbb{E}\left[M_u^\pi\right] \leq 1$, this recovers the minimax rate of $\mathcal{O}(\sqrt{\tau/n})$ given by Chandra et al. (2022). Moreover, the bound above is strictly stronger, since the RHS can be significantly smaller when the quantities $\{\mathbb{E}[M_u^\pi]\}_{u=0}^{4\tau-2}$ are small. Note that these are the expected count probability masses for the stationary distribution π — if the small count mass of the stationary distribution is small (i.e. most symbols appear many times), then our bound shows that WINGIT incurs a smaller error for missing mass estimation. This adaptivity is a distinct advantage of our approach. In contrast, the MSE bound provided in Pananjady et al. (2024) captures the correct scaling in terms of τ and n, but is not adaptive to this specific behavior of the stationary distribution π. This adaptivity property plays a central role in our final bound in Theorem 1, since the errors incurred across all the ζ values are aggregated to produce the claimed bound on TV error.

5 Numerical experiments

In this section, we provide a set of simulations on synthetically constructed Markov chains in order to demonstrate the effectiveness of the proposed estimator. We first describe the experimental set-up and discuss some of the estimators used in practice. We then compare the performance of these estimators against the proposed hybrid estimator.

5.1 Experimental Setup

We perform experiments on samples that we generate from different types of α-mixing processes. Specifically, we instantiate the stationary distribution of the mixing sequence from one of five different distributions. These distributions include: the uniform distribution, a step distribution with half of the symbols occurring with probability $1/(2|\mathcal{X}|)$ and the other half of symbols occurring with probability $3/(2|\mathcal{X}|)$, a Dirichlet distribution with the concentration parameter 1.5, and Zipf distributions with parameters 1.1 and 1.5. The Zipf distribution with parameter a assigns the probability $p(i) \propto i^{-a}$ for all $i \in \mathbb{N}$. We then layer a *random duplication model* on the i.i.d. samples generated from the stationary distribution to obtain the final α-mixing sequence. This process is described below. For notational convenience, we define $\mathsf{T}_{\mathsf{mix}} := \mathsf{t}_{\mathsf{mix}}(1/4)$, where $\mathsf{t}_{\mathsf{mix}}(\epsilon)$ is defined in Eq. (2).

1. First, we generate n_0 i.i.d. samples from the underlying stationary distribution. The number of symbols is set as $|\mathcal{X}| = n_0$. Thus, as we increase the number of samples, the number of symbols increases proportionally.

2. We then use a geometric duplication model in order to generate a mixing sequence from the i.i.d. samples. Specifically, for each sample in the i.i.d. sequence, we draw a geometric random variable with mean n_0^c (for some $c > 0$), which determines how many times that symbol is duplicated in the final sequence. This processcan be viewed as a draw from a sticky Markov chain (Chandra et al., 2022) in which the probability of observing an i.i.d. sample from the stationary distribution is given by $1/n_0^c$. The mixing time, $\mathsf{T}_{\mathsf{mix}}$, for this sequence is given by $\mathsf{T}_{\mathsf{mix}} = \mathcal{O}(n_0^c)$ (Pananjady et al., 2024). Since the duplication factors are random, the total length of the resulting sequence, denoted as n, is also a random variable with expected value $\mathbb{E}[n] = n_0^{c+1}$.

We compare the performance of our proposed estimator with several existing estimators that were originally designed for i.i.d. sequences. Specifically, we include the Good–Turing + Plug-in estimator analyzed by Acharya et al. (2013), as well as a family of popular add-β estimators defined as:

$$\widehat{M_\zeta}(\beta) = \frac{\varphi_\zeta(\zeta + \beta(\zeta))}{\nu},$$

where ν is a normalizing constant chosen to ensure that the estimated probabilities sum to 1. Different choices of the function $\beta(\zeta)$ yield well-known estimators:

- **Laplace estimator** (Laplace, 1814): $\beta(\zeta) = 1$ for all ζ,

- **Krichevsky–Trofimov estimator** (Krichevsky and Trofimov, 1981): $\beta(\zeta) = 0.5$ for all ζ,

- **Braess–Sauer estimator** (Braess and Sauer, 2004): $\beta(0) = 0.5$, $\beta(1) = 1$, and $\beta(\zeta) = 0.75$ for all $\zeta > 1$.

An important detail is that the normalization constant ν for the add-β estimators depends on the alphabet size $|\mathcal{X}|$. For the uniform, step, and Dirichlet distributions, the alphabet size is finite and set to $|\mathcal{X}| = n_0$. However, for the Zipf distribution, the alphabet is the set of all natural numbers, so that $|\mathcal{X}| = \infty$. To address this in our simulations, we approximate the alphabet size for the Zipf distribution using the number of unique symbols observed in the samples. This approximation is reasonable because the missing mass—corresponding to mass placed on unobserved symbols—is small with high probability for Zipf distribution and has a negligible effect on the TV distance.

We compare estimator performance by plotting the TV distance between the true and estimated vectors of count probabilities for each method. We present simulation results for three different values of c, the parameter controlling the mixing time of the sequences. All plots are averaged over 50 independent instances, and the shaded regions around each curve represent the standard deviation across these runs.

5.2   Results

First, in Figure 1 we plot the results for the case $c = 0.0$. Since the duplication factor for each sample in this case is simply 1, this represents the case of executing the estimators on an i.i.d. sequence and is a convenient reference point for evaluation. In this case, the WINGIT+Plug-in estimator is exactly equal to the Good–Turing+Plug-in estimator and thus the performance of these estimators overlap; we only include the evaluation of the WINGIT+Plug-in estimator to avoid redundancy. This estimator performs well on all distributions and is significantly better than the add-β estimators on all base distributions.

Figure 2 shows the results for $c = 0.2$, where the data is no longer i.i.d. The window size for the WINGIT+Plug-in estimator is set to $\tau = 10 \times n_0^c$, which is on the order of the mixing time of the sequence. In this setting, the WINGIT+Plug-in estimator continues to outperform all competing methods across all stationary distributions considered. It exhibits consistent estimation behavior, with the TV distance from the true count probabilities decreasing steadily as the number of samples increase. In contrast, the add-β estimators fail to achieve consistent estimation across the different base distributions.

Finally, Figure 3 presents the results for $c = 0.33$, where the data is even more dependent. Once again, we observe that the WINGIT+Plug-in estimator significantly outperforms the other estimators, highlighting its effectiveness in dealing with temporal dependencies in the data.

6   Proofs

We now provide proofs of all our results. We begin by recalling and defining notation. Let $n_0 = n/\tau$, and assume for readability that τ divides n. Recall the index sets $\{\mathcal{D}_i\}_{i=1}^n$ and $\{\mathcal{I}_i\}_{i=1}^n$ from Eq. (8). For $j \in [n_0]$ and $\ell = 0, 1, \ldots, 2\tau-1$, define the sets $\mathcal{B}_{j,\ell} := \mathcal{D}_{2\tau j-\ell}$ and $\mathcal{H}_{j,\ell} := \mathcal{I}_{2\tau j-\ell}$. Mnemonically, the set $\mathcal{B}_{j,\ell}$ represents a "block" around the $2\tau j - \ell$-th index of the sequence, and $\mathcal{H}_{j,\ell}$ represents the corresponding "hole", which contains all indices but with the block $\mathcal{B}_{j,\ell}$ removed. We drop floor and ceiling notation for readability, and assume divisibility of particular integers when convenient.
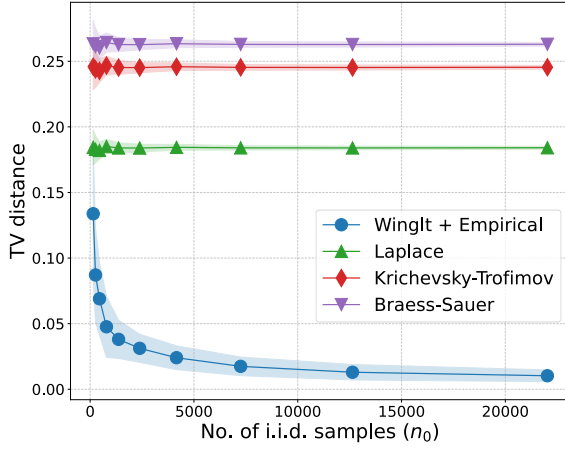
(a) Uniform
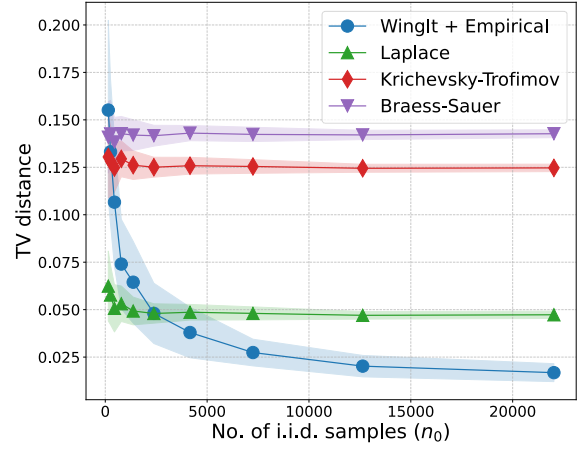
(b) Dirichlet
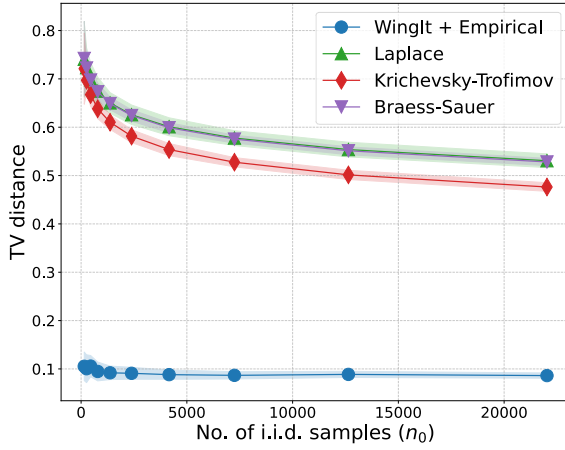
(c) Zipf with parameter 1.1

(d) Zipf with parameter 1.5

(e) Step

Figure 1: $c = 0.0$ corresponding to the IID sequence for n_0 samples, averaged over 50 instances.

14
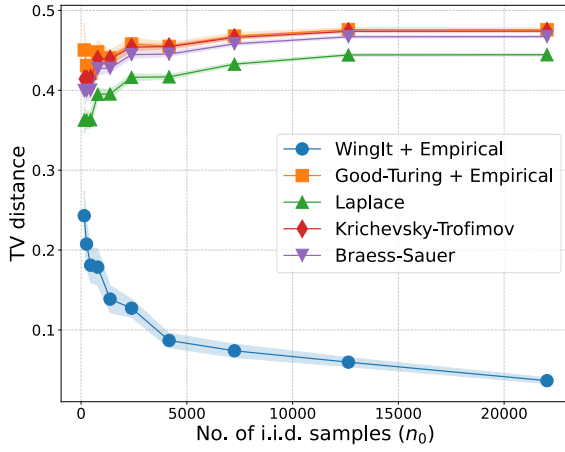
(a) Uniform

(b) Dirichlet

(c) Zipf with parameter 1.1

(d) Zipf with parameter 1.5

(e) Step

Figure 2. $c = 0.2$ corresponding to a mixing sequence with mixing time $\mathsf{T}_{\mathsf{mix}} = \mathcal{O}(n_0^{0.2})$ and expected number of samples $\mathbb{E}[n] = n_0^{1.2}$, averaged over 50 instances.
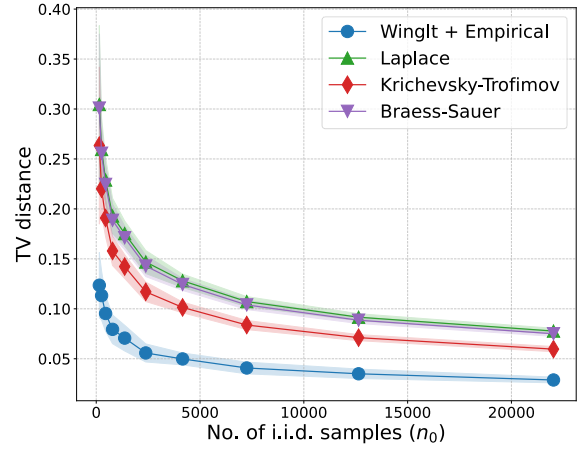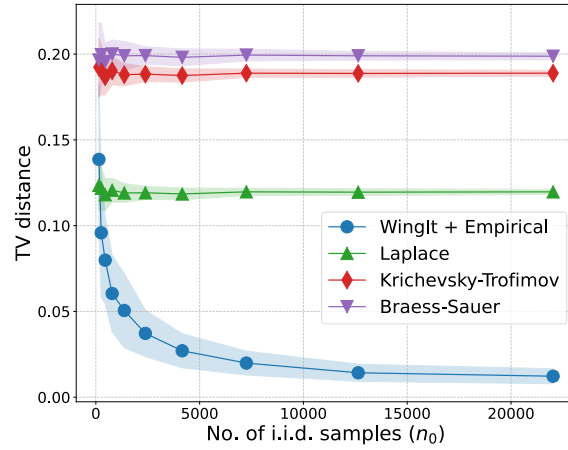
(a) Uniform

(b) Dirichlet
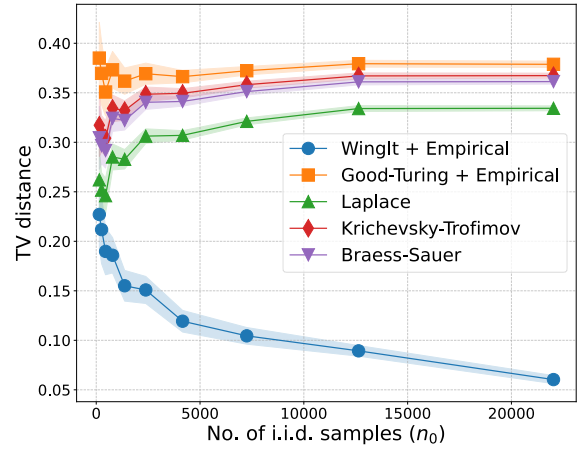
(c) Zipf with parameter 1.1

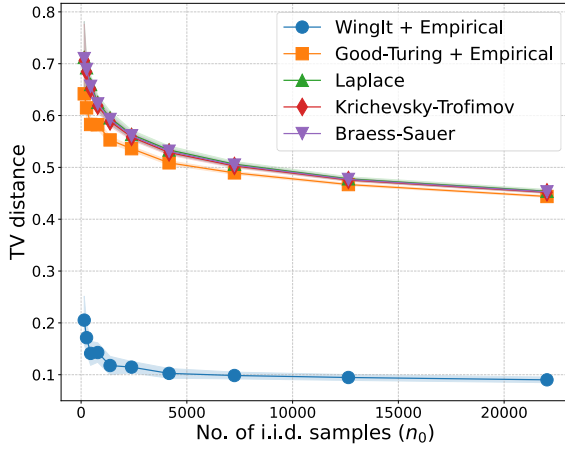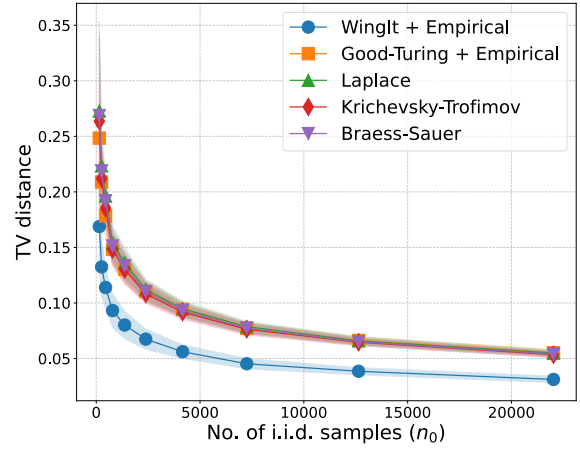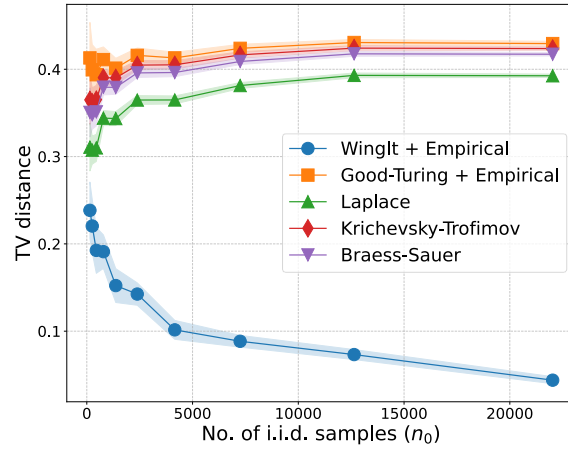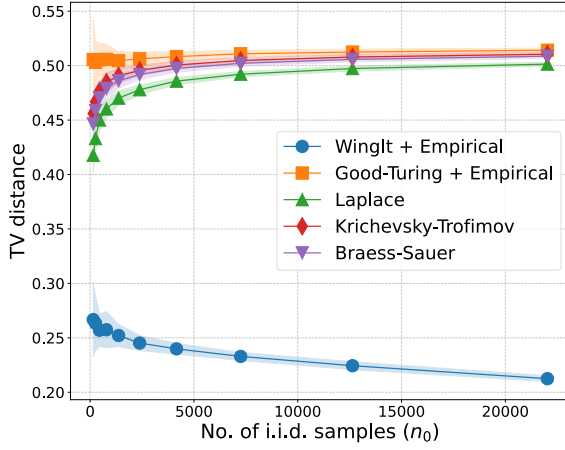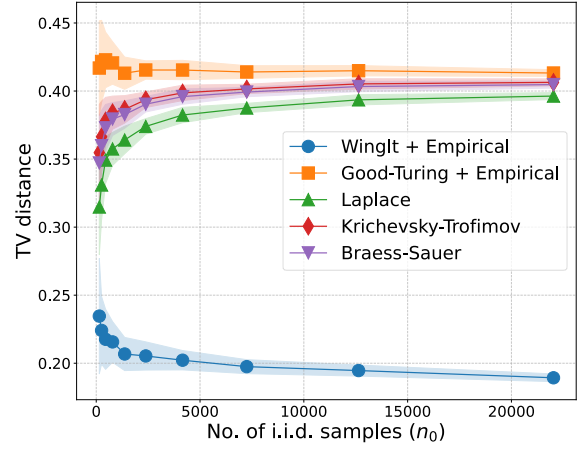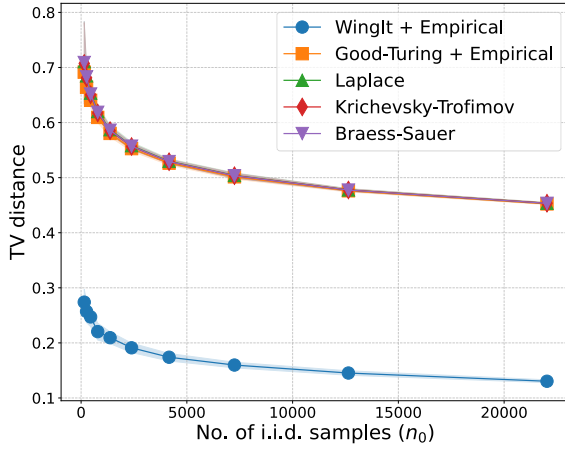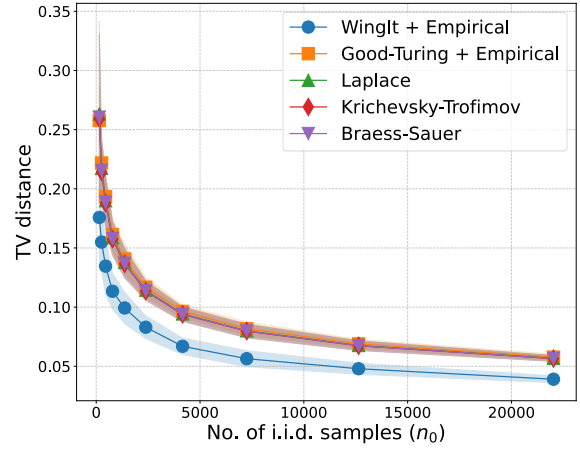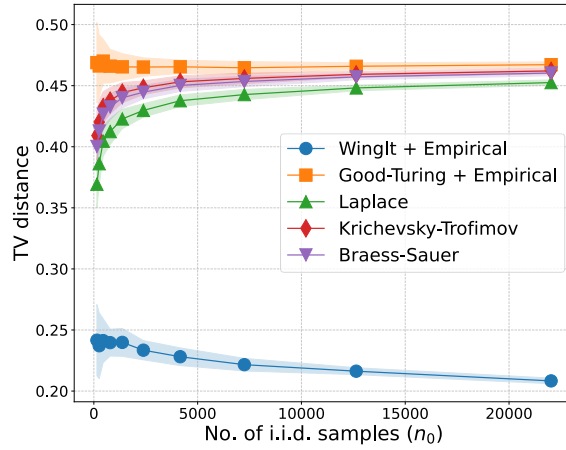(d) Zipf with parameter 1.5

(e) Step

Figure 3. $c = 0.2$ corresponding to a mixing sequence with mixing time $\mathsf{T}_{\mathsf{mix}} = \mathcal{O}(n_0^{0.33})$ expected number of samples $\mathbb{E}[n] = n_0^{1.33}$, averaged over 50 instances.

6.1 Proof of Lemma 1

Recall that the estimator $\widehat{q}$ is defined as $\widehat{q}_x = \widehat{M}_{N_x}/\varphi_{N_x}$ for all $x \in \mathcal{X}$. Also define the following *oracle-aided* natural estimator, which has knowledge of the stationary distribution π:

$$q_x^\pi = \frac{M_{N_x}^\pi}{\varphi_{N_x}} \quad \text{for all } x \in \mathcal{X}. \tag{14}$$

The estimator q^π calculates each count probability mass of the stationary distribution π and distributes it equally among all symbols having the same count in X^n. Denote the optimal natural estimator with respect to the TV error as

$$q^* = \arg \inf_{q \in \mathcal{Q}^{\mathsf{nat}}} d_{\mathsf{TV}}(\pi, q),$$

and for each $\zeta = 0, 1, \ldots, n$, define $M_\zeta^* = \sum_{x:N_x=\zeta} q_x^*$. Let $M^* \in \Delta(\{0, 1, \ldots, n\})$ denote the corresponding vector of count masses. Using the triangle inequality for TV distance, we have

$$d_{\mathsf{TV}}(\pi, \widehat{q}) \le d_{\mathsf{TV}}(\pi, q^\pi) + d_{\mathsf{TV}}(\widehat{q}, q^\pi). \tag{15}$$

We also have the following:

$$d_{\mathsf{TV}}(\pi, q^\pi) \le d_{\mathsf{TV}}(\pi, q^*) + d_{\mathsf{TV}}(q^*, q^\pi) \overset{\text{(i)}}{=} d_{\mathsf{TV}}(\pi, q^*) + d_{\mathsf{TV}}(M^*(X^n), M^\pi(X^n)),$$

where step (i) follows from Lemma 7 in the appendix, since both q^* and q^π are natural estimators. Finally, we have the sequence of inequalities

$$
\begin{aligned}
d_{\mathsf{TV}}(M^\pi(X^n), M^*(X^n)) &= \frac{1}{2} \sum_{u=0}^{n} |M_u^\pi(X^n) - M_u^*(X^n)| \\
&= \frac{1}{2} \sum_{u=0}^{n} \left| \sum_{x \in \mathcal{X}} \mathbb{I}\{N_x = u\}(\pi_x - q_x^*) \right| \\
&\le \frac{1}{2} \sum_{u=0}^{n} \sum_{x \in \mathcal{X}} \mathbb{I}\{N_x = u\} |\pi_x - q_x^*| \\
&= \frac{1}{2} \sum_{x \in \mathcal{X}} \left(|\pi_x - q_x^*| \sum_{u=0}^{n} \mathbb{I}\{N_x = u\} \right) \\
&\overset{\text{(i)}}{=} \frac{1}{2} \sum_{x \in \mathcal{X}} |\pi_x - q_x^*| = d_{\mathsf{TV}}(\pi, q^*),
\end{aligned}
$$

where step (i) follows because $\sum_{t=0}^{n} \mathbb{I}\{N_x = t\} = 1$. Putting together the pieces, we obtain

$$d_{\mathsf{TV}}(\pi, \widehat{q}) \le 2 \cdot d_{\mathsf{TV}}(\pi, q^*) + d_{\mathsf{TV}}(\widehat{q}, q^\pi) \overset{\text{(i)}}{=} 2 \cdot d_{\mathsf{TV}}(\pi, q^*) + d_{\mathsf{TV}}(\widehat{M}, M^\pi),$$

where step (i) follows since both $\widehat{q}$ and q^π are natural estimators and so we can again appeal to Lemma 7. This completes the proof of Lemma 1. $\square$

6.2 Proof of Theorem 1

We prove Theorem 1 from Propositions 1 and 2, which in turn are proved in Sections 6.4 and 6.5. Denote $\widehat{M}^{\mathrm{U}}$ as the *unnormalized* estimator, i.e. $\widehat{M}^{\mathrm{U}} := \widehat{M}(\tau; \overline{\zeta}) \cdot \nu$, with

$$\widehat{M}^{\mathrm{U}}_\zeta = \begin{cases} \widehat{M}_{\mathrm{WINGIT},\zeta}(\tau) & \text{if } \zeta \leq \overline{\zeta} \\ \widehat{M}_{\mathrm{PI},\zeta} & \text{if } \zeta > \overline{\zeta}, \end{cases}$$

We have

$$d_{\mathsf{TV}}(M^\pi(X^n), \widehat{M}(\tau; \overline{\zeta})) \overset{(i)}{\leq} \|M^\pi - \widehat{M}^{\mathrm{U}}\|_1$$

$$= \sum_{\zeta=0}^{\overline{\zeta}} \left| M_\zeta^\pi - \widehat{M}_{\mathrm{WINGIT},\zeta}(\tau) \right| + \sum_{\zeta=\overline{\zeta}+1}^{n} \left| M_\zeta^\pi - \widehat{M}_{\mathrm{PI},\zeta} \right|,$$

where step (i) follows from Lemma 8 in the appendix. The following algebraic inequalities will be useful in our bounds going forward:

$$\sum_{\zeta=\overline{\zeta}+1}^{n} \varphi_\zeta \overset{(a)}{\leq} \frac{n}{\overline{\zeta}+1} \quad \text{and} \quad \sum_{\zeta=\overline{\zeta}+1}^{n} \varphi_\zeta \sqrt{\zeta} \overset{(b)}{\leq} \frac{n}{\sqrt{\overline{\zeta}+1}}. \tag{16}$$

We defer a proof of Eq. (16) to the end of the section. We also assume that $n \gtrsim \tau^3 \log^3(Cn)$ for the rest of this section; the theorem is trivially true otherwise since the TV is always bounded above by 1.

Bounding WingIt error: Applying Proposition 2, we have

$$\sum_{\zeta=0}^{\overline{\zeta}} \mathbb{E}\left[\left| M_\zeta^\pi(X^n) - \widehat{M}_{\mathrm{WINGIT},\zeta} \right| \right]$$

$$\lesssim \sum_{\zeta=0}^{\overline{\zeta}} \sqrt{\frac{\tau}{n}} \left(\sqrt{\mathbb{E}[M_\zeta^\pi]} + \sqrt{\zeta \log(2\tau) \mathbb{E}\left[\mathbb{I}\left\{N_Y = \zeta\right\}\right]} + \sqrt{\sum_{u=1}^{4\tau-2} \frac{(\zeta+u)}{u} \mathbb{E}\left[M_{\zeta+u}^\pi\right]} \right) + \sum_{\zeta=0}^{\overline{\zeta}} \frac{(\zeta+1)\tau}{n}$$

$$\leq \sqrt{\frac{\tau}{n}(\overline{\zeta}+1)} + \sqrt{\frac{\tau}{n}} \underbrace{\sum_{\zeta=0}^{\overline{\zeta}} \sqrt{\zeta \log(2\tau) \mathbb{E}\left[\mathbb{I}\left\{N_Y = \zeta\right\}\right]}}_{E_1} + \sqrt{\frac{\tau}{n}} \underbrace{\sum_{\zeta=0}^{\overline{\zeta}} \sqrt{\sum_{u=1}^{4\tau-2} \frac{(\zeta+u)}{u} \mathbb{E}\left[M_{\zeta+u}^\pi\right]}}_{E_2} + \frac{(\overline{\zeta}+1)^2 \tau}{n}.$$

The first term in the above inequality is lower order compared to E_1 and E_2. Thus we now focus on providing the bounds for E_1 and E_2 separately. We have

$$E_1 = \sum_{\zeta=0}^{\overline{\zeta}} \sqrt{\zeta \log(2\tau) \mathbb{E}\left[\mathbb{I}\left\{N_Y = \zeta\right\}\right]} \overset{(i)}{\leq} (\overline{\zeta}+1) \sqrt{\log(2\tau)},$$

18

where step (i) follows because the Cauchy-Schwarz inequality and the fact $\sum_{\zeta=0}^{\overline{\zeta}} \mathbb{E}\left[\mathbb{I}\left\{N_Y = \zeta\right\}\right] \leq 1$. For the term E_2, we have

$$
\begin{aligned}
E_2 &= \sum_{\zeta=0}^{\overline{\zeta}} \sqrt{\sum_{u=1}^{4\tau-2} \frac{(\zeta+u)}{u} \mathbb{E}\left[M_{\zeta+u}^{\pi}\right]} \\
&= \sum_{\zeta=0}^{\overline{\zeta}} \sqrt{\mathbb{E}\left[\sum_{u=1}^{\lceil 4\tau-2\rceil} \mathbb{I}\left\{N_Y = \zeta+u\right\}\frac{(\zeta+u)}{u}\right]} \\
&\overset{(i)}{\leq} \sqrt{\sum_{\zeta=0}^{\overline{\zeta}} 1} \cdot \sqrt{\sum_{\zeta=0}^{\overline{\zeta}} \mathbb{E}\left[\sum_{u=1}^{\lceil 4\tau-2\rceil} \mathbb{I}\left\{N_Y = \zeta+u\right\}\frac{(\zeta+u)}{u}\right]}
\end{aligned}
$$

where step (i) follows from the Cauchy–Schwarz inequality. Continuing, we have

$$
\begin{aligned}
E_2 &\leq \sqrt{\overline{\zeta}+1} \cdot \sqrt{\mathbb{E}\left[\sum_{u=1}^{\lceil 4\tau-2\rceil} (1/u) \sum_{\zeta=0}^{\overline{\zeta}} \mathbb{I}\left\{N_Y = \zeta+u\right\}(\zeta+u)\right]} \\
&\leq \sqrt{\overline{\zeta}+1} \cdot \sqrt{\mathbb{E}\left[\sum_{u=1}^{\lceil 4\tau-2\rceil} \frac{(\overline{\zeta}+u)}{u}\right]} \leq \sqrt{(\overline{\zeta}+1)(\overline{\zeta}+4\tau-2)\log(4\tau)}.
\end{aligned}
$$

Putting together the pieces, we have

$$
\sum_{\zeta=0}^{\overline{\zeta}} \mathbb{E}\left[\left|M_{\zeta}^{\pi}(X^n) - \widehat{M}_{\text{WINGIT},\zeta}\right|\right] \lesssim (\overline{\zeta}+1)\sqrt{\frac{\tau \log(2\tau)}{n}} + \sqrt{\frac{\tau}{n}}\sqrt{(\overline{\zeta}+1)(\overline{\zeta}+4\tau-2)\log(4\tau)} + \frac{(\overline{\zeta}+1)^2\tau}{n}.
$$

In the statement of Theorem 1, we set the value $\overline{\zeta}+1 = n^{1/3}$. Since we have assumed that $n^{1/3} \gtrsim \tau \log(Cn)$, we can write $(\overline{\zeta}+4\tau-2) \lesssim \overline{\zeta}+1$ and substituting above yields

$$
\sum_{\zeta=0}^{\overline{\zeta}} \mathbb{E}\left[\left|M_{\zeta}^{\pi}(X^n) - \widehat{M}_{\text{WINGIT},\zeta}\right|\right] \lesssim (\overline{\zeta}+1)\sqrt{\frac{\tau \log(4\tau)}{n}} + \frac{(\overline{\zeta}+1)^2\tau}{n}.
$$

Bounding plug-in error: For bounding the error of the plug-in estimator one thing to note is that we use this estimator only for $\zeta \geq \overline{\zeta}+1 = n^{1/3}$. Applying the statement of Proposition 1, the following holds for each $\zeta \geq \overline{\zeta}+1$ provided that $\tau \geq \mathsf{t}_{\text{mix}}\left(\frac{\epsilon}{n^2}\right)$:

$$
\left|M_{\zeta}^{\pi}(X^n) - \widehat{M}_{\text{PI},\zeta}\right| \lesssim \frac{\sqrt{\tau\zeta} \cdot \varphi_{\zeta}(X^n)}{n}\sqrt{\log\left(\frac{Cn}{\delta}\right)}.
$$

probability at least $1 - \delta - 3\epsilon$.

Taking the union bound of the above inequality for all $\zeta \geq \overline{\zeta}+1$ we have

$$
\left|M_{\zeta}^{\pi}(X^n) - \widehat{M}_{\text{PI},\zeta}\right| \lesssim \frac{\sqrt{\tau\zeta} \cdot \varphi_{\zeta}(X^n)}{n}\sqrt{\log\left(\frac{Cn}{\delta}\right)}.
$$

19

probability at least $1 - n\delta - 3n\epsilon$. Substituting the above inequality into the error for the plug-in estimator we have

$$\sum_{\zeta=\overline{\zeta}+1}^{n} \left| M_{\zeta}^{\pi}(X^n) - \widehat{M}_{\mathrm{PI},\zeta} \right| \lesssim \sum_{\zeta=\overline{\zeta}+1}^{n} \frac{\sqrt{\tau\zeta} \cdot \varphi_{\zeta}(X^n)}{n} \sqrt{\log\left(\frac{Cn}{\delta}\right)}$$

$$\overset{(i)}{\leq} \sqrt{\frac{\tau}{\overline{\zeta}+1}} \sqrt{\log\left(\frac{Cn}{\delta}\right)},$$

probability at least $1 - n\delta - 3n\epsilon$. Above, step (i) follows from Eq. (16). Substituting $\epsilon' = 3n\epsilon$ and $\delta' = n\delta$, we have

$$\sum_{\zeta=\overline{\zeta}+1}^{n} \left| M_{\zeta}^{\pi}(X^n) - \widehat{M}_{\mathrm{PI},\zeta} \right| \lesssim \sqrt{\frac{\tau}{\overline{\zeta}+1}} \sqrt{\log\left(\frac{Cn^2}{\delta'}\right)},$$

with probability at least $1 - \delta' - \epsilon'$. Choosing $\epsilon' = \frac{1}{n^2}$ and then integrating the tail bound we obtain

$$\sum_{\zeta=\overline{\zeta}+1}^{n} \mathbb{E}\left| M_{\zeta}^{\pi}(X^n) - \widehat{M}_{\mathrm{PI},\zeta} \right| \lesssim \sqrt{\frac{\tau}{\overline{\zeta}+1}} \sqrt{\log(Cn^2)}. \tag{17}$$

Recall that Proposition 1 requires $\tau \geq \mathsf{t}_{\mathrm{mix}}(\epsilon/n^2)$. We set $\epsilon = \epsilon'/3n$ and finally substitute $\epsilon' = 1/n^2$. Thus we require the condition $\tau \geq \mathsf{t}_{\mathrm{mix}}(1/n^5)$.

Combining the pieces: Combining the two bounds, we have

$$\mathbb{E}\left[d_{\mathsf{TV}}(M^{\pi}(X^n), \widehat{M}(\tau))\right] \lesssim (\overline{\zeta}+1)\sqrt{\frac{\tau\log(4\tau)}{n}} + \frac{(\overline{\zeta}+1)^2\tau}{n} + \sqrt{\frac{\tau}{\overline{\zeta}+1}}\sqrt{\log(Cn^2)}.$$

Note that we have set $\overline{\zeta} + 1 = n^{1/3}$ and also that $n \geq C'\tau^3 \log^3(Cn)$ for a large enough constant C'. This allows us to simplify the above expression and obtain

$$\mathbb{E}\left[d_{\mathsf{TV}}(M^{\pi}(X^n), \widehat{M}(\tau))\right] \lesssim \frac{\sqrt{\tau\log(Cn)}}{n^{1/6}},$$

as claimed. We conclude by proving claim (16).

Proof of claim (16): First, note that $n = \sum_{\zeta=0}^{n} \zeta\varphi_{\zeta}$. Part (a) of the inequality then follows by noting that

$$n \geq \sum_{\zeta=\overline{\zeta}+1}^{n} \zeta\varphi_{\zeta} \geq (\overline{\zeta}+1) \sum_{\zeta=\overline{\zeta}+1}^{n} \varphi_{\zeta}. \tag{18}$$

To prove part (b), we use the Cauchy–Schwarz inequality to obtain

$$\sum_{\overline{\zeta}+1}^{n} \varphi_{\zeta}\sqrt{\zeta} = \sum_{\zeta=\overline{\zeta}+1}^{n} \sqrt{\varphi_{\zeta}}\sqrt{\zeta\varphi_{\zeta}} \leq \sqrt{\sum_{\zeta=\overline{\zeta}+1}^{n} \varphi_{\zeta}}\sqrt{\sum_{\zeta=\overline{\zeta}+1}^{n} \zeta\varphi_{\zeta}}.$$

20

Using Eq. (18) then yields

$$\sum_{\zeta=\bar{\zeta}+1}^{n} \varphi_\zeta \sqrt{\zeta} \le \sqrt{\frac{n}{\bar{\zeta}+1}} \cdot \sqrt{n} = \frac{n}{\sqrt{\bar{\zeta}+1}},$$

which completes the proof of the claim and therefore the proof of Theorem 1. $\qquad\square$

6.3   Proof of Lemma 2

We structure the proof as follows. Both parts of the lemma rely on a certain blocking argument, which we present first. Then, we prove each part of the lemma in turn. For the rest of this proof, suppose τ is some arbitrary integer, and assume for simplicity in exposition that τ divides n. In part (b), we will choose a particular value of τ.

Blocking argument: We start by dividing the entire sequence U^n into $n_0 = n/\tau$ intervals, each of length τ. Now we write the quantity of interest $\left| \mathbb{E}[U_1] - \frac{1}{n} \sum_{i=1}^{n} U_i \right|$ as a sum over disjoint even and odd-numbered blocks of the sequence. Let us denote the set of indices inside each of these intervals as S_k for $k \in [n_0]$. Note, by definition, that $|S_k| = \tau$ for all $k \in [n_0]$. We define the following portions of the original sequence U^n:

$$U_{\mathsf{odd}} := \bigoplus_{k \text{ odd}} \bigoplus_{j \in S_k} U_j, \quad \text{and}$$

$$U_{\mathsf{even}} := \bigoplus_{k \text{ even}} \bigoplus_{j \in S_k} U_j.$$

Using the triangle inequality, we have

$$\left| \mathbb{E}[U_1] - \frac{1}{n} \sum_{i=1}^{n} U_i \right| = \left| \frac{\mathbb{E}[U_1]}{2} - \frac{1}{n} \sum_{k \text{ odd}} \sum_{j \in S_k} U_j + \frac{\mathbb{E}[U_1]}{2} - \frac{1}{n} \sum_{k \text{ even}} \sum_{j \in S_k} U_j \right|$$

$$\le \left| \frac{\mathbb{E}[U_1]}{2} - \frac{1}{n} \sum_{k \text{ odd}} \sum_{j \in S_k} U_j \right| + \left| \frac{\mathbb{E}[U_1]}{2} - \frac{1}{n} \sum_{k \text{ even}} \sum_{j \in S_k} U_j \right|. \qquad (19)$$

We now construct a new auxiliary stochastic process as follows. We choose the tuple $(Y_1, \ldots, Y_\tau)$ independently of everything else, and according to the law of $(U_1, \ldots, U_\tau)$. Similarly, choose the blocks $(Y_{\tau+1}, \ldots, Y_{2\tau}), \ldots, (Y_{n-\tau+1}, \ldots, Y_n)$ IID from the same distribution. In words, at the start of each interval S_k, we restart the process from a state sampled from the stationary distribution π. Let us denote this sequence as

$$Y^n := (Y_j)_{j \in [n]}.$$

We again decompose the sequence Y^n over odd and even blocks similar to U^n.

$$Y_{\mathsf{odd}} = \bigoplus_{k \text{ odd}} \bigoplus_{j \in S_k} Y_j, \text{ and}$$

$$Y_{\mathsf{even}} = \bigoplus_{k \text{ even}} \bigoplus_{j \in S_k} Y_j.$$

Note that the stochastic processes U_{odd} and Y_{odd} are each of length $n/2$ and correspond directly to the definitions of W_j and W_j' in the statement of Lemma 13. Formally, we can apply Lemma 13, with the function $f : \mathcal{X}^{n/2} \to \{0, 1\}$ defined as below for an arbitrary sequence $V^{n/2} := (V_1, \ldots, V_{n/2})$:

$$f(V^{n/2}) := \mathbb{I}\left\{ \left| \mathbb{E}[V_1] - \frac{2}{n} \sum_{i=1}^{n/2} V_i \right| > t \right\}.$$

Applying the lemma yields

$$\mathbb{E}\left[f(U_{\mathsf{odd}}) \right] \leq \mathbb{E}\left[f(Y_{\mathsf{odd}}) \right] + \frac{n_0}{2} \mu \rho^\tau$$

for any $t \in \mathbb{R}$. Substituting the definition of f in the above inequality yields

$$\mathbb{P}\left(\left| \frac{\mathbb{E}[U_1]}{2} - \frac{1}{n} \sum_{k \text{ odd}} \sum_{j \in S_k} U_j \right| > \frac{t}{2} \right) \leq \mathbb{P}\left(\left| \frac{\mathbb{E}[Y_1]}{2} - \frac{1}{n} \sum_{k \text{ odd}} \sum_{j \in S_k} Y_j \right| > \frac{t}{2} \right) + \frac{n_0}{2} \mu \rho^\tau. \qquad (20)$$

An identical statement can be shown for the even subsequence.

Substituting Eq. (20) for odd and even sequences into Eq. (19) and applying the union bound, we obtain

$$\mathbb{P}\left(\left| \mathbb{E}[U_1] - \frac{1}{n} \sum_{j=1}^{n} U_j \right| > t \right) \leq \mathbb{P}\left(\left| \frac{\mathbb{E}[Y_1]}{2} - \frac{1}{n} \sum_{k \text{ odd}} \sum_{j \in S_k} Y_j \right| > \frac{t}{2} \right)$$

$$+ \mathbb{P}\left(\left| \frac{\mathbb{E}[Y_1]}{2} - \frac{1}{n} \sum_{k \text{ even}} \sum_{j \in S_k} Y_j \right| > \frac{t}{2} \right) + n_0 \mu \rho^\tau. \qquad (21)$$

We will now use Eq. (21) to prove each part of the lemma separately.

Proof of part (a): For notational convenience, define the random variables $J_k := \frac{1}{\tau} \sum_{j \in S_k} Y_j$ for $k \in [n_0]$. We collect some relevant properties of these random variables. First, since $Y_j \in [0, B]$, we have $J_k \in [0, B]$ for all $k \in [n_0]$. By construction of Y^n, the random variables $\{J_k\}_{k=1}^{n_0}$ are also independent and identically distributed. Moreover, because the process Y^n is initialized at the stationary distribution π, we have $\mathbb{E}[Y_1] = \mathbb{E}[J_1]$. Finally, recall that we defined $v^2 := \frac{1}{\tau^2} \mathsf{var}\left(\sum_{i=1}^{\tau} U_i \right)$ in the statement of the lemma. It is easy to verify that

$$v^2 := \frac{1}{\tau^2} \mathsf{var}\left(\sum_{i=1}^{\tau} U_i \right) = \frac{1}{\tau^2} \mathsf{var}\left(\sum_{i=1}^{\tau} Y_i \right) = \mathsf{var}(J_1) = \frac{2}{n_0} \sum_{k \text{ odd}} \mathsf{var}(J_k).$$

We now bound the two terms on the RHS of Eq. (21). Let us present the argument for the first (i.e. odd) term without loss of generality. We have

$$\left| \mathbb{E}[Y_1] - \frac{2}{n} \sum_{k \text{ odd}} \sum_{j \in S_k} Y_j \right| = \left| \mathbb{E}[J_1] - \frac{2}{n_0} \sum_{k \text{ odd}} J_k \right| \overset{\text{(i)}}{\leq} \sqrt{\frac{4 \tau v^2 \log(1/\delta)}{n}} + \frac{4 B \tau \log(1/\delta)}{3n},$$

22

where inequality (i) holds with probability at least $1 - 2\delta$ by the Bernstein inequality. Repeating the argument for the even subsequence and combining the bounds with Eq. (21), we obtain

$$\mathbb{P}\left(\left|\mathbb{E}[U_1] - \frac{1}{n}\sum_{j=1}^{n} U_j\right| > \sqrt{\frac{4\tau v^2 \log(1/\delta)}{n}} + \frac{4B\tau \log(1/\delta)}{3n}\right) \leq 4\delta + n_0\mu\rho^\tau.$$

For any $\tau \geq \mathsf{t}_{\mathsf{mix}}(\epsilon/n)$, we have $n_0\mu\rho^\tau \leq \epsilon$, so that

$$\left|\mathbb{E}[U_1] - \frac{1}{n}\sum_{j=1}^{n} U_j\right| \leq \sqrt{\frac{4\tau v^2 \log(1/\delta)}{n}} + \frac{4B\tau \log(1/\delta)}{3n},$$

with probability at least $1 - 4\delta - \epsilon$. This completes the proof of part (a).

Proof of part (b): For the proof of this part only, we abuse notation slightly and let $\tau = \mathsf{t}_{\mathsf{mix}}(\epsilon/n)$ for convenience, still letting $n_0 = n/\tau$. It is convenient to define the shorthand $L_k := \frac{1}{\tau B}\sum_{j \in S_k} Y_j = \frac{J_k}{B}$ for each $k \in [n_0]$. Since $Y_j \in [0, B]$, we have $L_k \in [0, 1]$ for all $k \in [n_0]$. As before, the collection of random variables $\{L_k\}_{k=1}^{n_0}$ is independent and identically distributed.

Using the empirical Bernstein bound from Maurer and Pontil (2009, Theorem 4), we have

$$\frac{1}{B}\left|\mathbb{E}[Y_1] - \frac{2}{n}\sum_{k \text{ odd}}\sum_{j \in S_k} Y_j\right| = \left|\mathbb{E}[L_1] - \frac{2}{n_0}\sum_{k \text{ odd}} L_k\right| \overset{(i)}{\leq} \sqrt{\frac{4\mathsf{sp}(L)\log(2/\delta)}{n_0}} + \frac{14\log(2/\delta)}{3(n_0 - 2)}, \qquad (22)$$

where step (i) holds with probability at least $1 - 2\delta$. Recall that we defined the spread as

$$\mathsf{sp}(L) = \frac{1}{(n_0/2)(n_0/2 - 1)}\sum_{\substack{p,q \text{ odd} \\ p<q}} (L_p - L_q)^2,$$

since the number of odd blocks is equal to $n_0/2$. We can upper bound $\mathsf{sp}(L)$ as below:

$$\mathsf{sp}(L) = \frac{4}{\tau^2 n_0(n_0 - 2)}\sum_{\substack{p,q \text{ odd} \\ p<q}}\left(\sum_{i \in S_p}\frac{Y_i}{B} - \sum_{j \in S_q}\frac{Y_j}{B}\right)^2$$

$$\overset{(i)}{\leq} \frac{4}{\tau n_0(n_0 - 2)}\sum_{\substack{p,q \text{ odd} \\ p<q}}\left(\sum_{i \in S_p}\frac{Y_i}{B} + \sum_{j \in S_q}\frac{Y_j}{B}\right)$$

$$\leq \frac{4}{\tau B(n_0 - 2)}\sum_{k \text{ odd}}\sum_{j \in S_k} Y_j.$$

where step (i) follows from Young's inequality and the fact that $Y_i/B \in [0, 1]$. Substituting expressions for $\mathsf{sp}(L)$, L_k and n_0 into Eq. (22), we obtain

$$\left|\frac{\mathbb{E}[Y_1]}{2} - \frac{1}{n}\sum_{k \text{ odd}}\sum_{j \in S_k} Y_j\right| \leq \frac{2\sqrt{\tau B \log(2/\delta)\sum_{k \text{ odd}}\sum_{j \in S_k} Y_j}}{n - 2\tau} + \frac{7\tau B \log(2/\delta)}{3(n - 2\tau)} \qquad (23)$$

23

with probability at least $1 - 2\delta$. Executing the same argument for the even subsequence, combining the pieces with the inequality $\sqrt{\sum_{\text{k odd}} \sum_{j \in S_k} Y_j} + \sqrt{\sum_{\text{k even}} \sum_{j \in S_k} Y_j} \leq \sqrt{2 \sum_{k=1}^{n_0} \sum_{j \in S_k} Y_j}$ and $2\sqrt{2} < 3$, and noting that $\sum_{k=1}^{n_0} \sum_{j \in S_k} Y_j = \sum_{j=1}^{n} Y_j$, we have

$$\left| \mathbb{E}[U_1] - \frac{1}{n} \sum_{j=1}^{n} U_j \right| \leq \frac{3\sqrt{\tau B \log(2/\delta) \sum_{j=1}^{n} Y_j}}{n - 2\tau} + \frac{14\tau B \log(2/\delta)}{3(n - 2\tau)} \tag{24}$$

with probability at least $1 - 4\delta - n_0 \mu \rho^\tau$.

Next, we claim for $\tau = \mathsf{t}_{\mathsf{mix}}(\epsilon/n)$ that if $\sum_{j=1}^{n} U_j \geq 36\tau B \log(2/\delta)$ and $n \geq 24\tau$, then the inequality

$$\sqrt{\sum_{j=1}^{n} Y_j} \leq 11 \sqrt{\sum_{j=1}^{n} U_j} + 14\sqrt{\tau B \log(2/\delta)}, \tag{25}$$

holds with probability at least $1 - 6\delta - n_0 \mu \rho^\tau$. We defer the proof of Eq. (25) to the end of this section.

Substituting Eq. (25) into Eq. (24), we obtain

$$\left| \mathbb{E}[U_1] - \frac{1}{n} \sum_{j=1}^{n} U_j \right| \leq \frac{33\sqrt{\tau B \log(2/\delta) \sum_{j=1}^{n} U_j}}{n - 2\tau} + \frac{140\tau B \log(2/\delta)}{3(n - 2\tau)},$$

with probability at least $1 - 10\delta - 2n_0 \mu \rho^\tau$. Now setting $\tau = \mathsf{t}_{\mathsf{mix}}(\epsilon/n)$, recall that in the statement of the lemma, we assumed the conditions $\sum_{j=1}^{n} U_j \geq 36B\tau \log(2/\delta)$ and $n \geq 24\tau$. From this, we obtain

$$\left| \mathbb{E}[U_1] - \frac{1}{n} \sum_{j=1}^{n} U_j \right| \leq \frac{734\sqrt{\tau B \log(2/\delta) \sum_{j=1}^{n} U_j}}{9n},$$

with probability at least $1 - 10\delta - 2n_0 \mu \rho^\tau$. For $\tau = \mathsf{t}_{\mathsf{mix}}(\epsilon/n)$, we have $n_0 \mu \rho^\tau \leq \epsilon$, and this completes the proof.

<u>Proof of claim (25)</u>: Executing the empirical Bernstein argument for the auxiliary sequence[4] Y^n we have

$$\left| \mathbb{E}[Y_1] - \frac{1}{n} \sum_{k=1}^{n_0} \sum_{j \in S_k} Y_j \right| \leq \frac{2\sqrt{\tau B \log(2/\delta) \sum_{k=1}^{n_0} \sum_{j \in S_k} Y_j}}{n - 2\tau} + \frac{7\tau B \log(2/\delta)}{3(n - 2\tau)},$$

with probability at least $1 - 2\delta$. Simplifying notation, we have

$$\left| \mathbb{E}[Y_1] - \frac{1}{n} \sum_{j=1}^{n} Y_j \right| \leq \frac{2\sqrt{\tau B \log(2/\delta) \sum_{j=1}^{n} Y_j}}{n - 2\tau} + \frac{7\tau B \log(2/\delta)}{3(n - 2\tau)}, \tag{26}$$

---

[4]Note that in this case, we do not need to split over odd and even subsequences; just a direct application of the empirical Bernstein bound suffices, and we obtain a slightly better constant.

Combing the Eq. (26) and Eq. (24) with the triangle inequality, we have

$$\left| \frac{1}{n} \sum_{j=1}^n Y_j - \frac{1}{n} \sum_{j=1}^n U_j \right| \leq \frac{5\sqrt{\tau B \log(2/\delta) \sum_{j=1}^n Y_j}}{n - 2\tau} + \frac{7\tau B \log(2/\delta)}{n - 2\tau}, \tag{27}$$

with probability at least $1 - 6\delta - n_0 \mu \rho^\tau$.

At the same time, using the identity $\sqrt{a} \leq \sqrt{b} + \frac{|a-b|}{\sqrt{b}}$ for all $a, b \geq 0$, we obtain

$$\frac{\sqrt{\sum_{j=1}^n Y_j}}{n - 2\tau} \leq \frac{\sqrt{\sum_{j=1}^n U_j}}{n - 2\tau} + \frac{1}{n - 2\tau} \frac{\left| \sum_{j=1}^n Y_j - \sum_{j=1}^n U_j \right|}{\sqrt{\sum_{j=1}^n U_j}}. \tag{28}$$

Substituting Eq. (28) into Eq. (27), we obtain

$$\left| \frac{1}{n} \sum_{j=1}^n Y_j - \frac{1}{n} \sum_{j=1}^n U_j \right| \leq \frac{5\sqrt{\tau B \log(2/\delta) \sum_{j=1}^n U_j}}{n - 2\tau} + \frac{7\tau B \log(2/\delta)}{n - 2\tau}$$

$$+ \frac{5\sqrt{\tau B \log(2/\delta)}}{n - 2\tau} \frac{\left| \sum_{j=1}^n Y_j - \sum_{j=1}^n U_j \right|}{\sqrt{\sum_{j=1}^n U_j}},$$

with probability at least $1 - 6\delta - n_0 \mu \rho^\tau$.

For $\tau = t_{\text{mix}}(\epsilon/n)$, we may now use the conditions $\sum_{j=1}^n U_j \geq 36\tau B \log(2/\delta)$ and $n \geq 24\tau$ to obtain

$$\left| \sum_{j=1}^n Y_j - \sum_{j=1}^n U_j \right| \leq 60 \sqrt{\tau B \log(2/\delta) \sum_{j=1}^n U_j} + 84\tau B \log(2/\delta),$$

with probability at least $1 - 6\delta - n_0 \mu \rho^\tau$. Substituting the above display into Eq. (28), we obtain

$$\sqrt{\sum_{j=1}^n Y_j} \leq 11 \sqrt{\sum_{j=1}^n U_j} + 14\sqrt{\tau B \log(2/\delta)},$$

with probability at least $1 - 6\delta - n_0 \mu \rho^\tau$. This completes the proof of the claim and therefore of the lemma. $\qquad \square$

6.4 Proof of Proposition 1

Recall that Proposition 1 bounds the ℓ_1 risk between the plug-in estimator $\widehat{M}_{\text{PI},\zeta}$ and the true count probability mass M_ζ^π for sufficiently large values of count ζ. Also recall the shorthand notation $\tau_0 = t_{\text{mix}}(\epsilon/n^2)$. We denote by $\mathcal{X}_\zeta$ the set of all symbols appearing exactly ζ times in X^n and by $\mathcal{X}_{\geq \zeta}$ the set of all symbols appearing at least ζ times in X^n. Clearly, we have $\mathcal{X}_\zeta \subseteq \mathcal{X}_{\geq \zeta}$. Using

this fact and the triangle inequality, we have

$$\left| M_\zeta^\pi - \widehat{M}_{\mathrm{PI},\zeta} \right| = \left| M_\zeta^\pi - \frac{\zeta \varphi_\zeta}{n} \right|$$

$$= \left| \sum_{x \in \mathcal{X}} \mathbb{I}\left\{ N_x(X^n) = \zeta \right\} \left(\pi_x - \frac{\zeta}{n} \right) \right|$$

$$\leq \sum_{x \in \mathcal{X}} \mathbb{I}\left\{ N_x(X^n) = \zeta \right\} \left| \pi_x - \frac{N_x(X^n)}{n} \right|$$

$$\leq \sum_{x \in \mathcal{X}_{\geq \zeta}} \mathbb{I}\left\{ N_x(X^n) = \zeta \right\} \left| \pi_x - \frac{N_x(X^n)}{n} \right|. \tag{29}$$

At this juncture, one possible approach is to prove a high-probability bound on the deviation term $\left| \pi_x - \frac{N_x(X^n)}{n} \right|$ for each $x \in \mathcal{X}$, and then take a union bound over the state space. However, this would induce a dependence on $|\mathcal{X}|$ and the resulting bound would no longer be universal in n. Consequently, we reduce the sum over $\mathcal{X}_{\geq \zeta}$ to the sum over a deterministic set of size at most $\mathcal{O}(n)$ using the following lemma.

Lemma 3. *Recall that $\tau_0 = \mathsf{t}_{\mathsf{mix}}(\epsilon/n^2)$. Then*

$$\mathbb{P}\left(\exists x \in \mathcal{X} : N_x(X^n) \geq 1 + \sqrt{(4 + 8\tau_0)\log(n/\delta)} + \frac{4\tau_0 \log(n/\delta)}{3} \text{ and } \pi_x \leq 1/n \right) \leq \delta + \epsilon.$$

Lemma 3 is proved in Section 6.4.1. Let us also define the deterministic set $\mathcal{Y} := \left\{ x \in \mathcal{X} : \pi_x \geq \frac{1}{n} \right\}$.

By Lemma 3, we have that for any $\zeta \geq 1 + \sqrt{(4 + 8\tau_0)\log(n/\delta)} + \frac{4\tau_0 \log(n/\delta)}{3}$, the inclusion

$$\mathcal{X}_{\geq \zeta} \subseteq \mathcal{Y} \text{ holds with probability at least } 1 - \delta - \epsilon. \tag{30}$$

Next, we observe that $N_x(X^n)$ can be written as follows:

$$N_x(X^n) = \sum_{j=1}^{n} \mathbb{I}\left\{ X_j = x \right\},$$

which takes the form of a sum of indicator random variables over the sequence X^n. Since $\mathcal{Y}$ contains at most n (nonrandom) elements, we may apply the union bound over these elements in conjunction with Lemma 2. This implies that the following event occurs *uniformly* over all $x \in \mathcal{Y}$ with probability at least $1 - 10\delta - 2\epsilon$ (note that we defined $\tau_0 = \mathsf{t}_{\mathsf{mix}}(\epsilon/n^2)$):

$$\left| \pi_x - \frac{N_x(X^n)}{n} \right| \lesssim \begin{cases} \frac{\sqrt{\tau_0 N_x(X^n) \log(2n/\delta)}}{n} & \text{if } N_x(X^n) \geq 36\tau_0 \log(2n/\delta) \\ \sqrt{\frac{4\pi_x(1 + 2\tau_0)\log(n/\delta)}{n}} + \frac{4\tau_0 \log(n/\delta)}{3n} & \text{otherwise.} \end{cases} \tag{31}$$

We make particular use of the first case in the proof — the second case is stated for completeness. Working now on the intersection of the high-probability events in (30) and (31), which in turn occurs with probability at least $1 - 11\delta - 3\epsilon$, and considering

$$\zeta \geq \max \left\{ 36\tau_0 \log(2n/\delta), 1 + \sqrt{(4 + 8\tau_0)\log(n/\delta)} + \frac{4\tau_0 \log(n/\delta)}{3} \right\},$$

we have

$$\sum_{x\in\mathcal{X}_{\geq\zeta}}\mathbb{I}\{N_x(X^n)=\zeta\}\left|\pi_x-\frac{N_x(X^n)}{n}\right|$$

$$=\sum_{x\in\mathcal{X}_{\geq\zeta}\cap\mathcal{Y}}\mathbb{I}\{N_x(X^n)=\zeta\}\left|\pi_x-\frac{N_x(X^n)}{n}\right|+\sum_{x\in\mathcal{X}_{\geq\zeta}\cap\mathcal{Y}^c}\mathbb{I}\{N_x(X^n)=\zeta\}\left|\pi_x-\frac{N_x(X^n)}{n}\right|$$

$$\stackrel{(i)}{=}\sum_{x\in\mathcal{X}_{\geq\zeta}\cap\mathcal{Y}}\mathbb{I}\{N_x(X^n)=\zeta\}\left|\pi_x-\frac{N_x(X^n)}{n}\right|$$

$$\stackrel{(ii)}{\lesssim}\sum_{x\in\mathcal{X}_{\geq\zeta}}\mathbb{I}\{N_x(X^n)=\zeta\}\frac{\sqrt{\tau_0 N_x(X^n)\log(2n/\delta)}}{n}$$

$$\leq\sum_{x\in\mathcal{X}}\mathbb{I}\{N_x(X^n)=\zeta\}\frac{\sqrt{\tau_0 N_x(X^n)\log(2n/\delta)}}{n},$$

where step (i) follows by Eq. (30) and step (ii) by Eq. (31).

Now, using the fact that $\sum_{x\in\mathcal{X}}\mathbb{I}\{N_x(X^n)=\zeta\}\sqrt{N_x(X^n)}=\varphi_\zeta(X^n)\sqrt{\zeta}$, we obtain

$$\left|M_\zeta^\pi-\widehat{M}_{\mathrm{PI},\zeta}\right|\lesssim\frac{\sqrt{\tau_0}}{n}\sqrt{\log\left(\frac{2n}{\delta}\right)}\varphi_\zeta(X^n)\sqrt{\zeta}\tag{32}$$

on the same event occurring with probability at least $1-11\delta-3\epsilon$. Adjusting constants in the definitions of δ and ϵ and substituting the value of τ_0 yields the statement of the proposition. $\square$

6.4.1 Proof of Lemma 3

In order to prove this lemma, we start by carefully partitioning the state space. We number the elements in $\mathcal{X}$ as $1,\ldots,|\mathcal{X}|$ and assume without loss of generality that they are ordered such that $\pi_1\leq\pi_2\leq\cdots\leq\pi_{|\mathcal{X}|}$. Let us define $j^*+1:=\min\{i:\pi_i>1/n\}$. Also define the cumulative probabilities $c_i=\sum_{m=1}^i\pi_i$, and let $i_k:=\min\{i:c_i\geq k/n\}$ denote a sequence of $n+1$ indices with $i_0:=0$. Now, partition the first j^* indices into disjoint sets $T_1,\ldots,T_L$ given by

$$T_\ell=\{i_{\ell-1}+1,\ldots,i_\ell\},\tag{33}$$

where the set is considered to be empty if $i_\ell=i_{\ell-1}$, and L is defined such that $i_L=j^*$. We overload notation and denote by π_{T_ℓ} the total stationary mass of all symbols appearing in the partition T_ℓ, i.e. $\pi_{T_\ell}:=\sum_{x\in T_\ell}\pi_x$. Note by the definition of $\{i_\ell\}_{\ell=1}^L$ and $\{T_\ell\}_{\ell=1}^L$ that each block T_ℓ contains elements whose stationary probabilities sum up to at most $1/n$. In other words, we have $\pi_{T_\ell}\leq 1/n$ for all $\ell\in[L]$. Thus, we have $\sum_{\ell=1}^L\pi_{T_\ell}\leq\frac{L}{n}\leq 1$, which implies that $L\leq n$.

Next, we define

$$N_{T_\ell}(X^n):=\sum_{x\in T_\ell}N_x(X^n).$$

27

We also set $\zeta_0 := 1 + \sqrt{(4 + 8\tau_0)\log(n/\delta)} + \frac{4\tau_0\log(n/\delta)}{3}$, where we defined $\tau_0 := \mathsf{t}_{\mathsf{mix}}(\epsilon/n^2)$ in the statement of Lemma 3. By construction of the partitions $\{T_\ell\}_{\ell=1}^L$, we obtain

$$\mathbb{P}\left(\exists x : N_x(X^n) \geq \zeta_0 \text{ and } \pi_x \leq 1/n\right) = \mathbb{P}\left(\exists x \in \cup_{\ell=1}^L T_\ell : N_x(X^n) \geq \zeta_0\right)$$
$$\leq \mathbb{P}\left(\exists \ell \in [L] : N_{T_\ell}(X^n) \geq \zeta_0\right)$$
$$\leq \sum_{\ell=1}^L \mathbb{P}\left(N_{T_\ell}(X^n) \geq \zeta_0\right). \tag{34}$$

It remains to characterize each of the terms $\mathbb{P}\left(N_{T_\ell}(X^n) \geq \zeta_0\right)$, for which we will use the statement of Bernstein's inequality from Lemma 2(a). Consider the random variable

$$B_i = \mathbb{I}\{X_i \in T_\ell\} - \mathbb{P}\left(X_i \in T_\ell\right).$$

We note that $|B_i| \leq 1$ and $\mathbb{E}[B_i] = 0$. Applying the Bernstein inequality from Lemma 2(a) — in its tail bound form — to the process $(B_1, \ldots, B_n)$ with $\tau_0 = \mathsf{t}_{\mathsf{mix}}(\epsilon/n^2)$ yields

$$\mathbb{P}\left(\sum_{i=1}^n B_i \geq t\right) \leq 4\exp\left(\frac{-t^2}{4n\tau_0 v^2 + \frac{4\tau_0}{3}t}\right) + \frac{\epsilon}{n} \tag{35}$$

for any $t > 0$. Recall from the statement of the lemma that $v^2 := \frac{1}{\tau_0^2}\mathsf{var}\left(\sum_{j=1}^{\tau_0} B_j\right) = \frac{1}{\tau_0^2}\sum_{j=1}^{\tau_0}\mathsf{var}\left(B_j\right) + \frac{2}{\tau_0^2}\sum_{j>i}^{\tau_0}\mathsf{cov}(B_i, B_j)$. We have

$$\mathsf{var}(B_i) = \mathbb{P}(X_i \in T_\ell) - \mathbb{P}(X_i \in T_\ell)^2 \leq 1/n \text{ and}$$
$$\sum_{j>i}^{\tau_0}\mathsf{cov}(B_i, B_j) \overset{(i)}{\leq} \sum_{j>i}^{\tau_0}\sqrt{\mathsf{var}(B_i)\mathsf{var}(B_j)} \leq \frac{\tau_0^2}{n},$$

where step (i) follows from the Cauchy-Schwarz inequality. This directly yields

$$v^2 \leq \frac{1}{n\tau_0} + \frac{2}{n} \implies 4n\tau_0 v^2 \leq 4(1 + 2\tau_0).$$

Continuing from Eq. (35) and substituting the definition of the random variables $\{B_i\}_{i=1}^n$, we have

$$\mathbb{P}\left(\sum_{i=1}^n \mathbb{I}\{X_i \in T_\ell\} \geq t + \sum_{i=1}^n \mathbb{P}\left(X_i \in T_\ell\right)\right) \leq \exp\left(\frac{-t^2}{4 + 8\tau_0 + \frac{4\tau_0}{3}t}\right) + \frac{\epsilon}{n}.$$

Next we note that, because the sets $\{T_\ell\}_{\ell=1}^L$ are disjoint, we have $\sum_{i=1}^n \mathbb{P}\{X_i \in T_\ell\} \leq 1$. From this, we obtain

$$\mathbb{P}\left(N_{T_\ell}(X^n) \geq t + 1\right) \leq \exp\left(\frac{-t^2}{4 + 8\tau_0 + \frac{4\tau_0}{3}t}\right) + \frac{\epsilon}{n}.$$

Substituting $t = \sqrt{(4 + 8\tau_0)\log(n/\delta)} + \frac{4\tau_0\log(n/\delta)}{3}$ in the above yields

$$\mathbb{P}\left(N_{T_\ell}(X^n) \geq 1 + \sqrt{(4 + 8\tau_0)\log(n/\delta)} + \frac{4\tau_0\log(n/\delta)}{3}\right) = \mathbb{P}\left(N_{T_\ell}(X^n) \geq \zeta_0\right) \leq \frac{\delta + \epsilon}{n}. \tag{36}$$

Combining Ineqs. (36) and (34), we obtain

$$\mathbb{P}\left(\exists x : N_x(X^n) \geq \zeta_0 \text{ and } \pi_x \leq 1/n\right) \leq L\left(\frac{\delta + \epsilon}{n}\right) \overset{(i)}{\leq} \delta + \epsilon,$$

where step (i) follows because we established that $L \leq n$. This completes the proof of the lemma. $\square$

6.5 Proof of Proposition 2

In this section, we bound the error of the WINGIT estimator for small frequencies. We first derive bounds on the mean squared error (MSE) for each ζ, and then use these to obtain bounds on the ℓ_1 error. Throughout this proof, note that τ should be thought of as an arbitrary integer that divides n, with $n_0 = n/(2\tau)$.

As in Pananjady et al. (2024), we can relate the MSE of the WINGIT estimator to a "skipped" version as below:

$$\mathsf{MSE}\left(M_\zeta^\pi(X^n), \widehat{M}_{\text{WINGIT},\zeta}(\tau)\right) \leq \frac{1}{2\tau} \sum_{\ell=0}^{2\tau-1} \mathbb{E}\left(M_\zeta^\pi(X^n) - \widehat{M}_{\text{WINGIT},\zeta}(\tau;\ell)\right)^2, \tag{37}$$

where

$$\widehat{M}_{\text{WINGIT},\zeta}(\tau;\ell) := \frac{1}{n_0} \sum_{j=1}^{n_0} \widehat{M}_{\tau,\zeta}^{(2\tau j-\ell)} \text{ for each } \ell = 0, \ldots, 2\tau - 1. \tag{38}$$

$\widehat{M}_{\text{WINGIT},\zeta}(\tau;\ell)$ represents the skipped estimator used in our analysis.

We now focus on bounding the term $\mathbb{E}\left(M_\zeta^\pi(X^n) - \widehat{M}_{\text{WINGIT},\zeta}(\tau;0)\right)^2$. An MSE bound proven for $\ell = 0$ can be identically transferred to bound the MSE of $\widehat{M}_{\text{WINGIT},\zeta}(\tau;\ell)$ for any $\ell = 0, \ldots, 2\tau-1$ (Pananjady et al., 2024, Proposition 5). Thus, without loss of generality, we address the case $\ell = 0$ and analyze the estimator $\widehat{M}_{\text{WINGIT}}(\tau;0)$. Consequently, we use the shorthand $\mathcal{B}_j := \mathcal{B}_{j,0}$ and $\mathcal{H}_j := \mathcal{H}_{j,0}$. For more details and a pictorial representation of the skipped estimators and the corresponding "dependent" and "independent" sets, refer to Pananjady et al. (2024, Section 7.1).

Adding and subtracting the term $\mathbb{E}_{\substack{Y \sim \pi \\ Y \perp\!\!\!\perp X^n}} \mathbb{I}\left\{N_Y(X_{\mathcal{H}_j}) = \zeta\right\}$ to the quantity of interest given by $\frac{1}{2}|\widehat{M}_{\text{WINGIT},\zeta}(\tau;0) - M_\zeta^\pi(X^n)|^2$, we obtain

$$\frac{1}{2} \cdot \left| \frac{1}{n_0} \sum_{j=1}^{n_0} \mathbb{I}\left\{N_{X_{2\tau j}}(X_{\mathcal{H}_j}) = \zeta\right\} - \mathbb{E}_{\substack{Y \sim \pi \\ Y \perp\!\!\!\perp X^n}} \mathbb{I}\left\{N_Y(X^n) = \zeta\right\} \right|^2$$

$$\leq \underbrace{\left| \frac{1}{n_0} \sum_{j=1}^{n_0} \left(\mathbb{E}_{\substack{Y \sim \pi \\ Y \perp\!\!\!\perp X^n}} \mathbb{I}\left\{N_Y(X_{\mathcal{H}_j}) = \zeta\right\} - \mathbb{E}_{\substack{Y \sim \pi \\ Y \perp\!\!\!\perp X^n}} \mathbb{I}\left\{N_Y(X^n) = \zeta\right\} \right) \right|^2}_{T_1}$$

$$+ \underbrace{\left| \frac{1}{n_0} \sum_{j=1}^{n_0} \left(\mathbb{I}\left\{N_{X_{2\tau j}}(X_{\mathcal{H}_j}) = \zeta\right\} - \mathbb{E}_{\substack{Y \sim \pi \\ Y \perp\!\!\!\perp X^n}} \mathbb{I}\left\{N_Y(X_{\mathcal{H}_j}) = \zeta\right\} \right) \right|^2}_{T_2}.$$

The rest of the proof proceeds by bounding $\mathbb{E}[T_1]$ and $\mathbb{E}[T_2]$.

Bounding $\mathbb{E}[T_1]$: In order to bound T_1 we define the random variable $P_j = \mathbb{I}\{N_Y(X_{\mathcal{H}_j}) = \zeta\} - \mathbb{I}\{N_Y(X^n) = \zeta\}$. We have

$$T_1 = \frac{1}{n_0^2}\left|\sum_{j=1}^{n_0}\mathbb{E}_Y P_j\right|^2.$$

We obtain $P_j \leq \mathbb{I}\{Y \in X_{\mathcal{B}_j}\}\mathbb{I}\{N_Y(X_{\mathcal{H}_j}) = \zeta\} \leq 1$ in a manner similar to the argument in (Pananjady et al., 2024, Lemma 12). We briefly outline the argument here for completeness. Since $P_j \in \{-1, 0, 1\}$ and $\mathbb{I}\{Y \in X_{\mathcal{B}_j}\} \cdot \mathbb{I}\{N_Y(X_{\mathcal{H}_j}) = \zeta\} \in \{0, 1\}$, the inequality is only non-trivial for the case where $P_j = 1$. In this case, we have $N_Y(X_{\mathcal{H}_j}) = \zeta$ and $N_Y(X^n) \neq \zeta$, which implies that $Y \in X_{\mathcal{B}_j}$. Thus, we have $\mathbb{I}\{Y \in X_{\mathcal{B}_j}\}\mathbb{I}\{N_Y(X_{\mathcal{H}_j}) = \zeta\} = 1$ and so the above inequality holds.

Since the blocks $\{\mathcal{B}_j\}_{j=1}^{n_0}$ are non-overlapping, we have

$$\bigsqcup_{j' \in [n_0] \setminus j} \mathcal{B}_{j'} \subset \mathcal{H}_j.$$

Now, suppose that for some j we have $\mathbb{I}\{Y \in X_{\mathcal{B}_j}\} \cdot \mathbb{I}\{N_Y(X_{\mathcal{H}_j}) = \zeta\} = 1$. This means that Y occurs at least once in $\mathcal{B}_j$, but its number of occurrences outside of $\mathcal{B}_j$ is equal to ζ. Then, we have

$$\sum_{j' \in [n_0] \setminus j}\mathbb{I}\{Y \in X_{\mathcal{B}_{j'}}\} \cdot \mathbb{I}\{N_Y(X_{\mathcal{H}_{j'}}) = \zeta\} \leq \sum_{j' \in [n_0] \setminus j}\mathbb{I}\{Y \in X_{\mathcal{B}_{j'}}\} \leq N_Y(X_{\mathcal{H}_j}) = \zeta.$$

Putting these pieces together yields $\sum_{j=1}^{n_0} P_j \leq \zeta + 1$ point-wise. Ultimately, this yields

$$T_1 = \frac{1}{n_0^2}\left(\mathbb{E}_Y \sum_{j=1}^{n_0} P_j\right)^2 \leq \left(\frac{\zeta+1}{n_0}\right)^2, \tag{39}$$

and so $\mathbb{E}[T_1] \leq \left(\frac{\tau(\zeta+1)}{n}\right)^2$. This ultimately turns out to be a lower-order term and so this worst-case analysis suffices.

Bounding $\mathbb{E}[T_2]$: The term T_2, which captures a certain conditional variance, will be the dominant term in the analysis. First, we have

$$T_2 = \frac{1}{n_0^2}\sum_{j,k=1}^{n_0} Z_j Z_k, \tag{40}$$

where we define

$$Z_j := \mathbb{I}\{N_{X_{2j\tau}}(X_{\mathcal{H}_j}) = \zeta\} - \mathbb{E}_{\substack{Y \sim \pi_n \\ Y \perp\!\!\!\perp X^n}}\mathbb{I}\{N_Y(X_{\mathcal{H}_j}) = \zeta\} \text{ for all } j \in [n_0].$$

We thus have

$$T_2 = \frac{1}{n_0^2}\sum_{j=1}^{n_0} Z_j^2 + \frac{1}{n_0^2}\sum_{j=1}^{n_0}\sum_{\substack{k=1 \\ k \neq j}}^{n_0} Z_j Z_k.$$

We first bound the cross terms in T_2. We define the following random variables for each $j, k \in [n_0]$ with $j \neq k$ and conditioned on a random variable $Y \sim \pi$:

$$Q_{j,k} := \mathbb{I}\left\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\right\} - \mathbb{I}\left\{N_Y(X_{\mathcal{H}_k}) = \zeta\right\}. \tag{41}$$

Observe that $Q_{j,k} \in [-1, 1]$ point-wise and so $|Q_{j,k}| \in [0, 1]$ point-wise. The following lemma relates the terms $Z_j Z_k$ and $|Q_{j,k}| + |Q_{k,j}|$ in expectation.

Lemma 4. *For each $j \neq k$, we have*

$$\mathbb{E}[Z_j Z_k] \leq 4\mathbb{E}[|Q_{j,k}|] + 4\mathbb{E}[|Q_{k,j}|] + 12\mu\rho^\tau,$$

where $Q_{j,k}$ is defined in Eq. (41).

We take Lemma 4 as given for the moment and prove it in Section 6.5.1. Further, auxiliary Lemma 10 yields the following point-wise bound on $|Q_{j,k}|$ for any $Y \in \mathcal{X}$:

$$|Q_{j,k}| \leq \mathbb{I}\left\{Y \in X_{\mathcal{B}_j}\right\} \cdot \left(\mathbb{I}\left\{N_Y(X_{\mathcal{H}_k}) = \zeta\right\} + \mathbb{I}\left\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\right\}\right).$$

Define the shorthand

$$R_Y := \sum_{j=1}^{n_0} \sum_{\substack{k=1 \\ k \neq j}}^{n_0} \mathbb{I}\left\{Y \in X_{\mathcal{B}_j}\right\} \mathbb{I}\left\{N_Y(X_{\mathcal{H}_k}) = \zeta\right\} + \sum_{j=1}^{n_0} \sum_{\substack{k=1 \\ k \neq j}}^{n_0} \mathbb{I}\left\{Y \in X_{\mathcal{B}_j}\right\} \cdot \mathbb{I}\left\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\right\}. \tag{42}$$

Combining Lemma 4 and Eq. (42), we obtain the following bound for the cross terms in T_2:

$$\frac{1}{n_0^2} \sum_{j=1}^{n_0} \sum_{\substack{k=1 \\ k \neq j}}^{n_0} \mathbb{E}[Z_j Z_k] \overset{(i)}{\leq} \frac{8}{n_0^2} \mathbb{E}[R_Y] + 12\mu\rho^\tau.$$

It remains to provide a satisfactory bound for $\mathbb{E}[R_Y]$, which is the technical crux of our analysis. The following lemma provides an adaptive, point-wise bound on R_Y.

Lemma 5. *Consider the quantity R_Y defined in Eq. (42). For all $Y \in \mathcal{X}$, we have*

$$R_Y \leq \zeta n_0 \cdot \mathbb{I}\left\{N_Y = \zeta\right\} \log(2\tau) + 12 n_0 \sum_{u=1}^{4\tau - 2} \mathbb{I}\left\{N_Y = \zeta + u\right\} \frac{(\zeta + u)}{u}.$$

We take Lemma 5 as given for the moment and prove it in Section 6.5.2.
For the diagonal terms in T_2, we have

$$\frac{1}{n_0^2} \sum_{j=1}^{n_0} \mathbb{E}[Z_j^2] = \frac{1}{n_0^2} \sum_{j=1}^{n_0} \mathbb{E}\left[\left|\mathbb{I}\left\{N_{X_{2\tau j}}(X_{\mathcal{H}_j}) = \zeta\right\} - \mathbb{E}_{\substack{Y \sim \pi \\ Y \perp\!\!\!\perp X^n}} \mathbb{I}\left\{N_Y(X_{\mathcal{H}_j}) = \zeta\right\}\right|^2\right]. \tag{43}$$

Using the property of indicator random variables taking values in $[0, 1]$, we obtain

$$\frac{1}{n_0^2} \sum_{j=1}^{n_0} \mathbb{E}[Z_j^2] \leq \frac{1}{n_0^2} \sum_{j=1}^{n_0} \mathbb{E}\left[\mathbb{I}\left\{N_{X_{2\tau j}}(X_{\mathcal{H}_j}) = \zeta\right\} + \mathbb{E}_{\substack{Y \sim \pi \\ Y \perp\!\!\!\perp X^n}} \mathbb{I}\left\{N_Y(X_{\mathcal{H}_j}) = \zeta\right\}\right]$$

31

Next, we apply Lemma 11 to obtain

$$\frac{1}{n_0^2} \sum_{j=1}^{n_0} \mathbb{E}[Z_j^2] \leq \frac{2}{n_0^2} \sum_{j=1}^{n_0} \mathbb{E}\left[\mathbb{I}\left\{N_Y(X_{\mathcal{H}_j}) = \zeta\right\}\right] + \frac{1}{n_0^2} \sum_{j=1}^{n_0} 3\mu\rho^\tau$$

$$= \frac{2}{n_0^2} \sum_{j=1}^{n_0} \mathbb{E}\left[\mathbb{I}\left\{N_Y(X_{\mathcal{H}_j}) = \zeta\right\}\right] + \frac{3\mu\rho^\tau}{n_0}.$$

The following lemma provides a point-wise, adaptive bound on the term $\sum_{j=1}^{n_0} \mathbb{I}\left\{N_Y(X_{\mathcal{H}_j}) = \zeta\right\}$.

Lemma 6. *We have*

$$\sum_{j=1}^{n_0} \mathbb{I}\left\{N_Y(X_{\mathcal{H}_j}) = \zeta\right\} \leq \sum_{u=1}^{2\tau-1} \mathbb{I}\left\{N_Y = \zeta + u\right\} \cdot \frac{\zeta + u}{u} + n_0 \cdot \mathbb{I}\left\{N_Y = \zeta\right\}.$$

We take Lemma 6 as given for the moment and prove it in Section 6.5.3. Applying the lemma yields

$$\frac{1}{n_0^2} \sum_{j=1}^{n_0} \mathbb{E}[Z_j^2] \leq \frac{2}{n_0^2} \sum_{u=1}^{2\tau-1} \mathbb{E}[\mathbb{I}\left\{N_Y = \zeta + u\right\}]\frac{\zeta + u}{u} + \frac{2}{n_0}\mathbb{E}[\mathbb{I}\left\{N_Y = \zeta\right\}] + \frac{3\mu\rho^\tau}{n_0}.$$

Combining the cross and diagonal terms for T_2, we have

$$T_2 \leq \frac{2}{n_0^2} \sum_{u=1}^{2\tau-1} \mathbb{E}[\mathbb{I}\left\{N_Y = \zeta + u\right\}]\frac{\zeta + u}{u} + \frac{2}{n_0}\mathbb{E}[\mathbb{I}\left\{N_Y = \zeta\right\}] + \frac{3\mu\rho^\tau}{n_0} + \frac{8}{n_0^2}\mathbb{E}[R_Y] + 12\mu\rho^\tau$$

$$\overset{(i)}{\leq} \frac{2}{n_0^2} \sum_{u=1}^{2\tau-1} \mathbb{E}[M_{\zeta+u}^\pi]\frac{\zeta + u}{u} + \frac{2}{n_0}\mathbb{E}[M_\zeta^\pi] + \frac{8}{n_0}\zeta \cdot \mathbb{E}[M_\zeta^\pi] \cdot \log(2\tau)$$

$$+ \frac{96}{n_0} \sum_{u=1}^{4\tau-2} \mathbb{E}[M_{\zeta+u}^\pi]\frac{(\zeta + u)}{u} + 15\mu\rho^\tau$$

$$\lesssim \frac{1}{n_0}\mathbb{E}[M_\zeta^\pi] + \frac{1}{n_0}\mathbb{E}[M_\zeta^\pi] \cdot (\zeta) \cdot \log(2\tau) + \frac{1}{n_0} \sum_{u=1}^{4\tau-2} \mathbb{E}[M_{\zeta+u}^\pi]\frac{(\zeta + u)}{u} + \mu\rho^\tau, \tag{44}$$

where step (i) uses the relation

$$\mathbb{E}\left[\sum_{u=1}^{4\tau-2} \mathbb{I}\left\{N_Y = \zeta + u\right\}\right] = \mathbb{E}\left[\sum_{x\in\mathcal{X}} \pi_x \sum_{u=1}^{4\tau-2} \mathbb{I}\left\{N_x = \zeta + u\right\}\right]$$

$$= \mathbb{E}\left[\sum_{u=1}^{4\tau-2} \sum_{x\in\mathcal{X}} \pi_x\mathbb{I}\left\{N_x = \zeta + u\right\}\right]$$

$$= \sum_{u=1}^{4\tau-2} \mathbb{E}\left[M_{\zeta+u}^\pi\right].$$

32

Combining the bounds for T_1 and T_2 (Ineqs. (39) and (44) respectively), we have

$$\mathbb{E}\left[|\widehat{M}_{\text{WingIt},\zeta}(\tau;0) - M_\zeta^\pi(X^n)|^2\right] \lesssim \left(\frac{(\zeta+1)\tau}{n}\right)^2 + \frac{1}{n_0}\mathbb{E}[M_\zeta^\pi] + \frac{1}{n_0}\mathbb{E}[M_\zeta^\pi] \cdot \zeta \cdot \log(2\tau)$$

$$+ \frac{1}{n_0}\sum_{u=1}^{4\tau-2}\mathbb{E}[M_{\zeta+u}^\pi]\frac{(\zeta+u)}{u} + \mu\rho^\tau. \tag{45}$$

We now relate the ℓ_1-risk of the WingIt estimator to its MSE for each value of ζ. We have

$$\mathbb{E}|M_\zeta^\pi(X^n) - \widehat{M}_{\text{WingIt},\zeta}(\tau)|$$

$$\overset{(i)}{\leq} \sqrt{\mathbb{E}|M_\zeta^\pi(X^n) - \widehat{M}_{\text{WingIt},\zeta}(\tau)|^2}$$

$$\overset{(ii)}{\lesssim} \sqrt{\left(\frac{(\zeta+1)\tau}{n}\right)^2 + \frac{1}{n_0}\mathbb{E}[M_\zeta^\pi] + \frac{1}{n_0}\mathbb{E}[M_\zeta^\pi] \cdot \zeta \cdot \log(2\tau) + \frac{1}{n_0}\sum_{u=1}^{4\tau-2}\mathbb{E}[M_{\zeta+u}^\pi]\frac{(\zeta+u)}{u} + \mu\rho^\tau}$$

$$\lesssim \frac{(\zeta+1)\tau}{n} + \sqrt{\frac{1}{n_0}\mathbb{E}[M_\zeta^\pi]} + \sqrt{\frac{\mathbb{E}[M_\zeta^\pi] \cdot \zeta \cdot \log(2\tau)}{n_0}} + \sqrt{\frac{1}{n_0}\sum_{u=1}^{4\tau-2}\mathbb{E}[M_{\zeta+u}^\pi]\frac{(\zeta+u)}{u}} + \sqrt{\mu\rho^\tau}. \tag{46}$$

where step (i) follows from Jensen's inequality and step (ii) follows from Eq. (45). Selecting $\tau \geq t_{\text{mix}}(n^{-2})$ yields the statement of Proposition 2. $\qquad\square$

It remains to prove Lemmas 4, 5 and 6, which we do next.

6.5.1 Proof of Lemma 4

The proof of this lemma is analogous to the proof of Lemma 1 in Pananjady et al. (2024). Define, for convenience,

$$\overline{Q}_{j,k} := \mathbb{E}_Y[Q_{j,k}] = \mathbb{E}_Y\left[\mathbb{I}\left\{N_Y(X_{\mathcal{H}_j\cap\mathcal{H}_k}) = \zeta\right\} - \mathbb{I}\left\{N_Y(X_{\mathcal{H}_k}) = \zeta\right\}\right].$$

Note that by auxiliary Lemma 9, we have

$$\overline{Q}_{j,k} \leq \mathbb{E}_Y\left[\mathbb{I}\left\{Y \in X_{\mathcal{H}_k\setminus\mathcal{H}_j}\right\} \cdot \mathbb{I}\left\{N_Y(X_{\mathcal{H}_j\cap\mathcal{H}_k}) = \zeta\right\}\right] \leq 1.$$

We then have the decomposition

$$Z_j Z_k$$
$$= \left(\mathbb{I}\left\{N_{X_{2\tau j}}(X_{\mathcal{H}_j}) = \zeta\right\} - \mathbb{E}_Y[\mathbb{I}\left\{N_Y(X_{\mathcal{H}_j\cap\mathcal{H}_k}) = \zeta\right\}] + \overline{Q}_{k,j}\right) \cdot$$
$$\left(\mathbb{I}\left\{N_{X_{2\tau k}}(X_{\mathcal{H}_k}) = \zeta\right\} - \mathbb{E}_{Y'}[\mathbb{I}\left\{N_{Y'}(X_{\mathcal{H}_j\cap\mathcal{H}_k}) = \zeta\right\}] + \overline{Q}_{j,k}\right)$$
$$\leq \left(\mathbb{I}\left\{N_{X_{2\tau j}}(X_{\mathcal{H}_j}) = \zeta\right\} - \mathbb{E}_Y[\mathbb{I}\left\{N_Y(X_{\mathcal{H}_j\cap\mathcal{H}_k}) = \zeta\right\}]\right) \cdot \left(\mathbb{I}\left\{N_{X_{2\tau k}}(X_{\mathcal{H}_k}) = \zeta\right\} - \mathbb{E}_{Y'}[\mathbb{I}\left\{N_{Y'}(X_{\mathcal{H}_j\cap\mathcal{H}_k}) = \zeta\right\}]\right)$$
$$+ \overline{Q}_{j,k} + \overline{Q}_{k,j} + \overline{Q}_{j,k} \cdot \overline{Q}_{k,j}$$
$$\leq \underbrace{\left(\mathbb{I}\left\{N_{X_{2\tau j}}(X_{\mathcal{H}_j}) = \zeta\right\} - \mathbb{E}_Y[\mathbb{I}\left\{N_Y(X_{\mathcal{H}_j\cap\mathcal{H}_k}) = \zeta\right\}]\right) \cdot \left(\mathbb{I}\left\{N_{X_{2\tau k}}(X_{\mathcal{H}_k}) = \zeta\right\} - \mathbb{E}_{Y'}[\mathbb{I}\left\{N_{Y'}(X_{\mathcal{H}_j\cap\mathcal{H}_k}) = \zeta\right\}]\right)}_{U_{j,k}}$$
$$+ |\overline{Q}_{j,k}| + |\overline{Q}_{k,j}| + |\overline{Q}_{j,k}| \cdot |\overline{Q}_{k,j}|$$
$$\overset{(i)}{\leq} U_{j,k}' + \frac{3}{2}(|\overline{Q}_{j,k}| + |\overline{Q}_{k,j}|),$$

33

where $Y' \sim \pi$ is an independent copy of Y. Above, step (i) follows due to the following algebraic inequalities: Since each $|\overline{Q}_{j,k}| \in [0,1]$, we have $|\overline{Q}_{j,k}| \cdot |\overline{Q}_{k,j}| \leq \sqrt{|\overline{Q}_{j,k}| \cdot |\overline{Q}_{k,j}|} \leq \frac{1}{2}(|\overline{Q}_{j,k}| + |\overline{Q}_{k,j}|)$.

Taking the expectation with respect to the sequence X^n, we have

$$\mathbb{E}_{X^n}[Z_j Z_k] \leq \mathbb{E}_{X^n}[U_{j,k}] + \frac{3}{2}\mathbb{E}_{X^n}[(|\overline{Q}_{j,k}| + |\overline{Q}_{k,j}|)].$$

Now we establish that $\mathbb{E}_{X^n}[U_{j,k}] \leq \frac{5}{2}\mathbb{E}_{X^n}[|\overline{Q}_{j,k}|] + \frac{5}{2}\mathbb{E}_{X^n}[|\overline{Q}_{k,j}|] + 12\mu\rho^\tau$. We have the further decomposition

$$\mathbb{E}_{X^n}[U_{j,k}]$$
$$= \underbrace{\mathbb{E}[\mathbb{I}\{N_{X_{2\tau j}}(X_{\mathcal{H}_j}) = \zeta\} \cdot \mathbb{I}\{N_{X_{2\tau k}}(X_{\mathcal{H}_k}) = \zeta\}]}_{U_1} - \underbrace{\mathbb{E}_{X^n}\left[\mathbb{I}\{N_{X_{2\tau j}}(X_{\mathcal{H}_j}) = \zeta\} \cdot \mathbb{E}_{Y'}[\mathbb{I}\{N_{Y'}(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\}]\right]}_{U_2}$$
$$- \underbrace{\mathbb{E}_{X^n}\left[\mathbb{I}\{N_{X_{2\tau k}}(X_{\mathcal{H}_k}) = \zeta\} \cdot \mathbb{E}_{Y}[\mathbb{I}\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\}]\right]}_{U_3}$$
$$+ \underbrace{\mathbb{E}_{X^n}\left[\mathbb{E}_{Y}[\mathbb{I}\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\}] \cdot \mathbb{E}_{Y'}[\mathbb{I}\{N_{Y'}(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\}]\right]}_{U_4}. \tag{47}$$

We now bound each of the above terms in turn. First, we bound U_1 as

$$U_1 = \mathbb{E}[\mathbb{I}\{N_{X_{2\tau j}}(X_{\mathcal{H}_j}) = \zeta\} \cdot \mathbb{I}\{N_{X_{2\tau k}}(X_{\mathcal{H}_k}) = \zeta\}]$$
$$\overset{(i)}{\leq} \mathbb{E}[\mathbb{I}\{N_Y(X_{\mathcal{H}_j}) = \zeta\} \cdot \mathbb{I}\{N_{Y'}(X_{\mathcal{H}_k}) = \zeta\}] + 6\mu\rho^\tau$$
$$= \mathbb{E}\left[\mathbb{E}_Y\left[\mathbb{I}\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\} - Q_{k,j}\right] \cdot \mathbb{E}_{Y'}\left[\mathbb{I}\{N_{Y'}(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\} - Q_{j,k}\right]\right] + 6\mu\rho^\tau$$
$$\overset{(ii)}{\leq} \mathbb{E}\left[\left(\mathbb{E}_Y\left[\mathbb{I}\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\}\right] - \overline{Q}_{k,j}\right) \cdot \left(\mathbb{E}_{Y'}\left[\mathbb{I}\{N_{Y'}(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\}\right] - \overline{Q}_{j,k}\right)\right] + 6\mu\rho^\tau$$
$$= \mathbb{E}\left[\mathbb{E}_{Y'}\left[\mathbb{I}\{N_{Y'}(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\}\right] \cdot \mathbb{E}_Y\left[\mathbb{I}\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\}\right]\right] + \mathbb{E}\left[\overline{Q}_{j,k} \cdot \overline{Q}_{j,k}\right]$$
$$- \mathbb{E}\left[\overline{Q}_{j,k}\mathbb{E}_Y[\mathbb{I}\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\}]\right] - \mathbb{E}\left[\overline{Q}_{k,j}\mathbb{E}_{Y'}[\mathbb{I}\{N_{Y'}(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\}]\right] + 6\mu\rho^\tau, \tag{48}$$

where step (i) uses Lemma 12 from the appendix (applied with $i_1 = 2\tau\min\{j,k\}, i_2 = 2\tau\max\{j,k\}$, and noting that $i_2 - i_1 \geq 2\tau$ as $j \neq k$), and step (ii) follows because Y and Y' are independent of everything else.

Proceeding to the next term, note that U_2 may be viewed as the expectation over X^n of

$$f(X_{2j\tau}; X_{\mathcal{H}_j}) := \mathbb{I}\{N_{X_{2\tau j}}(X_{\mathcal{H}_j}) = \zeta\} \cdot \mathbb{E}_Y[\mathbb{I}\{N_Y(X_{\mathcal{H}_k \cap \mathcal{H}_j}) = \zeta\}],$$

which is bounded in the range $[0,1]$. We may now apply Lemma 11 from the appendix (for the choice $i = 2\tau j$) to obtain $|\mathbb{E}[f(X_{2\tau j}; X_{\mathcal{H}_j})] - \mathbb{E}[f(Y'; X_{\mathcal{H}_j})]| \leq 3\mu\rho^\tau$. Thus, we have

$$U_2 \geq \mathbb{E}_{X^n}\left[\mathbb{E}_{Y'}[\mathbb{I}\{N_{Y'}(X_{\mathcal{H}_j}) = \zeta\}] \cdot \mathbb{E}_Y[\mathbb{I}\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\}]\right] - 3\mu\rho^\tau$$
$$\geq \mathbb{E}_{X^n}\left[\mathbb{E}_{Y'}[\mathbb{I}\{N_{Y'}(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\}] \cdot \mathbb{E}_Y[\mathbb{I}\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\}]\right] - \mathbb{E}_{X^n}[\overline{Q}_{k,j}] - 3\mu\rho^\tau. \tag{49}$$

By an identical argument to the above, we have

$$U_3 \geq \mathbb{E}_{X^n}\left[\mathbb{E}_{Y'}[\mathbb{I}\{N_{Y'}(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\}] \cdot \mathbb{E}_Y[\mathbb{I}\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\}]\right] - \mathbb{E}_{X^n}[\overline{Q}_{j,k}] - 3\mu\rho^\tau. \tag{50}$$

Putting Ineqs. (48), (49) and (50) together with the definition of U'_4 and performing the requisite cancellations, we have

$$
\begin{aligned}
\mathbb{E}[U'_{j,k}] &\leq (U'_1 - U'_2) + (-U'_3 + U'_4) \\
&\leq \mathbb{E}_{X^n}\left[\overline{Q}_{j,k} \cdot \overline{Q}_{j,k}\right] - \mathbb{E}_{X^n}\left[\overline{Q}_{j,k}\mathbb{E}_Y[\mathbb{I}\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\}]\right] - \mathbb{E}_{X^n}\left[\overline{Q}_{k,j}\mathbb{E}_{Y'}[\mathbb{I}\{N_{Y'}(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\}]\right] \\
&\quad + \mathbb{E}_{X^n}[\overline{Q}_{j,k}] + \mathbb{E}_{X^n}[\overline{Q}_{k,j}] + 12\mu\rho^\tau \\
&\overset{(i)}{\leq} \mathbb{E}_{X^n}\left[|\overline{Q}_{j,k}| \cdot |\overline{Q}_{j,k}|\right] + \mathbb{E}_{X^n}\left[|\overline{Q}_{j,k}|\right] + \mathbb{E}_{X^n}\left[|\overline{Q}_{k,j}|\right] + \mathbb{E}_{X^n}[|\overline{Q}_{j,k}|] + \mathbb{E}_{X^n}[|\overline{Q}_{k,j}|] + 12\mu\rho^\tau \\
&\overset{(ii)}{\leq} \frac{5}{2}\mathbb{E}_{X^n}[|\overline{Q}_{j,k}|] + \frac{5}{2}\mathbb{E}_{X^n}[|\overline{Q}_{j,k}|] + 12\mu\rho^\tau,
\end{aligned}
$$

where step (i) follows by taking the absolute value and using the triangle inequality and step (ii) follows due to the following algebraic inequalities: Since each $|\overline{Q}_{j,k}| \in [0,1]$, we have $|\overline{Q}_{j,k}| \cdot |\overline{Q}_{k,j}| \leq \sqrt{|\overline{Q}_{j,k}| \cdot |\overline{Q}_{k,j}|} \leq \frac{1}{2}(|\overline{Q}_{j,k}| + |\overline{Q}_{k,j}|)$. Substituting the bound for $\mathbb{E}[U_{j,k}]$ completes the proof of Lemma 4. $\qquad\square$

6.5.2 Proof of Lemma 5

From Eq. (42), we have

$$
R_Y = \underbrace{\sum_{j=1}^{n_0}\sum_{\substack{k=1 \\ k \neq j}}^{n_0} \mathbb{I}\{Y \in X_{\mathcal{B}_j}\} \cdot \mathbb{I}\{N_Y(X_{\mathcal{H}_k}) = \zeta\}}_{A_1} + \underbrace{\sum_{j=1}^{n_0}\sum_{\substack{k=1 \\ k \neq j}}^{n_0} \mathbb{I}\{Y \in X_{\mathcal{B}_j}\} \cdot \mathbb{I}\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\}}_{A_2}.
$$

First, we show an upper bound on A_1. It is convenient to introduce variables u and v that count the number of occurrences of Y in $\mathcal{B}_j$ and, $\mathcal{B}_k$ respectively. Then, we have

$$
\begin{aligned}
A_1 &= \sum_{j=1}^{n_0}\sum_{\substack{k=1 \\ k \neq j}}^{n_0} \mathbb{I}\{Y \in X_{\mathcal{B}_j}\} \cdot \mathbb{I}\{N_Y(X_{\mathcal{H}_k}) = \zeta\} \\
&= \sum_{j=1}^{n_0}\sum_{\substack{k=1 \\ k \neq j}}^{n_0}\sum_{u=1}^{2\tau-1} \mathbb{I}\{N_Y(X_{\mathcal{B}_j}) = u\} \cdot \mathbb{I}\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta - u\} \\
&= \sum_{j=1}^{n_0}\sum_{\substack{k=1 \\ k \neq j}}^{n_0}\sum_{u=1}^{2\tau-1}\sum_{v=0}^{2\tau-1} \mathbb{I}\{N_Y(X_{\mathcal{B}_j}) = u\} \cdot \mathbb{I}\{N_Y(X_{\mathcal{B}_k}) = v\} \cdot \mathbb{I}\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta - u\}.
\end{aligned}
$$

Next, we note that $N_Y = N_Y(X_{\mathcal{B}_j}) + N_Y(X_{\mathcal{B}_k}) + N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k})$. Using this along with a union bound over indicator random variables, we obtain

$$
A_1 \leq \sum_{j=1}^{n_0}\sum_{\substack{k=1 \\ k \neq j}}^{n_0}\sum_{u=1}^{2\tau-1}\sum_{v=0}^{2\tau-1} \mathbb{I}\{N_Y = \zeta + v\} \cdot \mathbb{I}\{N_Y(X_{\mathcal{B}_j} \cup X_{\mathcal{B}_k}) = u + v\} \cdot \mathbb{I}\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta - u\}.
$$

35

In order to obtain the final bound, we need to separate the indicator random variable over $X_{\mathcal{B}_j} \cup X_{\mathcal{B}_k}$ into indicators over $X_{\mathcal{B}_j}$ and $X_{\mathcal{B}_k}$ for each value of $u, v \in [2\tau - 1]$. Again applying a union bound over indicator random variables, we have

$$
\sum_{\substack{j=1 \\ }}^{n_0} \sum_{\substack{k=1 \\ k \neq j}}^{n_0} \mathbb{I}\left\{N_Y(X_{\mathcal{B}_j} \cup X_{\mathcal{B}_k}) = u + v\right\} \cdot \mathbb{I}\left\{N_Y = \zeta + v\right\}
$$

$$
\leq \sum_{\substack{j=1 \\ }}^{n_0} \sum_{\substack{k=1 \\ k \neq j}}^{n_0} \left(\mathbb{I}\left\{N_Y(X_{\mathcal{B}_j}) \geq (u+v)/2\right\} + \mathbb{I}\left\{N_Y(X_{\mathcal{B}_k}) \geq (u+v)/2\right\}\right) \cdot \mathbb{I}\left\{N_Y = \zeta + v\right\}
$$

$$
= 2 \sum_{\substack{j=1 \\ }}^{n_0} \sum_{\substack{k=1 \\ k \neq j}}^{n_0} \mathbb{I}\left\{N_Y(X_{\mathcal{B}_j}) \geq (u+v)/2\right\} \cdot \mathbb{I}\left\{N_Y = \zeta + v\right\}. \tag{51}
$$

Substituting Eq. (51) in the bound for A_1, we obtain

$$
A_1 \leq 2 \sum_{\substack{j=1 \\ }}^{n_0} \sum_{\substack{k=1 \\ k \neq j}}^{n_0} \sum_{u=1}^{2\tau-1} \sum_{v=0}^{2\tau-1} \mathbb{I}\left\{N_Y = \zeta + v\right\} \cdot \mathbb{I}\left\{N_Y(X_{\mathcal{B}_j}) \geq (u+v)/2\right\} \cdot \mathbb{I}\left\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta - u\right\}.
$$

In order to proceed further, we need to handle the cases for $v = 0$ and $v > 0$ slightly differently. In particular, we have

$$
A_1 \leq 2 \underbrace{\sum_{\substack{j=1 \\ }}^{n_0} \sum_{\substack{k=1 \\ k \neq j}}^{n_0} \sum_{u=1}^{2\tau-1} \sum_{v=1}^{2\tau-1} \mathbb{I}\left\{N_Y = \zeta + v\right\} \cdot \mathbb{I}\left\{N_Y(X_{\mathcal{B}_j}) \geq (u+v)/2\right\} \cdot \mathbb{I}\left\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta - u\right\}}_{A_1^{(1)}}
$$

$$
+ 2 \underbrace{\sum_{\substack{j=1 \\ }}^{n_0} \sum_{\substack{k=1 \\ k \neq j}}^{n_0} \sum_{u=1}^{2\tau-1} \mathbb{I}\left\{N_Y = \zeta\right\} \cdot \mathbb{I}\left\{N_Y(X_{\mathcal{B}_j}) \geq u/2\right\} \cdot \mathbb{I}\left\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta - u\right\}}_{A_1^{(2)}}.
$$

We bound the terms $A_1^{(1)}$ and $A_1^{(2)}$ separately. Beginning with the term $A_1^{(1)}$, we have

$$
A_1^{(1)} = 2 \sum_{\substack{j=1 \\ }}^{n_0} \sum_{\substack{k=1 \\ k \neq j}}^{n_0} \sum_{u=1}^{2\tau-1} \sum_{v=1}^{2\tau-1} \mathbb{I}\left\{N_Y = \zeta + v\right\} \cdot \mathbb{I}\left\{N_Y(X_{\mathcal{B}_j}) \geq (u+v)/2\right\} \cdot \mathbb{I}\left\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta - u\right\}
$$

$$
\overset{(i)}{\leq} 2 \sum_{\substack{j=1 \\ }}^{n_0} \sum_{\substack{k=1 \\ k \neq j}}^{n_0} \sum_{u=1}^{2\tau-1} \sum_{v=1}^{2\tau-1} \mathbb{I}\left\{N_Y = \zeta + v\right\} \cdot \mathbb{I}\left\{N_Y(X_{\mathcal{B}_j}) \geq v/2\right\} \cdot \mathbb{I}\left\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta - u\right\}
$$

$$
\overset{(ii)}{=} 2 \sum_{\substack{j=1 \\ }}^{n_0} \sum_{\substack{k=1 \\ k \neq j}}^{n_0} \sum_{v=1}^{2\tau-1} \mathbb{I}\left\{N_Y = \zeta + v\right\} \cdot \mathbb{I}\left\{N_Y(X_{\mathcal{B}_j}) \geq v/2\right\} \cdot \sum_{u=1}^{2\tau-1} \mathbb{I}\left\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta - u\right\}
$$

36

Above, step (i) follows because we have $\mathbb{I}\left\{N_Y(X_{\mathcal{B}_j}) \geq (u+v)/2\right\} \leq \mathbb{I}\left\{N_Y(X_{\mathcal{B}_j}) \geq v/2\right\}$, and step (ii) follows by again separating the summation over the cases $u = 0$ and $u > 0$. Using the fact that $\sum_{u=1}^{2\tau-1} \mathbb{I}\left\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta - u\right\} \leq 1$, we then obtain

$$A_1^{(1)} \leq 2 \sum_{j=1}^{n_0} \sum_{\substack{k=1 \\ k \neq j}}^{n_0} \sum_{v=1}^{2\tau-1} \mathbb{I}\left\{N_Y = \zeta + v\right\} \cdot \mathbb{I}\left\{N_Y(X_{\mathcal{B}_j}) \geq v/2\right\}$$

$$\overset{(i)}{\leq} 4n_0 \sum_{v=1}^{2\tau-1} \mathbb{I}\left\{N_Y = \zeta + v\right\} \cdot \frac{\zeta + v}{v}.$$

Above, step (i) follows because

$$2 \sum_{j=1}^{n_0} \sum_{\substack{k=1 \\ k \neq j}}^{n_0} \mathbb{I}\left\{N_Y(X_{\mathcal{B}_j}) \geq v/2\right\} \cdot \mathbb{I}\left\{N_Y = \zeta + v\right\} \leq 4n_0 \cdot \frac{(\zeta + v) \cdot \mathbb{I}\left\{N_Y = \zeta + v\right\}}{v}. \tag{52}$$

The inequality follows because under the condition $\mathbb{I}\left\{N_Y = \zeta + v\right\}$ and for a fixed $j \in [n_0]$, at most $\frac{N_Y}{v/2} = \frac{\zeta + v}{v/2}$ blocks can have at least $v/2$ occurrences of Y for any $Y \in \mathcal{X}$. Thus, the inner summation over k is bounded by $\frac{(\zeta + v)\mathbb{I}\{N_Y = \zeta + v\}}{v/2}$. Completing the outer summation over j yields the desired statement.

Next, we will bound the term $A_1^{(2)}$ which turns out to follow a similar but slightly simpler argument. Using the fact that $\mathbb{I}\left\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta - u\right\} \leq 1$, we obtain

$$A_1^{(2)} \leq 2 \sum_{u=1}^{2\tau-1} \sum_{j=1}^{n_0} \sum_{\substack{k=1 \\ k \neq j}}^{n_0} \mathbb{I}\left\{N_Y = \zeta\right\} \cdot \mathbb{I}\left\{N_Y(X_{\mathcal{B}_j}) \geq u/2\right\}$$

$$\overset{(i)}{\leq} 4n_0 \cdot \sum_{u=1}^{2\tau-1} \mathbb{I}\left\{N_Y = \zeta\right\} \cdot \frac{\zeta}{u} \leq 4n_0 \cdot \zeta \cdot \mathbb{I}\left\{N_Y = \zeta\right\} \cdot \log(2\tau).$$

Above, step (i) follows from Eq. (52) and then using the fact that $\sum_{u=1}^{2\tau-1} 1/u \leq \log(2\tau)$. Combining the bounds for $A_1^{(1)}$ and $A_1^{(2)}$ yields

$$A_1 \leq 4n_0 \cdot \zeta \cdot \mathbb{I}\left\{N_Y = \zeta\right\} \cdot \log(2\tau) + 4n_0 \sum_{v=1}^{2\tau-1} \mathbb{I}\left\{N_Y = \zeta + v\right\} \cdot \frac{\zeta + v}{v}.$$

Next, we will show the upper bound on A_2 through a similar argument. Fixing two indices $j < k \in [n_0]$ and using the union bound over indicators, we have

$$\mathbb{I}\left\{Y \in X_{\mathcal{B}_j}\right\} \cdot \mathbb{I}\left\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\right\} + \mathbb{I}\left\{Y \in X_{\mathcal{B}_k}\right\} \cdot \mathbb{I}\left\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\right\}$$

$$\overset{(i)}{\leq} 2 \cdot \mathbb{I}\left\{Y \in X_{\mathcal{B}_j} \cup X_{\mathcal{B}_k}\right\} \cdot \mathbb{I}\left\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\right\}$$

$$\overset{(ii)}{=} 2 \sum_{u=1}^{2(2\tau-1)} \mathbb{I}\left\{N_Y(X_{\mathcal{B}_j} \cup X_{\mathcal{B}_k}) = u\right\} \cdot \mathbb{I}\left\{N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta\right\}$$

$$= 2 \sum_{u=1}^{4\tau-2} \mathbb{I}\left\{N_Y = \zeta + u\right\} \cdot \mathbb{I}\left\{N_Y(X_{\mathcal{B}_j} \cup X_{\mathcal{B}_k}) = u\right\},$$

where step (i) follows from the union bound and step (ii) follows by introducing the summation over u which counts the number of occurrences of Y in $X_{\mathcal{B}_j} \cup X_{\mathcal{B}_k}$. Consequently, from the above statement, we have

$$A_2 \leq 2 \sum_{j=1}^{n_0} \sum_{\substack{k=1 \\ k \neq j}}^{n_0} \sum_{u=1}^{4\tau-2} \mathbb{I}\left\{N_Y = \zeta + u\right\} \cdot \mathbb{I}\left\{N_Y(X_{\mathcal{B}_j} \cup X_{\mathcal{B}_k}) = u\right\}$$

$$= 2 \sum_{u=1}^{4\tau-2} \mathbb{I}\left\{N_Y = \zeta + u\right\} \cdot \sum_{j=1}^{n_0} \sum_{\substack{k=1 \\ k \neq j}}^{n_0} \mathbb{I}\left\{N_Y(X_{\mathcal{B}_j} \cup X_{\mathcal{B}_k}) = u\right\}$$

$$\overset{(i)}{\leq} 8n_0 \sum_{u=1}^{4\tau-2} \mathbb{I}\left\{N_Y = \zeta + u\right\} \cdot \frac{(\zeta + u)}{u}.$$

Above, step (i) holds using the steps similar to the ones used to obtain Eq. (51) and Ineq (52). This completes the proof of the upper bound on A_2.

Combining the bounds for A_1 and A_2 completes the proof of the lemma. $\qquad\square$

6.5.3   Proof of Lemma 6

This lemma is proved in a manner similar to the proof of Lemma 5, but the argument is much simpler. We introduce the variable u which counts the number of occurrences of Y in $X_{\mathcal{B}_j}$. Then, we have

$$\sum_{j=1}^{n_0} \mathbb{I}\left\{N_Y(X_{\mathcal{H}_j}) = \zeta\right\} = \sum_{j=1}^{n_0} \sum_{u=0}^{2\tau-1} \mathbb{I}\left\{N_Y(X_{\mathcal{B}_j}) = u\right\} \cdot \mathbb{I}\left\{N_Y(X_{\mathcal{H}_j}) = \zeta\right\}$$

$$= \sum_{j=1}^{n_0} \sum_{u=0}^{2\tau-1} \mathbb{I}\left\{N_Y = \zeta + u\right\} \cdot \mathbb{I}\left\{N_Y(X_{\mathcal{B}_j}) = u\right\}$$

As in the proof of Lemma 5, we handle the cases for $u = 0$ and $u > 0$ slightly differently. We have

$$\sum_{j=1}^{n_0} \mathbb{I}\left\{N_Y(X_{\mathcal{H}_j}) = \zeta\right\} = \sum_{u=1}^{2\tau-1} \mathbb{I}\left\{N_Y = \zeta + u\right\} \cdot \sum_{j=1}^{n_0} \mathbb{I}\left\{N_Y(X_{\mathcal{B}_j}) = u\right\} + \mathbb{I}\left\{N_Y = \zeta\right\} \cdot \sum_{j=1}^{n_0} \mathbb{I}\left\{N_Y(X_{\mathcal{B}_j}) = 0\right\}$$

$$\overset{(i)}{\leq} \sum_{u=1}^{2\tau-1} \mathbb{I}\left\{N_Y = \zeta + u\right\} \cdot \frac{\zeta + u}{u} + n_0 \cdot \mathbb{I}\left\{N_Y = \zeta\right\},$$

where step (i) follows because under the condition $N_Y = \zeta + u$, at most $\frac{N_Y}{u} = \frac{\zeta+u}{u}$ blocks can have exactly u occurrences of Y. This completes the proof of the lemma. $\qquad\square$

7   Discussion

We have provided a flexible estimator and analysis of the vector of count probabilities $M^\pi(X^n)$ (and therefore, stationary distribution estimation) of any stationary exponentially α-mixing stochastic process. An explicit construction for the IID case (Acharya et al., 2013) reveals that our estimation

error rate is sharp in its dependence on n; we conjecture that the dependence on $\mathsf{t}_{\mathsf{mix}}$ is tight for our estimator, but showing this remains open. While we have obtained high-probability bounds on the error of the plug-in estimator, our bounds on the WINGIT are in expectation — obtaining corresponding high-probability bounds would be an interesting follow-up problem. Finally, our approach heavily uses the fact that TV decomposes as an ℓ_1 norm of frequency-by-frequency errors; obtaining corresponding analysis for the KL-divergence is an intriguing direction for future work.

More broadly, we believe that the de-correlation devices introduced through our design and analysis of the WINGIT and plug-in estimators for sequences of random variables may find applications in related problems, such as stochastic optimization (Mou et al., 2024; Li et al., 2023), uncertainty quantification with dependent data (Xu and Xie, 2023; Agrawal and Maguluri, 2024), and property testing problems for stochastic processes (Beran, 1992; Kalai and Vempala, 2024).

Acknowledgments

References

J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh. Optimal probability estimation with applications to prediction and classification. In *Conference on Learning Theory*, pages 764–796. PMLR, 2013.

S. Agrawal and S. T. Maguluri. Markov chain variance estimation: A stochastic approximation approach. *arXiv preprint arXiv:2409.05733*, 2024.

J. Beran. Statistical methods for data with long-range dependence. *Statistical science*, pages 404–416, 1992.

D. Berend and A. Kontorovich. The missing mass problem. *Statistics & Probability Letters*, 82(6): 1102–1110, 2012.

D. Berend and A. Kontorovich. On the concentration of the missing mass. *Electronic Communications in Probability*, 18(none):1 – 7, 2013. doi: 10.1214/ECP.v18-2359. URL https://doi.org/10.1214/ECP.v18-2359.

D. Braess and T. Sauer. Bernstein polynomials and learning theory. *Journal of Approximation Theory*, 128(2):187–206, 2004.

P. Chandra and A. Thangaraj. Missing mass under random duplications. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 522–526. IEEE, 2024.

P. Chandra, A. Thangaraj, and N. Rajaraman. Missing mass estimation from sticky channels. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 910–915. IEEE, 2022.

P. Chandra, A. Thangaraj, and N. Rajaraman. How good is Good–Turing for Markov samples? *Trans. Mach. Learn. Res.*, 2024.

S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394, 1999.

K. W. Church and W. A. Gale. Probability scoring for spelling correction. *Statistics and Computing*, 1:93–103, 1991.

V. H. de la Peña, M. J. Klass, and T. Leung Lai. Self-normalized processes: Exponential inequalities, moment bounds and iterated logarithm laws. *The Annals of Probability*, 32(3), July 2004. ISSN 0091-1798.

P. Doukhan. *Mixing: properties and examples*, volume 85. Springer Science & Business Media, 2012.

E. Drukh and Y. Mansour. Concentration bounds for unigram language models. *Journal of Machine Learning Research*, 6(8), 2005.

S. Favaro, A. Lijoi, and I. Prünster. A new estimator of the discovery probability. *Biometrics*, 68 (4):1188–1196, 2012.

R. A. Fisher, A. S. Corbet, and C. B. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, pages 42–58, 1943.

I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3/4):237–264, 1953.

Y. Han, S. Jana, and Y. Wu. Optimal prediction of Markov chains with and without spectral gap. *IEEE Transactions on Information Theory*, 69(6):3920–3959, 2023.

Y. Hao and A. Orlitsky. Doubly-competitive distribution estimation. In *International Conference on Machine Learning*, pages 2614–2623. PMLR, 2019.

Y. Hao, A. Orlitsky, and V. Pichapati. On learning Markov chains. *Advances in Neural Information Processing Systems*, 31, 2018.

F. Jelinek. Probability distribution estimation from sparse data. *IBM technical disclosure bulletin*, 28:2591–2594, 1985.

A. T. Kalai and S. S. Vempala. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 160–171, 2024.

R. Krichevsky and V. Trofimov. The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2):199–207, 1981.

P.-S. Laplace. Essai philosophique sur les probabilités (a philosophical essay on probabilities). *Paris, France: Veuve Courcier*, 1814.

D. A. Levin and Y. Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

T. Li, G. Lan, and A. Pananjady. Accelerated and instance-optimal policy evaluation with linear function approximation. *SIAM Journal on Mathematics of Data Science*, 5(1):174–200, 2023.

A. Lijoi, R. H. Mena, and I. Prünster. A Bayesian nonparametric method for prediction in EST analysis. *BMC bioinformatics*, 8:1–10, 2007.

A. Maurer and M. Pontil. Empirical Bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

D. McAllester and L. Ortiz. Concentration inequalities for the missing mass and for histogram rule error. *Journal of Machine Learning Research*, 4(Oct):895–911, 2003.

D. A. McAllester and R. E. Schapire. On the convergence rate of Good–Turing estimators. In *Conference on Learning Theory*, pages 1–6, 2000.

F. Merlevède, M. Peligrad, and E. Rio. Bernstein inequality and moderate deviations under strong mixing conditions. In *High dimensional probability V: the Luminy volume*, volume 5, pages 273–293. Institute of Mathematical Statistics, 2009.

W. Mou, A. Pananjady, M. J. Wainwright, and P. L. Bartlett. Optimal and instance-dependent guarantees for Markovian linear stochastic approximation. *Mathematical Statistics and Learning*, 7(1):41–153, 2024.

H. Ney, U. Essen, and R. Kneser. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech & Language*, 8(1):1–38, 1994.

M. I. Ohannessian and M. A. Dahleh. Rare probability estimation under regularly varying heavy tails. In *Conference on Learning Theory*, pages 21–1, 2012.

A. Orlitsky and A. T. Suresh. Competitive distribution estimation: Why is Good–Turing good. *Advances in Neural Information Processing Systems*, 28, 2015.

A. Painsky. Generalized Good–Turing improves missing mass estimation. *Journal of the American Statistical Association*, 118(543):1890–1899, 2023.

A. Pananjady, V. Muthukumar, and A. Thangaraj. Just Wing it: Near-optimal estimation of missing mass in a Markovian sequence. *Journal of Machine Learning Research*, 25(312):1–43, 2024.

I. Sason. On reverse Pinsker inequalities. *arXiv preprint arXiv:1503.07118*, 2015.

M. Skorski. Missing mass concentration for Markov chains. *arXiv preprint arXiv:2001.03603*, 2020.

G. Wolfer and A. Kontorovich. Minimax learning of ergodic Markov chains. In *Algorithmic Learning Theory*, pages 904–930. PMLR, 2019.

C. Xu and Y. Xie. Conformal prediction for time series. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11575–11587, 2023.

A   Auxiliary technical lemmas

In this appendix, we collect some auxiliary technical lemmas required in our main proofs.

A.1 Lemmas related to the TV distance

The following lemma relates the TV distance of two natural estimators/distributions to the TV distance between their count probability masses.

Lemma 7. *Let us denote by R_1 and R_2 any two natural distributions over $\mathcal{X}$. For a given sequence X^n, denote their vector of count probability masses as M_1 and M_2 respectively. Then, we have*

$$d_{\mathsf{TV}}(R_1, R_2) = d_{\mathsf{TV}}(M_1, M_2).$$

Proof. We have

$$
\begin{aligned}
d_{\mathsf{TV}}(R_1, R_2) &= \frac{1}{2} \sum_{x \in \mathcal{X}} |R_1(x) - R_2(x)| \\
&\overset{(i)}{=} \frac{1}{2} \sum_{x \in \mathcal{X}} \sum_{u=0}^{n} \mathbb{I}\{N_x = u\} \frac{|M_{u,1} - M_{u,2}|}{\varphi_u} \\
&= \frac{1}{2} \sum_{u=0}^{n} \sum_{x \in \mathcal{X}} \mathbb{I}\{N_x = u\} \frac{|M_{u,1} - M_{u,2}|}{\varphi_u} \\
&= \frac{1}{2} \sum_{u=0}^{n} |M_{u,1} - M_{u,2}| = d_{\mathsf{TV}}(M_1, M_2),
\end{aligned}
$$

where step (i) uses fact that both R_1 and R_2 are natural distributions. This completes the proof of the lemma. $\qquad\square$

The following lemma relates the total variation (TV) distance between the normalized and unnormalized versions of the vector of count probabilities for our estimator from M^π. This result allows us to simplify our analysis by exclusively working with the unnormalized estimator in all our proofs.

Lemma 8. *Let $\widehat{M}^{\mathsf{U}}$ denote a vector of nonnegative entries indexed by $\zeta = \{0, 1, \dots, n\}$, and define $\nu \coloneqq \sum_{\zeta=0}^{n} \widehat{M}_\zeta^{\mathsf{U}} > 0$. Define the normalized estimator $\widehat{M} = \widehat{M}^{\mathsf{U}}/\nu$ as a vector on the simplex $\Delta(\{0, 1, \dots, n\})$. Let M^π be the vector of count probability masses of the stationary distribution π. Then the following holds:*

$$d_{\mathsf{TV}}(\widehat{M}, M^\pi) \leq \|\widehat{M}^{\mathsf{U}} - M^\pi\|_1.$$

Proof. Using the definition $\nu \coloneqq \sum_{\zeta=0}^{n} \widehat{M}_\zeta^{\mathsf{U}}$, we have

$$
\begin{aligned}
d_{\mathsf{TV}}(\widehat{M}, M^\pi) &= \frac{1}{2} \sum_{\zeta=0}^{n} \left|\widehat{M}_\zeta - M_\zeta^\pi\right| \\
&= \frac{1}{2} \sum_{\zeta=0}^{n} \left|\frac{\widehat{M}_\zeta^{\mathsf{U}}}{\nu} - \widehat{M}_\zeta^{\mathsf{U}} + \widehat{M}_\zeta^{\mathsf{U}} - M_\zeta^\pi\right| \\
&\leq \frac{1}{2} \sum_{\zeta=0}^{n} \widehat{M}_\zeta^{\mathsf{U}} \left|\frac{1}{\nu} - 1\right| + \frac{1}{2} \sum_{\zeta=0}^{n} \left|\widehat{M}_\zeta^{\mathsf{U}} - M_\zeta^\pi\right| \\
&= \frac{|\nu - 1|}{2} + \frac{1}{2} \|\widehat{M}^{\mathsf{U}} - M^\pi\|_1.
\end{aligned}
$$

It remains to bound the term $|\nu - 1|$, which we do below:

$$|\nu - 1| = \left| \sum_{\zeta=0}^{n} (\widehat{M}_\zeta^{\mathrm{U}} - M_\zeta^\pi) \right| \le \sum_{\zeta=0}^{n} |\widehat{M}_\zeta^{\mathrm{U}} - M_\zeta^\pi| = \|\widehat{M}^{\mathrm{U}} - M^\pi\|_1.$$

Putting together the pieces completes the proof. $\qquad\square$

A.2 Lemmas concerning indicator random variables

In this section, we state and prove two simple lemmas concerning indicator random variables.

Lemma 9. *For any symbol $x \in \mathcal{X}$, let $Q_{j,k}$ be defined as in Eq. (41). Then, we have*

$$Q_{j,k} \le \mathbb{I}\left\{ x \in X_{\mathcal{H}_k \setminus \mathcal{H}_j} \right\} \cdot \mathbb{I}\left\{ N_x(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta \right\}$$
$$= \mathbb{I}\left\{ x \in X_{\mathcal{B}_j} \right\} \cdot \mathbb{I}\left\{ N_x(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta \right\}.$$

Proof. The only case when this inequality can be violated is when $Q_{j,k} = 1$ and $\mathbb{I}\left\{ x \in X_{\mathcal{B}_j} \right\} \cdot \mathbb{I}\left\{ N_x(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta \right\} = 0$. In every other case the inequality is satisfied since $Q_{j,k} \in \{-1, 0, 1\}$ and $\mathbb{I}\left\{ x \in X_{\mathcal{B}_j} \right\} \cdot \mathbb{I}\left\{ N_x(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta \right\} \in \{0, 1\}$. Let us thus analyze the case when the equality is violated and show that this cannot happen. Note that $Q_{j,k} = 1$ implies

$$N_x(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta \text{ and } N_x(X_{\mathcal{H}_k}) \ne \zeta.$$

We can write the set $X_{\mathcal{H}_k}$ as the union of the disjoint sets $X_{\mathcal{H}_j \cap \mathcal{H}_k}$ and $X_{\mathcal{B}_j}$. Thus, we have

$$N_x(X_{\mathcal{H}_k}) = N_x(X_{\mathcal{H}_j \cap \mathcal{H}_k}) + N_x(X_{\mathcal{B}_j}) = \zeta + N_x(X_{\mathcal{B}_j}).$$

Since $N_x(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta$ and $N_x(X_{\mathcal{H}_k}) \ne \zeta$, we can be sure that the symbol $x \in X_{\mathcal{B}_j}$. This implies that $\mathbb{I}\left\{ x \in X_{\mathcal{B}_j} \right\} \cdot \mathbb{I}\left\{ N_Y(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta \right\} = 1$. Hence, the inequality given in the statement of the lemma is always satisfied. $\qquad\square$

Lemma 10. *For any fixed symbol $x \in \mathcal{X}$, let $Q_{j,k}$ be defined as in Eq. (41). Then, we have*

$$|Q_{j,k}| \le \mathbb{I}\left\{ x \in X_{\mathcal{B}_j} \right\} \cdot \left(\mathbb{I}\left\{ N_x(X_{\mathcal{H}_k}) = \zeta \right\} + \mathbb{I}\left\{ N_x(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta \right\} \right).$$

Proof. From Lemma 9, we have

$$Q_{j,k} \le \mathbb{I}\left\{ x \in X_{\mathcal{B}_j} \right\} \cdot \mathbb{I}\left\{ N_x(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta \right\}.$$

Thus, $Q_{j,k} \ge 0$ implies that $|Q_{j,k}| = Q_{j,k}$. In this case, Lemma 9 yields

$$|Q_{j,k}| = Q_{j,k} \le \mathbb{I}\left\{ x \in X_{\mathcal{B}_j} \right\} \cdot \mathbb{I}\left\{ N_x(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta \right\}$$
$$\le \mathbb{I}\left\{ x \in X_{\mathcal{B}_j} \right\} \cdot \left(\mathbb{I}\left\{ N_x(X_{\mathcal{H}_k}) = \zeta \right\} + \mathbb{I}\left\{ N_x(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta \right\} \right).$$

The only case left to analyze is when $Q_{j,k} = -1$. This implies that

$$N_x(X_{\mathcal{H}_j \cap \mathcal{H}_k}) \ne \zeta \text{ and } N_x(X_{\mathcal{H}_k}) = \zeta.$$

This clearly implies that the symbol $x \in X_{\mathcal{B}_j}$. In this case, we have $|Q_{j,k}| = 1$ and $\mathbb{I}\left\{ x \in X_{\mathcal{B}_j} \right\} \cdot \left(\mathbb{I}\left\{ N_x(X_{\mathcal{H}_k}) = \zeta \right\} + \mathbb{I}\left\{ N_x(X_{\mathcal{H}_j \cap \mathcal{H}_k}) = \zeta \right\} \right) = 1$. This completes the proof of the lemma. $\qquad\square$

A.3   Surrogate Process Lemmas

We now state and prove a series of lemmas that relate the original stochastic process, or sub-process, to surrogate stochastic processes in which certain parts of the process are replaced by independent copies of random variables drawn from the stationary distribution.

Lemma 11. *For each $i \in [n]$, define the stochastic processes*

$$Z_i = (X_1, X_2, \ldots, X_{i-\tau}, X_i, X_{i+\tau}, X_{i+\tau+1}, \ldots, X_n),$$
$$Z_i' = (X_1, X_2, \ldots, X_{i-\tau}, X_i', X_{i+\tau}, X_{i+\tau+1}, \ldots, X_n),$$

where $X_i' \sim \pi$ is drawn independently of everything else. Suppose the stationary stochastic process X_t is exponentially α-mixing with parameters μ and ρ (see Eq. (1b)). Then, we have $d_{\mathsf{TV}}(Z_i, Z_i') \leq 3\mu \cdot \rho^\tau$.

Consequently, for any function f with range $[0, 1]$, we have

$$|\mathbb{E}[f(Z_i) - f(Z_i')]| \leq 3\mu \cdot \rho^\tau.$$

Proof. We prove the bound on total variation, noting that the consequence for bounded functions follows as an immediate corollary. We have

$$
\begin{aligned}
&d_{\mathsf{TV}}(Z_i, Z_i') \\
&= \sup_{S_1,\ldots,S_n \subseteq \mathcal{X}} \Big| \mathbb{P}(X_1 \in S_1, \ldots, X_{i-\tau} \in S_{i-\tau}, X_i \in S_i, X_{i+\tau} \in S_{i+\tau}, \ldots, X_n \in S_n) \\
&\qquad\qquad - \mathbb{P}(X_1 \in S_1, \ldots, X_{i-\tau} \in S_{i-\tau}, X_i' \in S_i, X_{i+\tau} \in S_{i+\tau}, \ldots, X_n \in S_n) \Big| \\
&\overset{(i)}{=} \sup_{S_1,\ldots,S_n \subseteq \mathcal{X}} \Big| \mathbb{P}(X_1 \in S_1, \ldots, X_{i-\tau} \in S_{i-\tau}, X_i \in S_i, X_{i+\tau} \in S_{i+\tau}, \ldots, X_n \in S_n) \\
&\qquad\qquad - \mathbb{P}(X_1 \in S_1, \ldots, X_{i-\tau} \in S_{i-\tau}) \mathbb{P}(X_i \in S_i, X_{i+\tau} \in S_{i+\tau}, \ldots, X_n \in S_n) \\
&\qquad\qquad + \mathbb{P}(X_1 \in S_1, \ldots, X_{i-\tau} \in S_{i-\tau}) \mathbb{P}(X_i \in S_i, X_{i+\tau} \in S_{i+\tau}, \ldots, X_n \in S_n) \\
&\qquad\qquad - \mathbb{P}(X_i \in S_i) \mathbb{P}(X_1 \in S_1, \ldots, X_{i-\tau} \in S_{i-\tau}, X_{i+\tau} \in S_{i+\tau}, \ldots, X_n = S_n) \Big| \\
&\overset{(ii)}{\leq} \alpha(\tau) + \sup_{S_1,\ldots,S_n \subseteq \mathcal{X}} \Big| \mathbb{P}(X_1 \in S_1, \ldots, X_{i-\tau} \in S_{i-\tau}) \mathbb{P}(X_i \in S_i, X_{i+\tau} \in S_{i+\tau}, \ldots, X_n \in S_n) \\
&\qquad\qquad - \mathbb{P}(X_1 \in S_1, \ldots, X_{i-\tau} \in S_{i-\tau}) \mathbb{P}(X_i \in S_i) \mathbb{P}(X_{i+\tau} \in S_{i+\tau}, \ldots, X_n \in S_n) \\
&\qquad\qquad + \mathbb{P}(X_1 \in S_1, \ldots, X_{i-\tau} \in S_{i-\tau}) \mathbb{P}(X_i \in S_i) \mathbb{P}(X_{i+\tau} \in S_{i+\tau}, \ldots, X_n \in S_n) \\
&\qquad\qquad - \mathbb{P}(X_i \in S_i) \mathbb{P}(X_1 \in S_1, \ldots, X_{i-\tau} \in S_{i-\tau}, X_{i+\tau} \in S_{i+\tau}, \ldots, X_n \in S_n) \Big| \\
&\overset{(iii)}{\leq} 2\alpha(\tau) + \sup_{S_1,\ldots,S_n \subseteq \mathcal{X}} \mathbb{P}(X_i \in S_i) \Big| \mathbb{P}(X_1 \in S_1, \ldots, X_{i-\tau} \in S_{i-\tau}) \mathbb{P}(X_{i+\tau} \in S_{i+\tau}, \ldots, X_n \in S_n) \\
&\qquad\qquad - \mathbb{P}(X_1 \in S_1, \ldots, X_{i-\tau} \in S_{i-\tau}, X_{i+\tau} \in S_{i+\tau}, \ldots, X_n \in S_n) \Big| \\
&\leq 2\alpha(\tau) + \alpha(2\tau) \leq 3\mu \cdot \rho^\tau,
\end{aligned}
$$

where (i) follows from independence and adding and subtracting the same term, and (ii) and (iii) follow from triangle inequality and the definition of the α-mixing coefficient. This completes the proof of the lemma. □

Lemma 12. *For each $i_1 < i_2 \in [n]$ with $i_2 - i_1 \geq 2\tau$, define the stochastic sub-processes*

$$Z_{i_1,i_2} = (X_1, X_2, \ldots, X_{i_1-\tau}, X_{i_1}, X_{i_1+\tau}, \ldots, X_{i_2-\tau}, X_{i_2}, X_{i_2+\tau}, \ldots, X_n),$$
$$Z'_{i_1,i_2} = (X_1, X_2, \ldots, X_{i_1-\tau}, X'_{i_1}, X_{i_1+\tau}, \ldots, X_{i_2-\tau}, X'_{i_2}, X_{i_2+\tau}, \ldots, X_n),$$

where $X'_{i_1}, X'_{i_2} \sim \pi$ are drawn independently of each other and of everything else. Suppose the stationary stochastic process X_t is exponentially α-mixing with parameters μ and ρ (see Eq. (1b)). Then, we have $d_{\mathsf{TV}}(Z_{i_1,i_2}, Z'_{i_1,i_2}) \leq 6\mu \cdot \rho^\tau$.

Consequently, for any function f with range $[0,1]$, we have

$$|\mathbb{E}[f(Z_{i_1,i_2}) - f(Z'_{i_1,i_2})]| \leq 6\mu \cdot \rho^\tau.$$

Proof. We prove the bound on total variation, noting that the consequence for bounded functions follows as an immediate corollary. The steps to bound the total variation are similar to the proof of Lemma 11. Define the following intermediate process:

$$\widetilde{Z} = (X_1, X_2, \ldots, X_{i_1-\tau}, X'_{i_1}, X_{i_1+\tau}, \ldots, X_{i_2-\tau}, X_{i_2}, X_{i_2+\tau}, \ldots, X_n).$$

Applying the triangle inequality, we have

$$d_{\mathsf{TV}}(Z_{i_1,i_2}, Z'_{i_1,i_2}) \leq d_{\mathsf{TV}}(Z_{i_1,i_2}, \widetilde{Z}) + d_{\mathsf{TV}}(\widetilde{Z}, Z'_{i_1,i_2}) \overset{(i)}{\leq} 6\mu \cdot \rho^\tau,$$

where step (i) follows from Lemma 11. This completes the proof of the lemma. $\qquad\square$

Lemma 13. *Let $n_0 = n/\tau$. For each $j \in [n_0]$, define the stochastic process $(X'_k)_{k \in \mathcal{D}_{j\tau}}$ as a $|\mathcal{D}_{j\tau}|$-length sequence and initial state sampled from the distribution π and independently of everything else. Thus, the tuple $(X'_1, \ldots, X'_\tau)$ is independent of everything else, and according to the law of $(X_1, \ldots, X_\tau)$. Similarly, choose the blocks $(X'_{\tau+1}, \ldots, X'_{2\tau}), \cdots, (X'_{n-\tau+1}, \ldots, X'_n)$ i.i.d. from the same distribution. We also assume the stationary stochastic process X_t is exponentially α-mixing with parameters μ and ρ (see Eq. (1b)). Now, define the stochastic processes*

$$W_j = \bigoplus_{\ell \in S_j} X_{\mathcal{B}_\ell} \quad and$$
$$W'_j = \bigoplus_{\ell \in S_j} X'_{\mathcal{B}_\ell}$$

where we define $S_j = \{\ell \in [n_0] : |\ell - j| \mod 2 \equiv 0\}$. Then, we have $d_{\mathsf{TV}}(W_j, W'_j) \leq n_0 \cdot \mu \cdot \rho^\tau$.

Consequently, for any function f having range $[0,1]$, we have

$$|\mathbb{E}[f(W_j) - f(W'_j)]| \leq n_0 \cdot \mu \cdot \rho^\tau.$$

Proof. We prove the bound on total variation, noting that the consequence for bounded functions follows as an immediate corollary. Denote $s = |S_j|$ as shorthand and index the elements of S_j as $\ell_1, \ldots, \ell_s$. Fix j and, for each $k = 0, 1, \ldots s$, define the auxiliary stochastic process

$$\widetilde{W}_j^{(k)} = \Big(\bigoplus_{k'=1}^{k} X_{\mathcal{B}_{\ell_{k'}}} \Big) \bigoplus \Big(\bigoplus_{k'=k+1}^{s} X'_{\mathcal{B}_{\ell_{k'}}} \Big).$$

noting that $\widetilde{W}_j^{(s)} = W_j$ and $\widetilde{W}_j^{(0)} = W_j'$. Then by the triangle inequality, we have

$$d_{\mathsf{TV}}(W_j, W_j') \leq \sum_{k=0}^{s-1} d_{\mathsf{TV}}(\widetilde{W}_j^{(k)}, \widetilde{W}_j^{(k+1)}).$$

We now claim that

$$d_{\mathsf{TV}}(\widetilde{W}_j^{(k)}, \widetilde{W}_j^{(k+1)}) \leq \mu \cdot \rho^\tau \text{ for all } k = 0, 1, \ldots, m-1. \tag{53}$$

Since $s = |S_j| \leq n_0$, this claim immediately yields the desired result. To prove claim (53), we note that

$d_{\mathsf{TV}}(\widetilde{W}_j^{(k)}, \widetilde{W}_j^{(k+1)})$

$= \sup_{S_1,\ldots,S_s \subseteq \mathcal{X}} |\mathbb{P}(X_{\mathcal{B}_{\ell_1}} \in S_1, \ldots, X_{\mathcal{B}_{\ell_k}} \in S_k, X'_{\mathcal{B}_{\ell_{k+1}}} \in S_{k+1}, \ldots, X'_{\mathcal{B}_{\ell_s}} \in S_s)$

$\quad\quad - \mathbb{P}(X_{\mathcal{B}_{\ell_1}} \in S_1, \ldots, X_{\mathcal{B}_{\ell_k}} \in S_k, X_{\mathcal{B}_{\ell_{k+1}}} \in S_{k+1}, X'_{\mathcal{B}_{\ell_{k+2}}} \in S_{k+2}, \ldots, X'_{\mathcal{B}_{\ell_s}} \in S_s)|$

$\overset{(i)}{=} \sup_{S_1,\ldots,S_s \subseteq \mathcal{X}} |\mathbb{P}(X_{\mathcal{B}_{\ell_1}} \in S_1, \ldots, X_{\mathcal{B}_{\ell_k}} \in S_k)\mathbb{P}(X'_{\mathcal{B}_{\ell_{k+1}}} \in S_{k+1})\mathbb{P}(X'_{\mathcal{B}_{\ell_{k+2}}} \in S_{k+2}, \ldots, X'_{\mathcal{B}_{\ell_s}} \in S_s)$

$\quad\quad - \mathbb{P}(X_{\mathcal{B}_{\ell_1}} \in S_1, \ldots, X_{\mathcal{B}_{\ell_k}} \in S_k, X_{\mathcal{B}_{\ell_{k+1}}} \in S_{k+1})\mathbb{P}(X'_{\mathcal{B}_{\ell_{k+2}}} \in S_{k+2}, \ldots, X'_{\mathcal{B}_{\ell_s}} \in S_s)|$

$\leq \sup_{S_1,\ldots,S_s \subseteq \mathcal{X}} |\mathbb{P}(X_{\mathcal{B}_{\ell_1}} \in S_1, \ldots, X_{\mathcal{B}_{\ell_k}} \in S_k)\mathbb{P}(X_{\mathcal{B}_{\ell_{k+1}}} \in S_{k+1}) - \mathbb{P}(X_{\mathcal{B}_{\ell_1}} \in S_1, \ldots, X_{\mathcal{B}_{\ell_k}} \in S_k, X_{\mathcal{B}_{\ell_{k+1}}} \in S_{k+1})|$

$\overset{(ii)}{\leq} \mu \cdot \rho^\tau,$

where step (i) follows from independence and step (ii) follows from the α-mixing assumption. This completes the proof of the lemma. $\qquad\square$