Journal of Machine Learning Research 25 (2024) 1-43 Submitted 4/24; Revised 8/24; Published 10/24

Just Wing It: Near-Optimal Estimation of Missing Mass in a Markovian Sequence

Ashwin Pananjady ASHWINPM@GATECH.EDU
Schools of Industrial and Systems Engineering
and Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, USA

Vidya Muthukumar VMUTHUKUMAR8@GATECH.EDU
Schools of Electrical and Computer Engineering
and Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, USA

Andrew Thangaraj ANDREW@EE.IITM.AC.IN
Department of Electrical Engineering
Indian Institute of Technology Madras
Chennai, India

Editor: Christian Shelton

Abstract

We study the problem of estimating the stationary mass—also called the unigram mass—that is missing from a single trajectory of a discrete-time, ergodic Markov chain. This problem has several applications—for example, estimating the stationary missing mass is critical for accurately smoothing probability estimates in sequence models. While the classical Good–Turing estimator from the 1950s has appealing properties for i.i.d. data, it is known to be biased in the Markovian setting, and other heuristic estimators do not come equipped with guarantees. Operating in the general setting in which the size of the state space may be much larger than the length n of the trajectory, we develop a linear-runtime estimator called *Windowed Good–Turing* (WingIt) and show that its risk decays as $\widetilde{\mathcal{O}}(\mathsf{T}_{\mathsf{mix}}/n)$, where $\mathsf{T}_{\mathsf{mix}}$ denotes the mixing time of the chain in total variation distance. Notably, this rate is independent of the size of the state space and minimax-optimal up to a logarithmic factor in $n/\mathsf{T}_{\mathsf{mix}}$. We also present an upper bound on the variance of the missing mass random variable, which may be of independent interest. We extend our estimator to approximate the stationary mass placed on elements occurring with small frequency in the trajectory. Finally, we demonstrate the efficacy of our estimators both in simulations on canonical chains and on sequences constructed from natural language text.

Keywords: missing mass, Good–Turing, Markov chains, minimax optimal

1. Introduction

Two classical problems in statistical analysis—relevant to both design of experiments and inference—are those of assessing *sample coverage* and *discovery probability*. Given a "training" sequence $X^n = (X_1, X_2, \ldots, X_n)$ of random examples in some unknown sample space,

the latter question concerns the probability with which an independent "test" observation Y will be a *discovery*, in that it is an element of the sample space that was unseen at training time. Equivalently, we are interested in estimating the *missing mass* in the training sample X^n, i.e. $\Pr\{Y \notin \{X_1, \ldots, X_n\}\}$.

This problem has roots in statistical analysis for ecology (Fisher et al., 1943), and also has important applications across genomics (Favaro et al., 2012) as well as speech and language modeling (Church and Gale, 1991; Chen and Goodman, 1999). Let us give a few operational examples. For a first example from genomics (Lijoi et al., 2007), suppose we have performed genome sequencing on several genes of an organism as part of training data, and we are now interested in whether there is value in performing additional sequencing. Then the missing mass exactly measures the probability that we discover a new gene with additional sequencing, and an accurate estimate of this quantity can guide decisions about whether or not to sequence further. For a second example, consider the problem of building a probability model for a language corpus (Ney et al., 1994). Many heuristic "smoothing" estimators have been developed for estimating these probability models (e.g., Ney et al., 1994; Jelinek, 1985; Gale and Sampson, 1995). A crucial component of these smoothing techniques is an estimate of the missing mass, since one would like to account for the (non-trivial) possibility that a word exists in the population corpus but has not yet been observed in the training data. Besides these examples, estimates of the missing mass are also used in so-called competitive distribution estimation (Orlitsky and Suresh, 2015) and in estimating other functionals of distributions such as their entropy (Vu et al., 2007). Recent connections have also been made between the missing mass in a fact sequence and the propensity of large language models to "hallucinate" spurious facts (Kalai and Vempala, 2024).

Many estimators with provable—and in fact minimax-optimal—guarantees exist for the case where the training data are exchangeable (Good, 1953; Lijoi et al., 2007; McAllester and Schapire, 2000). While the exchangeability assumption is reasonable in some applications, for example ecology (Shen et al., 2003; Colwell et al., 2012), it is clearly limiting in both genomics and speech or language applications, where temporal dependencies exist between the examples. The simplest form of such temporal dependence is Markovian structure, and, as articulated repeatedly in the literature (Hao et al., 2018; Chandra et al., 2021; Skorski, 2020), handling such structure in a principled fashion is an important first step for estimation of missing mass in temporally dependent training sequences. In spite of a significant body of work motivated by this topic, there still do not exist consistent estimators for missing mass functionals in general classes of Markovian sequences.

In this paper, we propose and theoretically analyze an estimator for the missing mass in a Markovian data sample, and variants for related problems. To make things concrete, suppose our stochastic process $X^n := (X_1, \ldots, X_n)$ is modeled by a stationary Markov chain $(\boldsymbol{P}, \pi)$ on a finite but *unknown* state space $\mathcal{X}$. We will make no assumptions on the alphabet size $|\mathcal{X}|$, and will be interested in also capturing the practically relevant large-alphabet setting, i.e. where $|\mathcal{X}| \gg n$. Here $\pi = (\pi_x)_{x \in \mathcal{X}}$ denotes the unique stationary distribution of the chain, and the matrix $\boldsymbol{P} \in [0,1]^{|\mathcal{X}| \times |\mathcal{X}|}$ denotes the transition probability matrix of the Markov chain. We assume for convenience that $X_1 \sim \pi$, but this assumption can be straightforwardly relaxed[1].

1. As is standard in the literature, one can handle the case of arbitrary X_1 by letting the chain burn in for a certain number of steps until the new "initial" distribution becomes close to the stationary measure.

As previously mentioned, our primary goal is to estimate the mass of the Markov chain that is missing from the random sample X^n. Motivated by the questions above, we focus on the *stationary* missing mass of the chain, given by

$$M_\pi(X^n) := \sum_{x \in \mathcal{X}} \pi_x \cdot \mathbb{I}\{x \notin \{X_1, \ldots, X_n\}\}, \tag{1}$$

where π_x is the probability assigned by the stationary distribution π to element $x \in \mathcal{X}$. Note that $M_\pi(X^n)$ is a *random functional*, as it depends not only on the parameters of the chain but also the random sample X^n. An equivalent definition—which resembles the description above—is given by

$$M_\pi(X^n) = \mathop{\mathbb{E}}_{\substack{Y \sim \pi \\ Y \perp\!\!\!\perp X^n}} \left[\mathbb{I}\{Y \notin \{X_1, \ldots, X_n\}\}\right], \tag{2}$$

where $U \perp\!\!\!\perp V$ denotes that random variables U and V are independent.

The missing mass is not the only functional that is relevant to discovery probabilities. A closely related functional is the *small-count* stationary probability, which measures the probability of seeing an element that had a frequency at most ζ in the training sequence (Lijoi et al., 2007; Favaro et al., 2012). In particular, consider the estimand

$$M_{\pi,\leq\zeta}(X^n) = \mathop{\mathbb{E}}_{\substack{Y \sim \pi \\ Y \perp\!\!\!\perp X^n}} \left[\mathbb{I}\{Y \text{ appears at most } \zeta \text{ times in } \{X_1, \ldots, X_n\}\}\right]. \tag{3}$$

We will present detailed results for estimating the functional $M_{\pi,\leq\zeta}$ in Section 5, focusing up until that point on the missing mass.

Our goal is to produce an estimator $\widehat{M} : \mathcal{X}^n \to [0,1]$ with minimum risk, where risk is measured using the mean squared error. In particular, for an estimand $M : \mathcal{X}^n \to [0,1]$ and estimator $\widehat{M}$, we write

$$\mathsf{MSE}(\widehat{M}, M) = \mathop{\mathbb{E}}_{X^n}\left[|\widehat{M}(X^n) - M(X^n)|^2\right]. \tag{4}$$

Above, the expectation is taken over any other sources of randomness in $\widehat{M}$ in addition to the randomness in the sequence X^n.

To set up some additional notation, we let $\|\mu - \nu\|_{\mathsf{TV}}$ denote the total variation distance between two probability measures μ and ν defined on the same space. Throughout, we assume that the Markov chain is ergodic and mixes in finite time. In particular, let $\mathsf{t}_{\mathsf{mix}}(\epsilon)$ denote the mixing time of the chain to within total variation $\epsilon \in (0, 1/2]$ of the stationary measure, i.e.

$$\mathsf{t}_{\mathsf{mix}}(\epsilon) := \min\left\{t \in \mathbb{N} : \max_{x \in \mathcal{X}} \|e_x^\top \boldsymbol{P}^t - \pi^\top\|_{\mathsf{TV}} \leq \epsilon\right\}, \tag{5}$$

where e_x denotes the indicator vector on element $x \in \mathcal{X}$ and π is viewed as a $|\mathcal{X}|$-dimensional column vector. The quantity $\mathsf{t}_{\mathsf{mix}}(1/4)$ is typically called the mixing time of the chain, and so we will write $\mathsf{T}_{\mathsf{mix}} := \mathsf{t}_{\mathsf{mix}}(1/4)$. It is straightforward to show (see, e.g. Levin and Peres (2017)) that

$$\mathsf{t}_{\mathsf{mix}}(\epsilon) \leq \mathsf{T}_{\mathsf{mix}} \cdot \log(1/\epsilon) \text{ for all } \epsilon < 1/4. \tag{6}$$

1.1 Related work

The problem of estimating missing mass of a random sequence, where each element is drawn from an arbitrarily large sample space, was studied as far back as the 1800s by Laplace (1814), who proposed the first among the class of "add-constant" estimators. These estimators have seen a line of theoretical and empirical follow-up work (Krichevsky and Trofimov, 1981; Gale and Church, 1994), with special attention being paid to the add-1/2-estimator (Krichevsky and Trofimov, 1981). Instead of outputting the normalized empirical frequencies of elements as a maximum-likelihood estimator would, these estimators add a constant to the (un-normalized) empirical frequency prior to normalization. In the process, they output a non-zero missing mass probability.

A notable and groundbreaking result of Good (1953)—attributed also to Turing—moved away from the class of add-constant estimators and proposed to estimate the missing mass via the normalized frequency of elements appearing *once* in the sequence. In particular, letting $\phi_s(X^n)$ denote the number of distinct elements of $\mathcal{X}$ that have appeared s times in the sample X^n, the celebrated Good–Turing estimator for the missing mass is given by

$$\widehat{M}_{\mathsf{GT}} = \frac{\phi_1(X^n)}{n}. \tag{7}$$

The estimator has been applied to diverse areas (see, e.g., Song and Croft, 1999; Gale et al., 1992; Church and Gale, 1991) and has also seen intense theoretical study in the last three decades (see, e.g., McAllester and Schapire, 2000; Drukh and Mansour, 2005; Orlitsky et al., 2003). In particular, several analyses of fine-grained properties of the estimator now exist for the i.i.d. setting (see, e.g., Chandra et al., 2019; Rajaraman et al., 2017; Acharya et al., 2018), and variants of the estimator have also been proposed and studied (Gandolfi and Sastri, 2004; Favaro et al., 2016; Painsky, 2022, 2023). While most analyses focus on additive error—e.g., the mean squared error of estimating the missing mass $M_\pi(X^n)$—the multiplicative error metric has also been studied (Ohannessian and Dahleh, 2012; Mossel and Ohannessian, 2019; Ayed et al., 2021; Ben-Hamou et al., 2017; Grabchak and Zhang, 2017). Besides estimation, the missing mass random variable $M_\pi(X^n)$ has itself generated a lot of interest in the i.i.d. setting—its concentration properties have been thoroughly studied, and several analysis techniques have been developed along the way (McAllester and Ortiz, 2003; Berend and Kontorovich, 2012, 2013).

In contrast to the i.i.d. setting, the Markovian setting has received relatively sparse treatment, in spite of being the main setting—smoothing in language models—that motivated some of the initial papers on the theory of the subject (McAllester and Schapire, 2000, 2001). Some such papers include results for sticky Markov chains (Chandra et al., 2022) and rank-2 Markov chains (Chandra et al., 2021). These papers mainly study the performance of the Good–Turing estimator and/or certain scaled variants of it, and also give sufficient conditions under which Good–Turing can succeed for Markovian data. However, these conditions are restrictive, and it is not yet known if one can perform consistent, let alone minimax-optimal estimation of the missing mass in the general Markovian setting. Concentration of missing mass in the Markovian setting has also received some recent interest (Skorski, 2020)—we will discuss this paper in greater detail in the sequel.

The problem of estimating the small-count probability (3) has also been studied in the literature (Lijoi et al., 2007), along with the related problem of estimating the *exact*

count probability, i.e., the probability of elements occurring *exactly* ζ times (Good, 1953). Estimators of the count probability have been developed for i.i.d. samples, and several theoretical results are also available for this setting (McAllester and Schapire, 2001; Drukh and Mansour, 2005; Acharya et al., 2013). The Markovian case, however, does not seem to have been theoretically studied.

Finally, we mention that besides the missing mass and count probabilities, other estimation and prediction problems (Hao et al., 2018; Han et al., 2023; Wolfer and Kontorovich, 2019) and bounds on "surprise" probabilities (Norris et al., 2017) have been studied for Markov chains. It is worth noting that the size of the state space appears explicitly as a parameter in these results.

1.2 Contributions and organization

Our contributions are summarized below:

- In Section 3, we propose an estimator for stationary missing mass in the Markov setting called the *Windowed Good–Turing*, or WingIt, estimator. Our estimator is based on the viewpoint of the Good–Turing estimator as a leave-one-out estimator, a perspective that we review (for the i.i.d. setting) and develop in Section 2.

- In Theorem 1 of Section 4, we provide a risk bound on the WingIt estimator, showing that it attains mean squared error on the order $\mathsf{T}_{\mathsf{mix}}/n$ up to a logarithmic factor in $n/\mathsf{T}_{\mathsf{mix}}$. This matches, up to this logarithmic factor, the minimax lower bound for missing mass estimation in mixing Markov chains (Chandra et al., 2022).

- Aside from providing an estimator for the missing mass, we also analyze the missing mass functional $M_\pi(X^n)$ as a random variable, and show in Theorem 2 that its variance is bounded on the order ${\mathsf{T}_{\mathsf{mix}}}^2/n$ up to a logarithmic factor. This bound follows from a stability property of the WingIt estimator and constitutes, to our knowledge, the first variance bound on the missing mass in the Markovian setting, complementing a one-sided bound due to Skorski (2020).

- In Section 5, we present an extension of our methodology for estimating the small-count probability (3). Note that this generalizes the problem of estimating the stationary missing mass, which corresponds to the case $\zeta = 0$. This result, stated as Theorem 3, appears to improve analogous guarantees from the literature even for i.i.d. samples.

- In Section 6, we provide simulations on some synthetic Markov chains and on natural language text. These experiments corroborate our theory while showing how the WingIt estimator can significantly outperform the vanilla Good–Turing estimator. We also (empirically) explore an automatic and data-dependent tuning method for the window size hyperparameter in our WingIt estimator.

Notation: For two real numbers a and b, let $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. Let $[n]$ denote the set of natural numbers less than or equal to n. For an index set $P \subseteq [n]$, let $\boldsymbol{X}_P = \{X_i\}_{i \in P}$ denote the set of random variables corresponding to that index set. The set $\boldsymbol{X}_{[n]}$ thus contains all random variables in the sequence X^n. With a slight abuse of

notation, we let $X_P = (X_i)_{i \in P}$ denote the sequence of random variables with indices in P, ordered canonically. For two sequences indexed by u, 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. 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.

2. From i.i.d. to Markov: Revisiting Good–Turing

First, let us revisit the special case where the samples X^n are i.i.d. and the stationary distribution π corresponds to the probability mass function from which each sample is drawn. Here, the celebrated Good–Turing estimator (7) of $M_\pi(X^n)$ is given by the number of symbols that appear *once* in X^n divided by the sample size n. As articulated in the literature (e.g., McAllester and Schapire, 2001), this quantity can be thought of as a *leave-one-out* estimate of the functional that repeatedly simulates a placeholder for Y from the given sample and approximately evaluates the indicator in Eq. (2). In particular, consider the collection of estimators given by the random variables

$$\widehat{M}^{(i)} = \mathbb{I}\left\{X_i \notin \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\}\right\} \quad \text{for} \quad i = 1, \ldots, n. \tag{8}$$

We can then equivalently write the Good–Turing estimator (7) as

$$\widehat{M}_{\mathsf{GT}} = \frac{1}{n} \sum_{i=1}^{n} \widehat{M}^{(i)}. \tag{9}$$

Clearly, the random variable X_i "simulates" drawing a fresh sample Y from π independently of the subsequence $(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$, and inspecting Eq. (2) yields that $\widehat{M}^{(i)}$ is an unbiased estimator of $M_\pi(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. Using the i.i.d. nature of the observations, one can then show that (McAllester and Schapire, 2000)

$$\left|\, \mathbb{E}[M_\pi((X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n))] - M_\pi(X^n)\right| \lesssim n^{-1},$$

so that we have a near-unbiased estimator of the quantity $\mathbb{E}[M_\pi(X^n)]$. Coupling this observation with additional arguments that bound the variance of the estimator $\widehat{M}_{\mathsf{GT}}$ and estimand M_π (McAllester and Schapire, 2000; McAllester and Ortiz, 2003), we obtain a bound on the mean-squared error of the Good–Turing estimator.

It is instructive to re-examine the central pitfall of the Good–Turing estimator for a Markov chain, which is that strong local dependencies between adjacent samples in the Markov chain induce non-vanishing bias. This argument has been sketched before (see, e.g. Chandra et al. (2022)), but we nevertheless give a brief, self-contained illustrative example and a heuristic calculation of the bias below. Let $\mathcal{X} = [k]$ for some $k \gg n$ and consider transition kernels $\boldsymbol{P} \in [0,1]^{k \times k}$ of the form

$$\boldsymbol{P} = (1-p)\boldsymbol{I} + p\mathbf{1}\pi^\top, \tag{10}$$

where $\boldsymbol{I}$ and $\mathbf{1}$ denote the identity matrix and all-1s column vector of suitable dimensions. Such a transition kernel gives rise to a so-called "sticky", or lazy, Markov chain having

stationary distribution π and mixing time $\frac{1}{2p} \leq \mathsf{T}_{\mathsf{mix}} \leq \frac{2}{p}$ for all $k \geq 2$ and $p \in (0, 1/2]$ (see Lemma 10). Thus, as the probability p becomes small, the mixing time becomes proportionally large.

Now suppose that $\pi = \frac{1}{k}\mathbf{1}$, so that the stationary distribution is the uniform distribution on k elements. Due to the stickiness of the chain—i.e., its propensity to remain in its current state—we will see $K \asymp np$ unique elements in a typical sample $(X_1, \ldots, X_n)$ of the chain. The stationary missing mass of the chain will hence be given, in expectation, by

$$\mathbb{E}[M_\pi(X^n)] = \frac{k - \mathbb{E}[K]}{k} \geq \frac{k - Cnp}{k}$$

for some universal constant $C > 0$. In particular, if $k \geq Cn$ and $p \leq 1/4$, we have $\mathbb{E}[M_\pi(X^n)] \geq 3/4$.

On the other hand, the Good–Turing estimator will obey $\mathbb{E}[\widehat{M}_{\mathsf{GT}}(X^n)] \leq p$. This is because for all $i \in [n]$, we have

$$\mathbb{E}[\widehat{M}_{\mathsf{GT}}(X^n)] = \mathbb{E}[\widehat{M}^{(i)}] = \Pr\{X_i \notin \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\} \leq \Pr\{X_i \neq X_{i-1}\} \leq p. \tag{11}$$

Consequently, for $k \geq Cn$ and $p \leq 1/4$, we have

$$\mathbb{E}[|M_\pi(X^n) - \widehat{M}_{\mathsf{GT}}(X^n)|] \overset{\text{(i)}}{\geq} |\,\mathbb{E}[M_\pi(X^n) - \widehat{M}_{\mathsf{GT}}(X^n)]| \geq \frac{3}{4} - p \geq \frac{1}{2}, \tag{12}$$

where step (i) follows by Jensen's inequality. In words, the Good–Turing estimator has constant, non-vanishing bias for sticky Markov chains in which the stationary distribution is uniform on a large state space. In addition, we have

$$\begin{aligned}\Pr\left\{|M_\pi(X^n) - \widehat{M}_{\mathsf{GT}}(X^n)| \geq 1/4\right\} &\overset{\text{(i)}}{\geq} \Pr\left\{|M_\pi(X^n) - \widehat{M}_{\mathsf{GT}}(X^n)| \geq \frac{1}{2} \cdot \mathbb{E}[Z]\right\} \\ &\overset{\text{(ii)}}{\geq} \frac{1}{4} \cdot \frac{1}{4} = \frac{1}{16},\end{aligned}$$

where step (i) follows by Eq. (12), and step (ii) follows from the Paley–Zygmund inequality $\Pr(Z \geq \theta\,\mathbb{E}[Z]) \geq (1-\theta)^2\,\mathbb{E}[Z]^2$ for any random variable $Z \in [0, 1]$, applied with $\theta = 1/2$. Thus, the Good–Turing estimator is *inconsistent*, in that its error $|M_\pi(X^n) - \widehat{M}_{\mathsf{GT}}(X^n)|$ cannot converge in probability to zero even as $n \to \infty$. This phenomenon is also empirically illustrated in Figure 2 in Section 6.

While the inconsistency of the Good–Turing estimator is unfortunate, we next build on some important design principles sketched here in order to develop a consistent estimator.

3. Methodology: The Windowed Good–Turing estimator

In this section, we describe a natural modification of the Good–Turing estimator, interpreted through the leave-one-out lens (8), that mitigates the above issues and estimates the stationary missing mass at a minimax-optimal rate.

3.1 A first step: Modifying the "leave-one-out" estimator to reduce its bias

The central issue with the original leave-one-out estimator $\widehat{M}^{(i)}$ (Eq. (8)) when applied to Markov chains lay in the strong dependencies induced between adjacent samples of the chain: As demonstrated by Eq. (11), successive samples X_i and X_{i-1} are tightly coupled through the structure of the transition kernel and very far from being independent. To mitigate this issue, we modify the leave-one-out estimator to a "leave-a-window-out" estimator that removes the samples that are adjacent to X_i before computing the corresponding estimator. We first illustrate the idea for $i = n$ for simplicity. Recall that $\widehat{M}^{(n)} = \mathbb{I}\{X_n \notin \{X_1, \ldots, X_{n-1}\}\}$. Instead of using the leave-one-out subsequence $(X_1, \ldots, X_{n-1})$ as a proxy for X^n, we use the slightly smaller subsequence $(X_1, \ldots, X_{n-\tau})$. As long as we choose some fixed $\tau \gg \mathsf{T}_{\mathsf{mix}}$, we should expect that $(X_1, \ldots, X_{n-\tau})$ is nearly independent of X_n. Accordingly, we define the estimator

$$\widehat{M}_\tau^{(n)} := \mathbb{I}\{X_n \notin \{X_1, \ldots, X_{n-\tau}\}\}. \tag{13}$$

To develop some heuristic intuition, suppose for the moment that X_n were exactly independent of $(X_1, \ldots, X_{n-\tau})$, and recall that the marginal distribution of X_n is the stationary measure π. Then by construction, the random variable $\widehat{M}_\tau^{(n)}$ would be an unbiased estimator of $M_\pi(X^{n-\tau})$, which one should expect, in turn, to be close to the desired missing mass $M_\pi(X^n)$ provided τ is not too large. In other words, the estimate $\widehat{M}_\tau^{(n)}$ will have a small bias that can be controlled via a suitable choice of window size τ.

3.2 WingIt: Averaging an ensemble of windowed leave-one-out estimators

Above, we have sketched how to produce a single random variable $\widehat{M}_\tau^{(n)}$ with small estimation bias. However, we still have the issue of variance. How do we construct multiple estimators like $\widehat{M}_\tau^{(n)}$ and average over them as in Eq. (9)? The natural idea to construct the i-th such estimator is to inspect the definition (13), replace X_n with X_i, and the set $\{X_1, \ldots, X_{n-\tau}\}$ with the portion of the sequence X^n that should behave as nearly independent of X_i. Concretely, for each $i \in [n]$, define the index sets

$$\mathcal{D}_i = \{k \in [n] : |k - i| < \tau\} \quad \text{and} \quad \mathcal{I}_i = [n] \setminus \mathcal{D}_i. \tag{14}$$

In words the set $\mathcal{D}_i$ contains indices that are close to i, so that if $\tau \gg \mathsf{t}_{\mathsf{mix}}$ then we should expect $\boldsymbol{X}_{\mathcal{D}_i}$ to be the set of random variables in the sequence that depend significantly on X_i. On the other hand, the complementary set $\mathcal{I}_i$ is the index set of random variables that are nearly independent of X_i. The above intuition then leads naturally to the estimator

$$\widehat{M}_\tau^{(i)} := \mathbb{I}\{X_i \notin \boldsymbol{X}_{\mathcal{I}_i}\}, \tag{15}$$

which generalizes $\widehat{M}_\tau^{(n)}$ to any index i. Finally, we combine these estimates to reduce variance, creating the final estimator

$$\widehat{M}_{\textsc{WingIt}}(\tau) = \frac{1}{n} \sum_{i=1}^{n} \widehat{M}_\tau^{(i)}. \tag{16}$$

Note that if $\tau = 1$, then we recover the Good–Turing estimator (9). See Figure 5 later in the paper for an illustration of the windowing procedure.

3.3 Linear time implementation of WingIt

As presented in Eqs. (15) and (16), the WingIt estimator can be naively implemented in $\mathcal{O}(n^2)$ time, since each estimator $\widehat{M}_\tau^{(i)}$ can be computed by searching through the entire sequence. In this section, we show that, in fact, the entire estimator $\widehat{M}_{\textsc{WingIt}}(\tau)$ can be computed in time $\mathcal{O}(n)$ time for any value of the window size τ. The computational complexity of the WingIt estimator is thus comparable to that of the vanilla Good–Turing estimator (7). Given a sequence X^n and a natural number τ, the WingIt estimator 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

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: $\widehat{M} \leftarrow 0$
9: **for** $i = 1, \ldots, n$ **do**
10: **if** `locations`$[X_i]$.first $> i - \tau$ **and** `locations`$[X_i]$.last $< i + \tau$ **then**
11: $\widehat{M} \leftarrow \widehat{M} + 1/n$
12: **end if**
13: **end for**
14: **return** $\widehat{M}$

Correctness: It suffices to show that the condition in Step 10 evaluates to `True` only when $\mathbb{I}\{X_i \notin \boldsymbol{X}_{\mathcal{I}_i}\} = 1$. By construction, the list `locations`$[X_i]$ contains indices $k \in [n]$ sorted in increasing order such that $X_k = X_i$. So, if the first (i.e. smallest) and last (i.e., largest) element of the list `locations`$[X_i]$ are within $(i - \tau, i + \tau)$, then we have that $X_i \notin \boldsymbol{X}_{\mathcal{I}_i}$.

Running time: The for loop in Steps 2–7 requires a single pass through the data. In the for loop in Steps 9–12, for each value of i, we access only the first and last element of the list `locations`$[X_i]$, which takes two operations in a Python implementation. The total running time of both loops is therefore $\mathcal{O}(n)$, resulting in an overall algorithm that runs in linear time.

Memory: The memory requirement is dominated by the dictionary creation in Steps 2–7. We create a list for each dictionary key, and there are at most n keys. The sum of sizes of all lists in the dictionary is equal to the number of elements observed, i.e., n. So the memory, assuming each element of $[n]$ can be stored in constant space, is $\mathcal{O}(n)$.

Having proved that the WingIt estimator can be computed using linear time and space, we now turn to studying its estimation error properties.

4. Theoretical results on missing mass estimation

This section presents a risk guarantee for $\widehat{M}_{\textsc{WingIt}}$ and a variance bound on the estimand $M_\pi(X^n)$.

4.1 Risk guarantee for the WingIt estimator

We first provide a guarantee for the risk of the estimator $\widehat{M}_{\textsc{WingIt}}$ (16). Recall the definition (5) of the mixing time up to arbitrary total variation $\mathsf{t}_{\mathsf{mix}}(\epsilon)$ and that we write $\mathsf{T}_{\mathsf{mix}} = \mathsf{t}_{\mathsf{mix}}(1/4)$.

Theorem 1 *Suppose we choose $\tau \geq \mathsf{t}_{\mathsf{mix}}\left((\mathsf{T}_{\mathsf{mix}}/n) \wedge 1/4\right)$, and let $\widehat{M}_{\textsc{WingIt}}(\tau)$ denote the estimator defined in Eq.* (16)*. Then there is an absolute positive constant C such that*

$$\mathsf{MSE}(\widehat{M}_{\textsc{WingIt}}(\tau), M_\pi) \leq C \cdot \frac{\tau}{n} \wedge 1. \tag{17}$$

In the special case of i.i.d sequences, the chain mixes in one step, our estimator specializes to Good–Turing, and we may set $\tau = 1$ in Theorem 1 to recover existing guarantees for the Good–Turing estimator of missing mass on i.i.d. sequences (McAllester and Schapire, 2000; Rajaraman et al., 2017). A few remarks on the general Markov case are now in order. We assume that $n \geq 4\mathsf{T}_{\mathsf{mix}}$ in all the remarks below; if $n < 4\mathsf{T}_{\mathsf{mix}}$, then the RHS in Eq. (17) reduces to a universal constant and conversely, it is straightforward to show that consistent estimation is impossible (see footnote 3 below). Let us now proceed to our discussion.

First, observe that if $n \geq 4\mathsf{T}_{\mathsf{mix}}$, it follows from the mixing condition in Eqs. (5) and (6) that $\mathsf{t}_{\mathsf{mix}}\left((\frac{\mathsf{T}_{\mathsf{mix}}}{n}) \wedge 1/4\right) \leq \mathsf{T}_{\mathsf{mix}} \cdot \log(n/\mathsf{T}_{\mathsf{mix}})$. Therefore, setting the window size $\tau \asymp \mathsf{T}_{\mathsf{mix}} \cdot \log(n/\mathsf{T}_{\mathsf{mix}})$, we obtain

$$\mathsf{MSE}(\widehat{M}_{\textsc{WingIt}}(\tau), M_\pi) \lesssim \frac{\mathsf{T}_{\mathsf{mix}}}{n} \cdot \log(n/\mathsf{T}_{\mathsf{mix}}),$$

so that the MSE is on the order $\mathcal{O}\left(\frac{\log n^*}{n^*}\right)$ with $n^* = n/\mathsf{T}_{\mathsf{mix}}$ denoting the effective sample size. Note that by Markov's inequality, we immediately obtain that for all $\epsilon > 0$,

$$\Pr\left\{|\widehat{M}_{\textsc{WingIt}}(\tau) - M_\pi(X^n)| \geq \epsilon\right\} \leq C\frac{\mathsf{T}_{\mathsf{mix}}}{n\epsilon^2} \cdot \log(n/\mathsf{T}_{\mathsf{mix}}),$$

thereby showing that $\widehat{M}_{\textsc{WingIt}}(\tau)$ is a consistent estimator of M_π provided τ is chosen appropriately.

It is worth remarking at this juncture that the window size τ is a hyperparameter in our algorithm, and Theorem 1 holds provided it is chosen in a data-independent fashion but satisfies the (non-random) bound $\tau \geq \mathsf{t}_{\mathsf{mix}}\left((\mathsf{T}_{\mathsf{mix}}/n) \wedge 1/4\right)$. In practice, we may not know $\mathsf{T}_{\mathsf{mix}}$ (even up to a constant factor) and one may need to tune τ in a data-dependent manner. In this case, Theorem 1 does not apply as is, but our proof techniques may still be useful in showing that a data-dependent procedure is valid. Note that estimators for the mixing time are available in the literature, e.g. for reversible Markov chains (Hsu et al., 2019), and these mixing time estimates could be used to tune τ. We sketch a different data-dependent tuning method in Section 6. Theoretically analyzing such estimators with data-dependent τ is an important direction for future work.

Next, we remark on the issue of optimality. Chandra et al. (2022) proved a minimax lower bound of order $\Omega((np)^{-1})$ on the mean squared error of estimating missing mass in sticky chains of the form (10). As remarked on before (see Lemma 10), such chains have a mixing time $\mathsf{T}_{\mathsf{mix}} \asymp p^{-1}$, so this yields a lower bound of $\Omega(\mathsf{T}_{\mathsf{mix}}/n)$ for such chains. Theorem 1 matches this lower bound up to the logarithmic factor $\log(n/\mathsf{T}_{\mathsf{mix}})$ and further holds for *all* chains of mixing time[2] at most $\mathsf{T}_{\mathsf{mix}}$.

More formally, define the class of Markov chains that mix in time at most T, as

$$\mathcal{P}_{\mathsf{mix}}(T) := \{\text{Markov chain } (\boldsymbol{P}, \pi) : \text{ mixing time of chain } \mathsf{T}_{\mathsf{mix}} \text{ is at most } T\}.$$

Theorem 1 implies that for a universal constant $C > 0$, we have the worst-case upper bound

$$\sup_{(\boldsymbol{P},\pi)\in\mathcal{P}_{\mathsf{mix}}(T)} \mathsf{MSE}(\widehat{M}_{\textsc{WingIt}}(2T\log n), M_\pi) \le C \cdot \frac{T\log(n/T)}{n}. \tag{18a}$$

On the other hand, we may state[3] the minimax lower bound (Chandra et al., 2022, Theorem 2) as the following: There is a universal constant $c > 0$ such that if $n \ge 2T\log n$ then for any estimator $\widehat{M}$ that is a measurable function of the observations X^n, we must have

$$\sup_{(\boldsymbol{P},\pi)\in\mathcal{P}_{\mathsf{mix}}(T)} \mathsf{MSE}(\widehat{M}, M_\pi) \ge c \cdot \frac{T}{n}. \tag{18b}$$

Taken together, Eqs. (18a) and (18b) thus imply that in the regime[4] $n \gtrsim T\log n$, the WingIt estimator is information-theoretically minimax optimal up to a logarithmic factor in n. Removing this logarithmic factor is an interesting open problem, and will likely require new ideas both in terms of algorithm design and analysis.

Finally, we comment on our analysis path, which is significantly different from the related literature on missing mass estimation from an i.i.d. sequence. As alluded to before, a natural and popular method to analyze estimators of the missing mass in the i.i.d. setting (McAllester and Schapire, 2000, 2001) is to exploit concentration of the estimand and write

$$\mathsf{MSE}(\widehat{M}, M_\pi) \le 3|\mathop{\mathbb{E}}_{X^n}[\widehat{M}(X^n)] - \mathop{\mathbb{E}}_{X^n}[M_\pi(X^n)]|^2 + 3\,\mathrm{var}(\widehat{M}(X^n)) + 3\,\mathrm{var}(M_\pi(X^n)), \tag{19}$$

which can be obtained by adding and subtracting terms and using the elementary inequality $(a+b+c)^2 \le 3(a^2+b^2+c^2)$. Operationally, therefore, analyzing the MSE of the estimator relies in itself on understanding the variance of the missing mass random variable $M_\pi(X^n)$, which has nothing to do with the estimator. In the Markovian case, it appears challenging to control $\mathrm{var}(M_\pi(X^n))$ by straightforward means. Other analysis techniques for missing mass

2. For the specific class of sticky chains, Theorem 1 would yield the rate of $\mathcal{O}((np)^{-1}\log(np))$, which matches the lower bound of Chandra et al. (2022) up to a factor $\log(np)$.
3. Such a statement follows from noting that in Chandra et al. (2022, Eq. (6)): (a) We can set $T = 2/(1-\alpha)$, and (b) When $n \ge \frac{2\log n}{1-\alpha} = 2T\log n$, the second term on the RHS can be made less than half the first term.
4. In the regime $n \le \mathsf{T}_{\mathsf{mix}}$, the worst case risk of any estimator can be shown to be lower bounded by a constant, so our estimator is also trivially minimax-optimal in this regime.

estimation in the i.i.d. setting (e.g. Rajaraman et al., 2017; Chandra et al., 2019) work with an exact decomposition of the MSE expressed as a sum of weighted indicators over pairs of elements $x, x' \in \mathcal{X}$, and use the i.i.d. assumption to bound these terms in a precise fashion. One such property that is used to show concentration is the negative associativity of certain random variables (McAllester and Ortiz, 2003), and we do not expect this property to hold for general Markov chains.

In contrast to these approaches, we begin with a nonstandard decomposition of the MSE by conditioning on the sequence X^n, and our argument deviates significantly from Eq. (19). Additionally, owing to the structure of our estimator (16), we must compare our random sequence X^n to suitably modified random sequences with windows of random variables left out and/or replaced by independent copies; we do so by proving certain total variation bounds for Markov bridges, in Lemmas 13 and 14.

4.2 Variance of missing mass functional $M_\pi(X^n)$

The analysis path that we sketched above circumvents needing to control the variance of the estimand, i.e. $\mathrm{var}(M_\pi(X^n))$. Nevertheless, and somewhat surprisingly, analyzing various properties of the estimator $\widehat{M}_{\textsc{WingIt}}$ allows us to indirectly upper bound $\mathrm{var}(M_\pi(X^n))$. We state this result as the following theorem, which could be of independent interest.

Theorem 2 *There is an absolute positive constant C such that*

$$\mathrm{var}(M_\pi(X^n)) \leq C \cdot \frac{\mathsf{T}_{\mathsf{mix}}{}^2 \cdot \log(1 + n/\mathsf{T}_{\mathsf{mix}})}{n} \wedge 1. \tag{20}$$

Theorem 2 is proved in Section 7.3. En route, we control the variance of the estimator $\widehat{M}_{\textsc{WingIt}}(\tau)$ by proving that it satisfies a certain stability (i.e. bounded differences) property for all values of τ—see Lemma 7, which may be of independent interest. Intuitively speaking, the random variable $\widehat{M}_{\textsc{WingIt}}(\tau)$ satisfies a bounded differences property with respect to the sequence X^n since if τ small, the impact of changing one coordinate is local[5]. Having said that, we conjecture that the bound (20) can be improved by replacing $\mathsf{T}_{\mathsf{mix}}{}^2$ by $\mathsf{T}_{\mathsf{mix}}$, but doing so will require different techniques since the bounded differences inequality in Lemma 7 is tight (see Remark 8).

The most related result to Theorem 2 was proved by Skorski (2020), who showed a *one-sided* tail bound on $M_\pi(X^n)$ in terms of the hitting time of large sets of the Markov chain. Even though the hitting time of large sets is comparable to the mixing time (Oliveira, 2012; Peres and Sousi, 2015), the main result of Skorski (2020) cannot, strictly speaking, be compared with Theorem 2. On the one hand, Theorem 2 implies the two-sided polynomial tail bound

$$\Pr\{|M_\pi(X^n) - \mathbb{E}[M_\pi(X^n)]| \geq \epsilon\} \leq C \cdot \frac{\mathsf{T}_{\mathsf{mix}}{}^2 \log(1 + n/\mathsf{T}_{\mathsf{mix}})}{n\epsilon^2} \text{ for all } \epsilon > 0,$$

which can be obtained via direct application of Markov's inequality. On the other hand, Skorski (2020, Corollary 1) provides a stronger exponentially decaying bound depending linearly

5. In the extreme case where the window length is 1, we recover the well-known bounded differences property of the vanilla Good–Turing estimator (McAllester and Schapire, 2000).

on $\mathsf{T}_{\mathsf{mix}}$, but only on the upper tail of the random variable and without centering it at its expectation. Consequently, a variance bound cannot be extracted from that result.

Having discussed estimation guarantees for the missing mass, we now turn to the problem of estimating small-count probabilities.

5. Estimating stationary mass of elements with frequency at most ζ

In this section, we show that the idea behind the WingIt estimator can be applied robustly to estimate not only the missing mass functional $M_\pi(X^n)$, but also the mass of all elements that occur *at most* ζ times (3). To define this functional formally, let $N_x(X_P) = N_x(\boldsymbol{X}_P) := \sum_{i\in P} \mathbb{I}\{X_i = x\}$ denote the number of occurrences of the element $x \in \mathcal{X}$ in the (sub)-sequence X_P and (sub)-set $\boldsymbol{X}_P$. Then, the mass of all elements that occur *exactly* ζ times is defined as

$$M_{\pi,\zeta}(X^n) := \sum_{x\in\mathcal{X}} \pi_x \cdot \mathbb{I}\{N_x(X^n) = \zeta\}. \tag{21}$$

We focus on the mass of all elements that occur *at most* ζ times (cf. the definition in Eq. (3))

$$M_{\pi,\le\zeta}(X^n) := \sum_{x\in\mathcal{X}} \pi_x \cdot \mathbb{I}\{N_x(X^n) \le \zeta\}. \tag{22}$$

Clearly, both Eq. (21) and Eq. (22) recover the missing mass functional $M_\pi(X^n)$ when $\zeta = 0$. For small ζ, both of these functionals provide more fine-grained information about the mass placed by the stationary distribution on low-frequency elements of X^n. It is worth noting that the functional (22) is directly related to the discovery probability (Lijoi et al., 2007; Favaro et al., 2012), whereby we are interested in the probability of "discovering" in the test sample an element that appeared rarely (i.e. $\le \zeta$ times) in the training sample .

We now define a natural extension of the WingIt estimator for the functional $M_{\pi,\le\zeta}(X^n)$ that retains the leave-a-window-out principle. In particular, recalling our notation from Eq. (14), we generalize the missing mass estimator $\widehat{M}_\tau^{(i)}$ (15) via

$$\widehat{M}^{(i)}_{\tau,\le\zeta} := \mathbb{I}\{N_{X_i}(\boldsymbol{X}_{\mathcal{I}_i}) \le \zeta\}, \quad \text{and construct the estimator} \quad \widehat{M}_{\textsc{WingIt},\le\zeta}(\tau) := \frac{1}{n}\sum_{i=1}^{n} \widehat{M}^{(i)}_{\tau,\le\zeta}. \tag{23}$$

The following theorem shows that the estimator $\widehat{M}_{\textsc{WingIt},\le\zeta}(\tau)$ has small MSE.

Theorem 3 *Suppose we choose $\tau \ge \mathsf{t}_{\mathsf{mix}}((\mathsf{T}_{\mathsf{mix}}/n) \wedge 1/4)$, and let $\widehat{M}_{\textsc{WingIt},\le\zeta}(\tau)$ denote the estimator defined in Eq. (23). Then there is an absolute positive constant C such that*

$$\mathsf{MSE}(\widehat{M}_{\textsc{WingIt},\le\zeta}(\tau), M_{\pi,\le\zeta}) \le C \cdot \frac{(\zeta+1)\tau}{n} \wedge 1. \tag{24}$$

Theorem 3 is proved in Section 7.4; a few remarks on the result follow. First, note that by setting $\zeta = 0$, Theorem 3 recovers Theorem 1, our result for missing mass, since the

estimator $\widehat{M}_{\textsc{WingIt},\leq 0}(\tau)$ exactly coincides with the missing mass estimator $\widehat{M}_{\textsc{WingIt}}(\tau)$. Accordingly, the proof of this theorem generalizes (and is structured similarly to) the proof of Theorem 1. Second, note that setting $\tau \asymp \mathsf{T}_{\mathsf{mix}} \log(1 + n/\mathsf{T}_{\mathsf{mix}})$ yields

$$\mathsf{MSE}(\widehat{M}_{\textsc{WingIt},\leq\zeta}(\tau), M_{\pi,\leq\zeta}) \lesssim \frac{(\zeta+1)\mathsf{T}_{\mathsf{mix}}}{n} \cdot \log(1 + n/\mathsf{T}_{\mathsf{mix}}), \tag{25}$$

so that the MSE is on the order $\mathcal{O}\left((\zeta+1) \cdot \frac{\log n^*}{n^*}\right)$ with $n^* = n/\mathsf{T}_{\mathsf{mix}}$ denoting the effective sample size. To our knowledge, a result equivalent to Eq. (25) with linear dependence on $\zeta + 1$ is not directly available in the literature, even in the i.i.d. setting. In fact, it is instructive to revisit the i.i.d. setting for small-count probabilities and compare with the Good–Turing estimator (Good, 1953).

Remark 4 *Recalling the notation $\phi_s(X^n)$ for the number of elements of the sample space $\mathcal{X}$ that occur s times in X^n, the Good–Turing estimator for the functional $M_{\pi,\zeta}(X^n)$ (21) is given by $\widehat{M}_{\mathsf{GT},\zeta} = \frac{\zeta+1}{n} \cdot \phi_{\zeta+1}(X^n)$. We can then derive an estimator for $M_{\pi,\leq\zeta}(X^n)$ by writing $\widehat{M}_{\mathsf{GT},\leq\zeta} = \sum_{s=0}^{\zeta} \widehat{M}_{\mathsf{GT},s}$. Conversely, we can derive an estimator of the exact-count functional $M_{\pi,\zeta}$ from our estimator $\widehat{M}_{\textsc{WingIt},\leq\zeta}$. In particular, we can construct the estimator $\widehat{M}_{\textsc{WingIt},\zeta}(\tau) := \widehat{M}_{\textsc{WingIt},\leq\zeta}(\tau) - \widehat{M}_{\textsc{WingIt},\leq(\zeta-1)}(\tau)$, which can be interpreted as writing*

$$\widehat{M}^{(i)}_{\tau,\zeta} := \mathbb{I}\{N_{X_i}(\boldsymbol{X}_{\mathcal{I}_i}) = \zeta\}, \quad \text{and constructing the estimator} \quad \widehat{M}_{\textsc{WingIt},\zeta}(\tau) = \frac{1}{n}\sum_{i=1}^{n} \widehat{M}^{(i)}_{\tau,\zeta}.$$

The leave-one-out perspective (McAllester and Schapire, 2001) then yields the following consequence for $\tau = 1$ in our estimator: We have $\widehat{M}_{\textsc{WingIt},\zeta}(1) = M_{\mathsf{GT},\zeta}$, and therefore $\widehat{M}_{\textsc{WingIt},\leq\zeta}(1) = M_{\mathsf{GT},\leq\zeta}$.

To our knowledge, existing analyses of the Good–Turing estimator that are tailored to exact-count estimation do not recover the small-count estimation error guarantee of Theorem 3 even in the special case of i.i.d. observations; simply translating these results to guarantees on estimating $\widehat{M}_{\mathsf{GT},\zeta}$ leads to weaker guarantees. To be concrete, applying the result of Drukh and Mansour (2005) (which is for the MSE of $\widehat{M}_{\mathsf{GT},\zeta}$) yields

$$\mathsf{MSE}(\widehat{M}_{\mathsf{GT},\leq\zeta}, M_{\pi,\leq\zeta}) \overset{\text{(i)}}{\leq} \zeta \sum_{s=0}^{\zeta} \mathsf{MSE}(\widehat{M}_{\mathsf{GT},s}, M_{\pi,s}) = \zeta \sum_{s=0}^{\zeta} \frac{\sqrt{s}}{n} + \left(\frac{s}{n}\right)^2 \lesssim \frac{\zeta^{5/2}}{n} \wedge 1, \tag{26a}$$

where step (i) follows from the (loose) inequality $(\sum_{s=0}^{\zeta} a_s)^2 \leq \zeta \sum_{s=0}^{\zeta} a_s^2$. However, setting $\tau = 1$ in Theorem 3, we see that Eq. (24) improves the guarantee (26a) even in the i.i.d. case, showing that

$$\mathsf{MSE}(\widehat{M}_{\mathsf{GT},\leq\zeta}, M_{\pi,\leq\zeta}) \overset{\text{(i)}}{=} \mathsf{MSE}(\widehat{M}_{\textsc{WingIt},\leq\zeta}(1), M_{\pi,\leq\zeta}) \lesssim (\zeta+1)/n, \tag{26b}$$

where step (i) follows from Remark 4.

Conversely, our bound on the exact-count Good–Turing estimator (obtained by setting $\tau = 1$ in Theorem 3 and appealing to Remark 4) is suboptimal in its dependence on ζ. Our bound specializes in the i.i.d. case to

$$\mathsf{MSE}(\widehat{M}_{\mathsf{GT},\zeta}, M_{\pi,\zeta}) \le 2\mathsf{MSE}(\widehat{M}_{\mathsf{GT},\le\zeta}, M_{\pi,\le\zeta}) + 2\mathsf{MSE}(\widehat{M}_{\mathsf{GT},\le(\zeta-1)}, M_{\pi,\le(\zeta-1)}) \lesssim \frac{(\zeta+1)}{n}. \tag{27a}$$

Note that the dependence on ζ is linear, compared to the following bound of Drukh and Mansour (2005) that has an improved dependence on ζ:

$$\mathsf{MSE}(\widehat{M}_{\mathsf{GT},\zeta}, M_{\pi,\zeta}) \lesssim \frac{\sqrt{\zeta+1}}{n} + \left(\frac{\zeta+1}{n}\right)^2. \tag{27b}$$

It is worth noting that the bounds (26) are on the same estimator, and the bounds (27) are on the same estimator. Our analysis technique appears to be better equipped to deal with estimation error on the small-count probabilities, i.e. $\mathsf{MSE}(\widehat{M}_{\mathsf{GT},\le\zeta}, M_{\pi,\le\zeta})$, while the analysis technique of Drukh and Mansour (2005) is better equipped to deal with estimation error on the exact-count probabilities, i.e., $\mathsf{MSE}(\widehat{M}_{\mathsf{GT},\zeta}, M_{\pi,\zeta})$.

The problems of whether the rate (26b) and its Markovian analog (25) are information-theoretically optimal for estimating the small-count[6] probability $M_{\pi,\le\zeta}$ are interesting and, to our knowledge, open, both in the i.i.d. and Markovian settings. To address this, it would be interesting to examine and carefully modify generalizations of Good–Turing estimators (e.g. Painsky (2023)) for the Markov setting.

6. Numerical experiments

In this section, we provide a set of simulations on synthetically constructed Markov chains and on natural language text in order to corroborate our theoretical results, in particular Theorem 1. Before proceeding to the experiments themselves, we describe in Section 6.1 a data-dependent tuning procedure of the window size τ in the estimator $\widehat{M}_{\textsc{WingIt}}(\tau)$.

Code and the text used for the simulations are available at Thangaraj et al. (2024).

6.1 Data-dependent tuning of window size τ

Theorem 1 prescribes that we choose the window size τ to be at least on the order $\mathsf{T}_{\mathsf{mix}} \log(1+ n/\mathsf{T}_{\mathsf{mix}})$. An important question to address is how to choose the window size τ when we do not have access to a valid upper bound on $\mathsf{T}_{\mathsf{mix}}$. We now propose a validation procedure to select τ, and test this procedure in the experiments in the sequel. For this section, assume n is divisible by 3 for notational convenience.

Given the sequence X^n, we choose a candidate window size $\widehat{\tau}$ via the following procedure. We first split the sequence into the first one-third $Z^{(1)} = (X_1, \ldots, X_{n/3})$ and the final one-third $Z^{(2)} = (X_{2n/3+1}, \ldots, X_n)$ and compute the random variable

$$\widetilde{M}(Z^{(1)}, Z^{(2)}) = \frac{1}{(n/3)} \sum_{i=2n/3+1}^{n} \mathbb{I}\{X_i \notin \{X_1, \ldots, X_{n/3}\}\}.$$

6. Note that the small ζ regime, i.e. $\zeta = o(n)$, is the interesting one; we should expect accurate estimation to be possible for large ζ since the corresponding elements appear many times.

Next, iterate $\tau = 1, 2, 4, \ldots, 2^{\lfloor \log_2(n/6) \rfloor}$ in increasing order, and compute $\widehat{M}_{\textsc{WingIt}}(\tau)$ on the sequence $Z^{(1)}$; denote this random variable by $\widehat{M}_{\textsc{WingIt}}(Z^{(1)}; \tau)$ for convenience. We then set $\widehat{\tau}$ to be the smallest τ among this set such that

$$\left| \widehat{M}_{\textsc{WingIt}}(Z^{(1)}; \tau) - \widetilde{M}(Z^{(1)}, Z^{(2)}) \right|^2 \leq \frac{C_{\mathsf{tune}}\, \tau}{(n/3)} \tag{28}$$

for a suitable choice of the constant $C_{\mathsf{tune}} > 0$. If such an inequality is not satisfied for any τ in the prescribed list, then we set $\widehat{\tau} = n/6$. While we do not prove theoretical guarantees for the tuned estimator, Appendix B provides some intuition for why this procedure is reasonable as an automatic tuning method. We next show that empirically, this tuning method is competitive with the optimally chosen window size.

6.2 Experiments on simulated Markov chains and natural language text

For the simulated Markov and natural language text sequences considered in this section, we vary the sequence length n and plot the MSE of the estimator $\widehat{M}_{\textsc{WingIt}}(\tau)$ as a function of n for different values of the window size τ. We also plot the MSE of the tuned estimator $\widehat{M}_{\textsc{WingIt}}(\widehat{\tau})$. Note once again that the special case $\tau = 1$ corresponds to the Good–Turing estimator (7). We also plot (in dashed lines) the result of the tuning procedure that we described in Section 6.1, with the constant C_{tune} in Eq. (28) set to be 1. Every point in these plots is generated by averaging the results of multiple sequences generated from the source. Throughout this section we denote $\widehat{M}(\cdot) := \widehat{M}_{\textsc{WingIt}}(\cdot)$ as shorthand.

First, and as a sanity check, we consider the case of the trivial Markov chain formed by i.i.d. samples. For generating n samples, we consider the uniform distribution over the state space $\mathcal{X} = \{1, 2, \ldots, \lfloor 1.2n \rfloor\}$, which is close to the worst-case distribution for the Good–Turing estimator. Moreover, this ensures that the missing mass $M_\pi(X^n)$ is significant. Figure 1 shows two plots. The plot on the left is that of the MSE of the estimator $\widehat{M}_{\textsc{WingIt}}(\tau)$ as a function of sequence length n for various values of the window size τ and for tuned $\widehat{\tau}$. The plot on the right shows the mean values (over the 100 runs) of the triple of random variables $(M_\pi, \widehat{M}(1), \widehat{M}(\widehat{\tau}))$ along with the 90 percentile confidence bar (5th to 95th percentile). Observe that on the one hand—and as expected in the i.i.d. setting with mixing time $\mathsf{T}_{\mathsf{mix}} = 1$—the minimum MSE is attained when $\tau = 1$, i.e., by the vanilla Good–Turing estimator (7). On the other hand, the MSE is only marginally higher for higher values of the window size τ, and all estimators appear to enjoy the same rate of decay of MSE in the sequence length n. The effect of misspecifying τ appears to become insignificant as n increases. The MSE of $\widehat{M}(\widehat{\tau})$ with tuned $\widehat{\tau}$ (dashed line) is close to the minimum attained MSE. In the plot to the right, the mean values of missing mass M_π and the estimators $\widehat{M}(1)$, $\widehat{M}(\widehat{\tau})$ are almost overlapping for all n. The confidence bars for the three quantities are shown as a wide blue bar for M_π, an orange bar for $\widehat{M}(1)$ and as a capped black line for $\widehat{M}(\widehat{\tau})$. The confidence bars, narrow to begin with, shrink to negligible lengths as n increases.

In our second experiment, we once again consider the state space $\mathcal{X} = \{1, \ldots, \lfloor 1.2n \rfloor\}$. We simulate the sticky Markov chain (10) with $p = 0.5$ and stationary distribution given by the uniform distribution on $\mathcal{X}$, i.e., $\pi_x = \frac{1}{\lfloor 1.2n \rfloor}$ for all $x \in \mathcal{X}$. As before, we simulate the performance of the WingIt estimator as a function of n for different values of window size

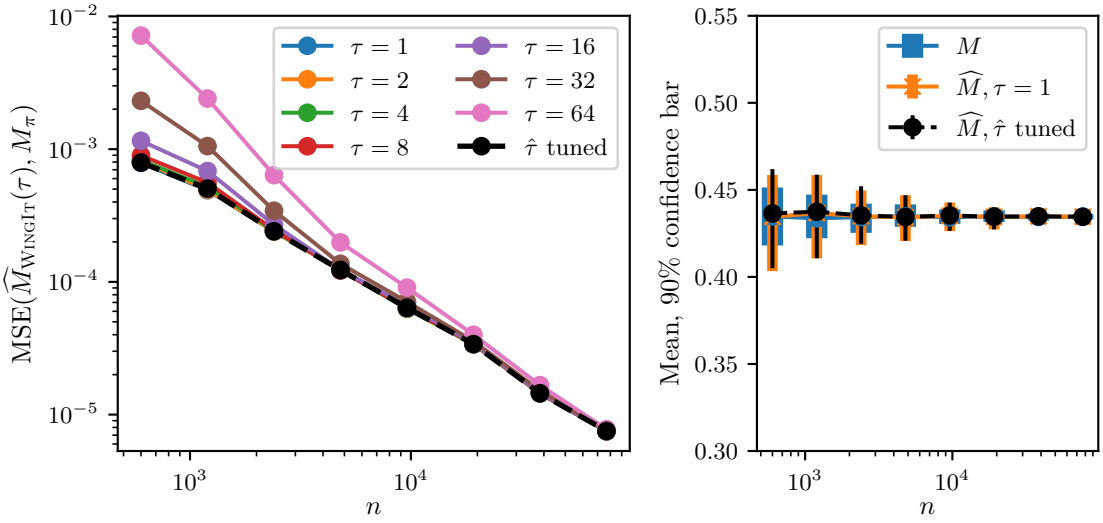

Figure 1: IID Uniform($[1.2n]$) for n samples, averaged over 100 trajectories.

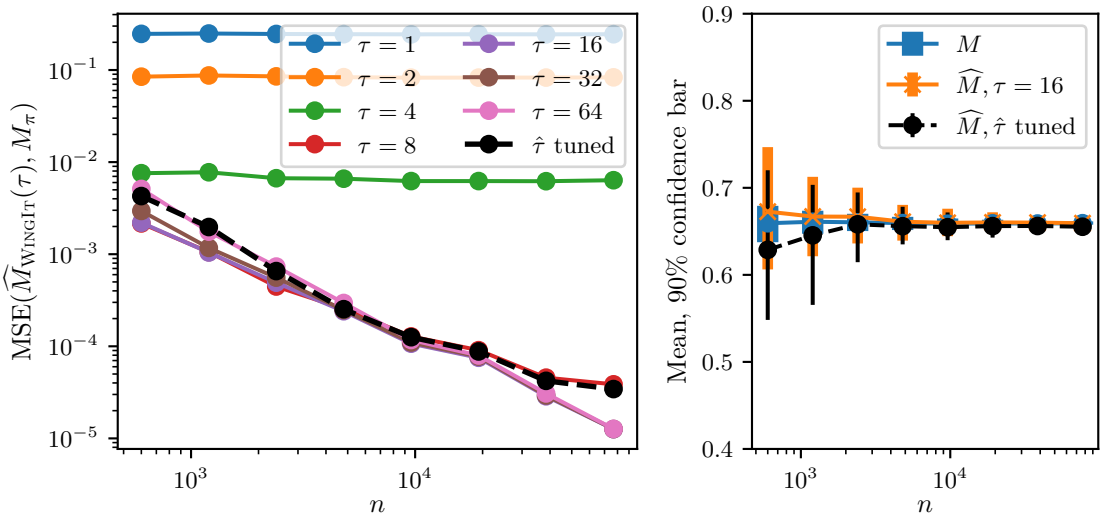

Figure 2. Sticky(0.5) with $\pi = \mathrm{Uniform}([1.2n])$ for n samples, averaged over 100 trajectories.

τ, and present our results in Figure 2. In the MSE plot to the left, we observe the following. In contrast to the previous i.i.d. example, we now find that the choice $\tau = 1$ is a poor one, and that the error of the estimator does not decay with the sample size n. As expected from the analysis presented in Section 2, the vanilla Good–Turing estimator (7) indeed suffers a constant MSE in this setting. Note that by Lemma 10, the mixing time of this chain is bounded as $\mathsf{T}_{\mathsf{mix}} \in [1, 4]$, and Theorem 1 predicts that the estimator should succeed when the window size τ is larger than $\mathsf{T}_{\mathsf{mix}}$. This prediction is borne out in simulation: While the estimator exhibits constant bias even when $\tau = 4$, when $\tau = 8$ the MSE suddenly decays in n. Further increases in the window size τ preserve the consistency of the estimator while affecting the MSE only slightly. In the plot to the right in Figure 2, we see that the mean value of missing mass coincides with the mean value of both the estimators (one with $\tau = 16$ and the other with tuned window size) for larger values of n. For $n < 2000$, the mean of the estimator $\widehat{M}_{\textsc{WingIt}}(16)$ almost coincides with missing mass, while the mean of $\widehat{M}_{\textsc{WingIt}}(\widehat{\tau})$ is smaller. The confidence bars for both estimators are wider than that of the missing mass for $n < 2000$ though they narrow quickly as n grows. These observations suggest that the tuning procedure may be improvable, particularly for small n.

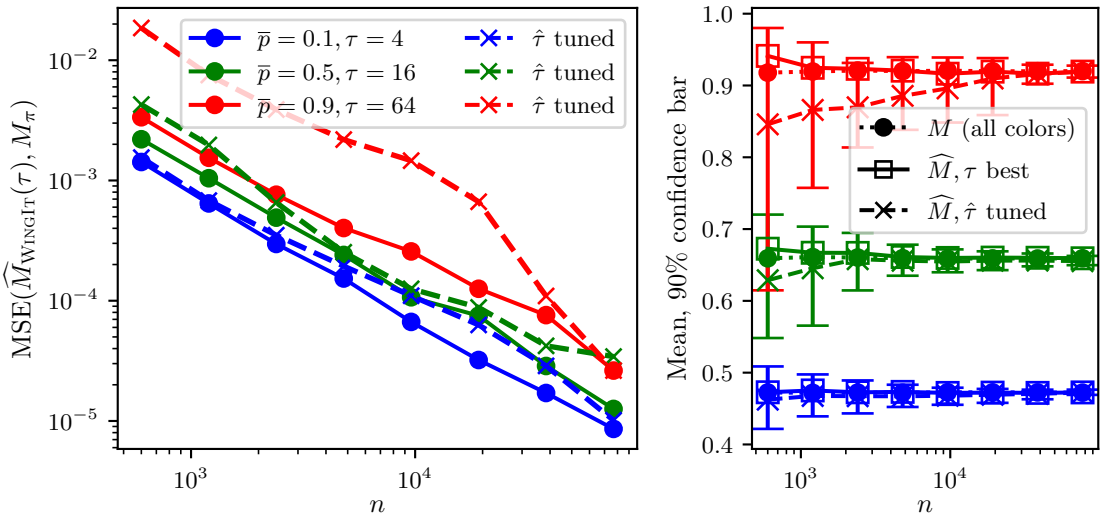

Figure 3. Sticky($\overline{p}$) with $\pi = \mathrm{Uniform}([1.2n])$ for n samples, 100 trajectories. Compared to Eq. (10), we have reparameterized as $\overline{p} = 1 - p$. To reduce clutter, confidence bar is shown only for $\widehat{M}_{\textsc{WingIt}}(\widehat{\tau})$.

In our next experiment, we simulate the sticky Markov chain (10) with $\overline{p} \triangleq 1 - p = 0.1, 0.5, 0.9$ setting the state space and stationary distribution as before. The aim of this experiment is to examine more closely the influence of the stickiness parameter $\overline{p}$ on the optimal choice of window size τ. The same simulations are performed and results are compared in Figure 3. The best (power of 2) window sizes for $\overline{p} = 0.1, 0.5, 0.9$ are observed through simulations to be, respectively, $\tau = 4, 16, 64$. The plot to the left shows the MSE of $\widehat{M}_{\textsc{WingIt}}(\tau)$ for the best observed window size and for the data-tuned window size $\widehat{\tau}$

(called "tuned"). For the least sticky chain ($\overline{p} = 0.1$), the tuned and the best estimators have overlapping MSEs for all n. As stickiness increases, the MSE of the tuned estimator marginally deviates from the best for small n. For highly sticky chains ($\overline{p} = 0.9$), the deviation persists for significantly longer. The plot to the right shows mean values and confidence bars (for the tuned estimator alone) for all three values of $\overline{p}$. The best estimator is close in mean to missing mass for all n for all $\overline{p}$. As stickiness increases, the tuning procedure appears to require increasingly larger values of n for good accuracy.

In our final experiment, presented in Figure 4, we consider the text of the novel *A Tale of Two Cities* by Charles Dickens accessed through Project Gutenberg (Dickens, 1994). All auxiliary content (preface, table of contents, chapter titles, Project Gutenberg related text) were removed and only the novel text was retained. This text was tokenized and all punctuation was removed. Titles (Miss, Mr. etc.) and names of characters that occurred as collocations with high frequency (10 of them) were merged into single tokens. The result was a sequence of $N = 136092$ tokens, numbered from 0 to $N-1$, with a vocabulary $\mathcal{X}$ (unique tokens) of size $|\mathcal{X}| = 10542$. For defining missing mass M_π, the overall frequency distribution of the N tokens was taken to be the stationary distribution π. A consecutive sequence of n tokens (Token $s+1$ to Token $s+n$ with starting point s) is considered as a trajectory of length n, i.e. X^n. For a given length n, approximately $15N/n$ trajectories were considered in the simulations with their starting points separated by $n/15$. An important feature of this data is that the Markov assumption (also known as the bigram assumption in this literature) is clearly violated, since we expect the text to have longer-range dependencies. In any case, we can run our estimator for the missing mass and compare it to the true missing mass, which can still be computed from the sequence once the stationary probability is fixed.

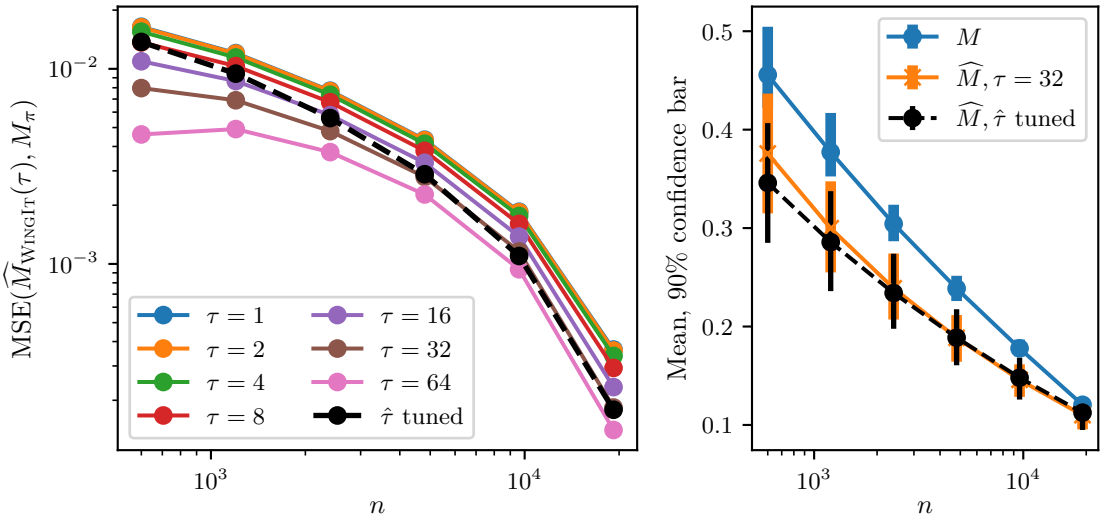

Figure 4. Text of 'A Tale of Two Cities' by Charles Dickens, Vocabulary: 10542 words, Trajectories: sequences of n words from text.

In Figure 4, we show the MSE in the plot to the left and the means with confidence bars to the right. Interestingly, all of the choices of τ that we consider yield similar MSE

performance for the WingIt estimator, and *all* of these choices appear to have a super-linear rate of decay with the sample size n. The reason for the difference in rate of decay could be because we hold $|\mathcal{X}|$ constant as we increase n in the text simulations resulting in a decrease in M_π with n. This decrease is confirmed in the plot on the right, where we see that the mean of missing mass falls from about 0.45 for $n = 600$ to about 0.1 for $n = 19200$. Further, from the right plot, we observe that the tuned estimator and the estimator with $\tau = 32$ are close to each other in mean, while deviating a bit from the mean of missing mass for small n. The aforementioned long-range dependencies in the text are likely to be causing this minor deviation. In any case, the estimator $\widehat{M}_{\text{WingIt}}(\widehat{\tau})$ appears to be accurate for all n. Overall, our experiments on this corpus demonstrate that the missing mass estimator is robust to model misspecification, and could work well even for non-Markovian sources.

7. Proofs

In this section, we present proofs of the main theorems. We begin with preliminaries in Section 7.1, which introduces additional notation and a useful reduction device that will help us analyze our estimators $\widehat{M}_{\text{WingIt}}(\tau)$ and $\widehat{M}_{\text{WingIt},\leq\zeta}(\tau)$. Technical lemmas are frequently referenced in our proofs, and their statements and proofs can be found in Appendix A.

7.1 Preliminary decompositions and notation

Suppose for convenience[7] that n is divisible by 2τ, and let $n_0 = n/(2\tau)$. Recall our single-sample estimators $\widehat{M}_\tau^{(i)}$ (Eq. (15)) and the definition (16) of the WingIt estimator. Define the "skipped" estimators

$$\widehat{M}_{\text{WingIt}}(\tau; \ell) := \frac{1}{n_0} \sum_{j=1}^{n_0} \widehat{M}_\tau^{(2\tau j - \ell)} \text{ for each } \ell = 0, \ldots, 2\tau - 1. \tag{29}$$

In words, each of these estimators averages only n_0 of the individual estimates $\widehat{M}_\tau^{(i)}$ by skipping 2τ indices at a time; this skipping induces further decorrelations. Note that we may write

$$\widehat{M}_{\text{WingIt}}(\tau) = \frac{1}{2\tau} \sum_{\ell=0}^{2\tau-1} \widehat{M}_{\text{WingIt}}(\tau; \ell)$$

by definition. Furthermore, we have

$$\begin{aligned} \mathsf{MSE}(\widehat{M}_{\text{WingIt}}(\tau), M_\pi) &= \mathbb{E}\left(\frac{1}{2\tau} \sum_{\ell=0}^{2\tau-1} \widehat{M}_{\text{WingIt}}(\tau; \ell) - M_\pi \right)^2 \\ &\overset{\text{(i)}}{\leq} \frac{1}{2\tau} \sum_{\ell=0}^{2\tau-1} \mathbb{E}\left(\widehat{M}_{\text{WingIt}}(\tau; \ell) - M_\pi \right)^2, \end{aligned} \tag{30}$$

where step (i) follows from Jensen's inequality applied to the convex function $z \mapsto z^2$.

7. Our argument extends straightforwardly without this assumption; we only make it to avoid carrying floor and ceiling notation.

Via a parallel argument, and introducing the objects

$$\widehat{M}_{\text{WINGIT},\leq\zeta}(\tau;\ell) := \frac{1}{n_0}\sum_{j=1}^{n_0} \widehat{M}_{\tau,\leq\zeta}^{(2\tau j-k)} \text{ for each } \ell = 0,\ldots,2\tau-1, \tag{31}$$

we have

$$\mathsf{MSE}(\widehat{M}_{\text{WINGIT},\leq\zeta}(\tau), M_{\pi,\leq\zeta}) \leq \frac{1}{2\tau}\sum_{\ell=0}^{2\tau-1} \mathbb{E}\left(\widehat{M}_{\text{WINGIT},\leq\zeta}(\tau;\ell) - M_{\pi,\leq\zeta}\right)^2. \tag{32}$$

Our argument to prove Theorems 1 and 3 will proceed by establishing the following proposition.

Proposition 5 *If $\tau \geq \mathsf{t}_{\mathsf{mix}}((\mathsf{T}_{\mathsf{mix}}/n)\wedge 1/4)$, then the following statements hold:*
(a) There is an absolute positive constant C such that we have

$$\mathbb{E}\left(\widehat{M}_{\text{WINGIT}}(\tau;\ell) - M_\pi\right)^2 \leq C\cdot\frac{\tau}{n}\wedge 1 \quad \textit{for all} \quad \ell = 0,\ldots,2\tau-1. \tag{33}$$

(b) There is an absolute positive constant C such that we have

$$\mathbb{E}\left(\widehat{M}_{\text{WINGIT},\leq\zeta}(\tau;\ell) - M_\pi\right)^2 \leq C\cdot\frac{(\zeta+1)\tau}{n}\wedge 1 \quad \textit{for all} \quad \ell = 0,\ldots,2\tau-1. \tag{34}$$

We prove part (a) of Proposition 5 in proving Theorem 1; see Section 7.2. We prove part (b) of Proposition 5 in proving Theorem 3; see Section 7.4.

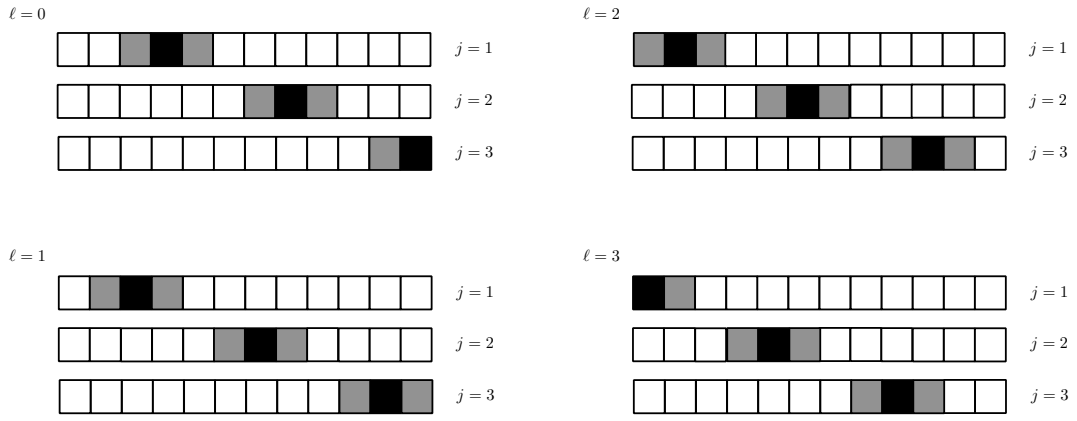

Figure 5. Schematic showing our estimator construction for $n = 12$ and $\tau = 2$, so that $n_0 = n/(2\tau) = 3$. For the various values $j \in [n_0]$ and $\ell = 0, 1, \ldots, 2\tau-1$, the index $2\tau j - \ell$ is shown in black and the window of size $\tau - 1$ on either size of it is excluded when computing the estimator $\widehat{M}_\tau^{(2\tau j-\ell)}$. The indices in color form the sets $\mathcal{D}_{2\tau j-\ell} = \mathcal{B}_{j,\ell}$, while the indices in white form the sets $\mathcal{I}_{2\tau j-\ell} = \mathcal{H}_{j,\ell}$. For each ℓ, the sets $\{\mathcal{D}_{j,\ell}\}_{j\in[n_0]}$ are non-overlapping.

Recall our index sets $\{\mathcal{D}_i\}_{i=1}^n$ and $\{\mathcal{I}_i\}_{i=1}^n$ from Eq. (14). 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 $\mathcal{H}_{j,\ell} := \mathcal{I}_{2\tau j-\ell}$ as the "dependent" and

"independent" indices, respectively, for the j-th window, or block, of size (at most) $2\tau - 1$. Mnemonically, one should view $\mathcal{B}_{j,\ell}$ as the j-th *block* of indices and $\mathcal{H}_{j,\ell} = [n] \setminus \mathcal{B}_{j,\ell}$ as the set of indices having a *hole* at block j.

In the sequel, we will pay special attention to 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 $\mathcal{H}_j := \mathcal{H}_{j,0}$. See Figure 5 for an illustration of our notation.

Recall that $\boldsymbol{X}_{[n]} = \{X_1, \ldots, X_n\}$ is the set (not sequence) of all random variables. In the sequel, we use the shorthand $\mathbb{E}_Y \equiv \mathbb{E}_{Y \sim \pi}$ and $\mathbb{E}_{Y'} \equiv \mathbb{E}_{Y' \sim \pi}$, where Y, Y' are drawn i.i.d. and independently of X^n. The notation $\mathbb{E}$ (without any subscript) is reserved for an expectation taken over all the randomness in the problem.

7.2 Proof of Theorem 1

Owing to Eq (30), it suffices to establish Proposition 5(a). As will be clear from the proof, our argument will apply to bound the MSE of the estimator $\widehat{M}_{\text{WingIt}}(\tau; \ell)$ for any $\ell = 0, \ldots, 2\tau - 1$, so we concentrate on establishing that for an absolute positive constant C and $\tau \geq \mathsf{t}_{\mathsf{mix}}\left((\mathsf{T}_{\mathsf{mix}}/n) \wedge 1/4\right)$:

$$\mathbb{E}\left(\widehat{M}_{\text{WingIt}}(\tau; 0) - M_\pi\right)^2 \leq C \cdot \frac{\tau}{n} \wedge 1. \tag{35}$$

Recall the shorthand $n_0 = n/(2\tau)$ and the notation $\mathcal{B}_j$ and $\mathcal{H}_j$ from above. Viewing X^n as fixed for the moment and writing out our estimator $\widehat{M}_{\text{WingIt}}(\tau; 0) = \frac{1}{n_0}\sum_{j=1}^{n_0} \widehat{M}_\tau^{(2\tau j)}$, we have that $\frac{1}{2}|\widehat{M}_{\text{WingIt}}(\tau; 0) - M_\pi(X^n)|^2$ is equal to

$$\begin{aligned}
&\frac{1}{2} \cdot \left| \frac{1}{n_0} \sum_{j=1}^{n_0} \mathbb{I}\left\{X_{2\tau j} \notin \boldsymbol{X}_{\mathcal{H}_j}\right\} - \mathop{\mathbb{E}}_{\substack{Y \sim \pi \\ Y \perp\!\!\!\perp X^n}} \mathbb{I}\left\{Y \notin \boldsymbol{X}_{[n]}\right\} \right|^2 \\
&\qquad \leq \underbrace{\left| \frac{1}{n_0} \sum_{j=1}^{n_0} \left(\mathop{\mathbb{E}}_{\substack{Y \sim \pi \\ Y \perp\!\!\!\perp X^n}} \mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j}\right\} - \mathop{\mathbb{E}}_{\substack{Y \sim \pi \\ Y \perp\!\!\!\perp X^n}} \mathbb{I}\left\{Y \notin \boldsymbol{X}_{[n]}\right\} \right) \right|^2}_{T_1} \\
&\qquad\qquad + \underbrace{\left| \frac{1}{n_0} \sum_{j=1}^{n_0} \left(\mathbb{I}\left\{X_{2\tau j} \notin \boldsymbol{X}_{\mathcal{H}_j}\right\} - \mathop{\mathbb{E}}_{\substack{Y \sim \pi \\ Y \perp\!\!\!\perp X^n}} \mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j}\right\} \right) \right|^2}_{T_2},
\end{aligned} \tag{36}$$

where we have added and subtracted the term $\frac{1}{n_0}\sum_{j=1}^{n_0} \mathop{\mathbb{E}}_{\substack{Y \sim \pi \\ Y \perp\!\!\!\perp X^n}} \mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j}\right\}$ inside the expression $|\widehat{M}_{\text{WingIt}}(\tau; 0) - M_\pi(X^n)|$ and used the inequality $\frac{1}{2}(a+b)^2 \leq (a^2 + b^2)$.

We now bound $\mathbb{E}[T_1]$ and $\mathbb{E}[T_2]$, in turn.

7.2.1 Bounding $\mathbb{E}[T_1]$

Note that T_1 resembles a *conditional* squared bias term. For each $j \in [n_0]$, define the random variable $P_j := \mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j}\right\} - \mathbb{I}\left\{Y \notin \boldsymbol{X}_{[n]}\right\}$. Applying Lemma 11, we have

$$P_j = \mathbb{I}\left\{Y \in \boldsymbol{X}_{\mathcal{B}_j}\right\} \cdot \mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j}\right\},$$

which is the indicator that Y appears in the sequence *only in* block $\mathcal{B}_j$.

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

$$\bigsqcup_{j' \in [n_0] \setminus j} \mathcal{B}_{j'} \subset \mathcal{H}_j. \tag{37}$$

Now suppose that for some j we have $\mathbb{I}\left\{Y \in \boldsymbol{X}_{\mathcal{B}_j}\right\} \cdot \mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j}\right\} = 1$, which implies $\mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j}\right\} = 1$ and $\mathbb{I}\left\{Y \in \boldsymbol{X}_{\mathcal{H}_j}\right\} = 0$. Then we must have

$$\begin{aligned}
\sum_{j' \in [n_0] \setminus j} \mathbb{I}\left\{Y \in \boldsymbol{X}_{\mathcal{B}_{j'}}\right\} \cdot \mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_{j'}}\right\} &\overset{\text{(i)}}{\leq} \sum_{j' \in [n_0] \setminus j} \mathbb{I}\left\{Y \in \boldsymbol{X}_{\mathcal{B}_{j'}}\right\} \\
&\leq \mathbb{I}\left\{Y \in \boldsymbol{X}_{\bigsqcup_{j' \in [n_0] \setminus j} \mathcal{B}_{j'}}\right\} \\
&\overset{\text{(ii)}}{\leq} \mathbb{I}\left\{Y \in \boldsymbol{X}_{\mathcal{H}_j}\right\} = 0,
\end{aligned}$$

where (i) follows because $\mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_{j'}}\right\} \leq 1$ and (ii) follows by Eq. (37). Thus, we have

$$\sum_{j=1}^{n_0} P_j = \sum_{j=1}^{n_0} \mathbb{I}\left\{Y \in \boldsymbol{X}_{\mathcal{B}_j}\right\} \cdot \mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j}\right\} \leq 1,$$

pointwise for every sequence X^n. Said another way, the term $\sum_{j=1}^{n_0} P_j$ is equal to the indicator that Y appears in exactly one block, and therefore must be at most equal to 1.

Putting together the pieces, we have

$$T_1 = \frac{1}{n_0^2} \left(\mathop{\mathbb{E}}_{Y} \sum_{j=1}^{n_0} P_j \right)^2 \leq \frac{1}{n_0^2},$$

and so $\mathbb{E}[T_1] \leq \frac{1}{n_0^2} \lesssim \left(\frac{\tau}{n}\right)^2 \wedge 1$.

7.2.2 Bounding $\mathbb{E}[T_2]$

We note that T_2 resembles a *conditional* variance term. Define as shorthand the random variables $Z_j := \mathbb{I}\left\{X_{2\tau j} \notin \boldsymbol{X}_{\mathcal{H}_j}\right\} - \mathbb{E}_Y \mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j}\right\}$ for all $j \in [n_0]$. Then we have

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

where the inequality follows since $Z_j \in [-1, 1]$ for all $j \in [n_0]$. Therefore, it suffices to bound the cross terms when $j \neq k$. For each $j, k \in [n_0]$ with $j \neq k$, define the random variables

$$Q_{j,k} = \mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\} - \mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_k}\right\} \tag{38}$$

The following lemma relates the expectation of the cross terms to expectations of the random variables defined above.

Lemma 6 *Suppose $\tau \geq \mathsf{t}_{\mathsf{mix}}(\epsilon)$. Then for each $j \neq k$, we have*

$$\mathbb{E}[Z_j Z_k] \leq \frac{5}{2}\,\mathbb{E}[Q_{j,k}] + \frac{5}{2}\,\mathbb{E}[Q_{k,j}] + 16\epsilon,$$

where the random variables $\{Q_{j,k}\}$ are as defined in Eq. (38)*.*

We take Lemma 6 as given for the moment and prove it in Section 7.2.3. Let us now use it to bound $\mathbb{E}[T_2]$. Applying Lemma 11, we may write $Q_{j,k} = \mathbb{I}\left\{Y \in \boldsymbol{X}_{\mathcal{H}_k \setminus \mathcal{H}_j}\right\} \cdot \mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}$.

Now consider some fixed $k \in [n_0]$. Since the sets $\{\mathcal{B}_j\}_{j=1}^{n_0}$ are non-overlapping, we have $\mathcal{H}_k \setminus \mathcal{H}_j = \mathcal{B}_j$, and

$$\mathcal{H}_j \cap \mathcal{H}_k \supset \bigsqcup_{j' \in [n_0] \setminus \{j,k\}} \mathcal{B}_{j'}.$$

If for some $j \neq k$, we have

$$\mathbb{I}\left\{Y \in \boldsymbol{X}_{\mathcal{H}_k \setminus \mathcal{H}_j}\right\} \cdot \mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\} = 1,$$

then

$$\begin{aligned}\sum_{j' \in [n_0] \setminus \{j,k\}} \mathbb{I}\left\{Y \in \boldsymbol{X}_{\mathcal{H}_k \setminus \mathcal{H}_{j'}}\right\} \cdot \mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_{j'} \cap \mathcal{H}_k}\right\} &\leq \sum_{j' \in [n_0] \setminus \{j,k\}} \mathbb{I}\left\{Y \in \boldsymbol{X}_{\mathcal{B}_{j'}}\right\} \\ &\leq \mathbb{I}\left\{Y \in \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\} = 0.\end{aligned}$$

Consequently, we have

$$\sum_{j \in [n_0] \setminus k} Q_{j,k} \leq 1.$$

Essentially, we have shown that $\sum_{j \in [n_0] \setminus k} Q_{j,k}$ is at most the indicator that Y appears in exactly one block (other than $\mathcal{B}_k$), which is at most 1.

Applying Lemma 6 and using the linearity of expectation then yields

$$\sum_{j=1}^{n_0} \sum_{\substack{k=1 \\ k \neq j}}^{n_0} \mathbb{E}[Z_j Z_k] \leq 5n_0 + 16n_0^2\epsilon.$$

Consequently, we have $\mathbb{E}[T_2] \lesssim \frac{1}{n_0} + \epsilon$. Substituting $\epsilon = \frac{\mathsf{T}_{\mathsf{mix}}}{n}$ and noting that $\tau \geq \mathsf{T}_{\mathsf{mix}}$ by assumption, we obtain $\mathbb{E}[T_2] \leq C \cdot \frac{\tau}{n} \wedge 1$.

Putting together our bounds on $\mathbb{E}[T_1]$ and $\mathbb{E}[T_2]$ establishes Theorem 1. It remains to prove Lemma 6.

7.2.3 Proof of Lemma 6

Define

$$\widetilde{Q}_{j,k} := \mathop{\mathbb{E}}_{Y}[Q_{j,k}] = \mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\} - \mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_k}\right\}] \tag{39}$$

for convenience. Note that by Lemma 11, we have $\widetilde{Q}_{j,k} = \mathbb{E}_Y[\mathbb{I}\left\{Y \in \boldsymbol{X}_{\mathcal{H}_k \setminus \mathcal{H}_j}\right\} \cdot \mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}]$, so that $\widetilde{Q}_{j,k} \in [0,1]$. We have the decomposition

$$\begin{aligned}
&Z_j Z_k \\
&= \left(\mathbb{I}\left\{X_{2\tau j} \notin \boldsymbol{X}_{\mathcal{H}_j}\right\} - \mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}] + \widetilde{Q}_{k,j}\right) \cdot \left(\mathbb{I}\left\{X_{2\tau k} \notin \boldsymbol{X}_{\mathcal{H}_k}\right\} - \mathop{\mathbb{E}}_{Y'}[\mathbb{I}\left\{Y' \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}] + \widetilde{Q}_{j,k}\right) \\
&\leq \underbrace{\left(\mathbb{I}\left\{X_{2\tau j} \notin \boldsymbol{X}_{\mathcal{H}_j}\right\} - \mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}]\right) \cdot \left(\mathbb{I}\left\{X_{2\tau k} \notin \boldsymbol{X}_{\mathcal{H}_k}\right\} - \mathop{\mathbb{E}}_{Y'}[\mathbb{I}\left\{Y' \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}]\right)}_{U_{j,k}} \\
&\qquad + \widetilde{Q}_{j,k} + \widetilde{Q}_{k,j} + \widetilde{Q}_{j,k} \cdot \widetilde{Q}_{k,j} \\
&\overset{\text{(i)}}{\leq} U_{j,k} + \frac{3}{2}(\widetilde{Q}_{j,k} + \widetilde{Q}_{k,j}).
\end{aligned}$$

Here step (i) follows by the following sequence of algebraic inequalities: Given that every $\widetilde{Q}_{j,k}$ is bounded in the range $[0,1]$, we have $\widetilde{Q}_{j,k} \cdot \widetilde{Q}_{k,j} \leq \sqrt{\widetilde{Q}_{j,k} \cdot \widetilde{Q}_{k,j}} \leq \frac{1}{2}(\widetilde{Q}_{j,k} + \widetilde{Q}_{k,j})$.

It remains to establish that $\mathbb{E}[U_{j,k}] \leq \mathbb{E}_{X^n}[\widetilde{Q}_{j,k}] + \mathbb{E}_{X^n}[\widetilde{Q}_{k,j}] + 16\epsilon$. We have the further decomposition

$$\begin{aligned}
\mathbb{E}[U_{j,k}] = &\underbrace{\mathbb{E}[\mathbb{I}\left\{X_{2\tau j} \notin \boldsymbol{X}_{\mathcal{H}_j}\right\} \cdot \mathbb{I}\left\{X_{2\tau k} \notin \boldsymbol{X}_{\mathcal{H}_k}\right\}]}_{U_1} - \underbrace{\mathop{\mathbb{E}}_{X^n}\left[\mathbb{I}\left\{X_{2\tau j} \notin \boldsymbol{X}_{\mathcal{H}_j}\right\} \cdot \mathop{\mathbb{E}}_{Y'}[\mathbb{I}\left\{Y' \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}]\right]}_{U_2} \\
&- \underbrace{\mathop{\mathbb{E}}_{X^n}\left[\mathbb{I}\left\{X_{2\tau k} \notin \boldsymbol{X}_{\mathcal{H}_k}\right\} \cdot \mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}]\right]}_{U_3} + \underbrace{\mathop{\mathbb{E}}_{X^n}\left[\mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}] \cdot \mathop{\mathbb{E}}_{Y'}[\mathbb{I}\left\{Y' \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}]\right]}_{U_4}
\end{aligned} \tag{40}$$

We now bound each of the above terms in turn.

To begin, we notice that $\mathbb{I}\left\{X_{2\tau j} \notin \boldsymbol{X}_{\mathcal{H}_j}\right\} \leq \mathbb{I}\left\{X_{2\tau j} \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}$ and $\mathbb{I}\{X_{2\tau k} \notin \boldsymbol{X}_{\mathcal{H}_k}\} \leq \mathbb{I}\left\{X_{2\tau k} \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}$ to bound U_1 as

$$\begin{aligned}
U_1 &\leq \mathbb{E}[\mathbb{I}\left\{X_{2\tau j} \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\} \cdot \mathbb{I}\left\{X_{2\tau k} \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}] \\
&\overset{\text{(i)}}{\leq} \mathbb{E}[\mathbb{I}\left\{Y' \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\} \cdot \mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}] + 8\epsilon \\
&\overset{\text{(ii)}}{=} \mathop{\mathbb{E}}_{X^n}\left\{\mathop{\mathbb{E}}_{Y'}[\mathbb{I}\left\{Y' \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}] \cdot \mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}]\right\} + 8\epsilon.
\end{aligned} \tag{41}$$

Here, step (i) uses Lemma 14 (applied with $i_1 = 2\tau \min\{j,k\}, i_2 = 2\tau \max\{j,k\}$, and noting that $i_2 - i_1 \geq 2\tau$ for all $j \neq k$) 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

$$f(X_{2\tau j}; \boldsymbol{X}_{\mathcal{H}_j}) := \mathbb{I}\left\{X_{2\tau j} \notin \boldsymbol{X}_{\mathcal{H}_j}\right\} \cdot \mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_k \cap \mathcal{H}_j}\right\}],$$

which is bounded in the range $[0,1]$. Since $\tau \geq \mathsf{t}_{\mathsf{mix}}(\epsilon)$, we may apply Lemma 13 (applied with $i = 2j\tau$) to obtain $|\mathbb{E}[f(X_{2\tau j}; \boldsymbol{X}_{\mathcal{H}_j})] - \mathbb{E}[f(Y'; \boldsymbol{X}_{\mathcal{H}_j})]| \leq 4\epsilon$. Thus,

$$\begin{aligned} U_2 &\geq \mathop{\mathbb{E}}_{X^n}\left[\mathop{\mathbb{E}}_{Y'}[\mathbb{I}\left\{Y' \notin \boldsymbol{X}_{\mathcal{H}_j}\right\}] \cdot \mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}]\right] - 4\epsilon \\ &\overset{\text{(i)}}{\geq} \mathop{\mathbb{E}}_{X^n}\left[\mathop{\mathbb{E}}_{Y'}[\mathbb{I}\left\{Y' \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}] \cdot \mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}]\right] - \mathop{\mathbb{E}}_{X^n}[\widetilde{Q}_{k,j}] - 4\epsilon, \end{aligned} \tag{42}$$

where step (i) follows because

$$\begin{aligned} &\mathop{\mathbb{E}}_{Y'}[\mathbb{I}\left\{Y' \notin \boldsymbol{X}_{\mathcal{H}_j}\right\}] \cdot \mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}] \\ &\quad = \mathop{\mathbb{E}}_{Y'}[\mathbb{I}\left\{Y' \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}] \cdot \mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}] + \mathop{\mathbb{E}}_{Y'}[\mathbb{I}\left\{Y' \notin \boldsymbol{X}_{\mathcal{H}_j}\right\}] \cdot \mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}] \\ &\qquad\qquad - \mathop{\mathbb{E}}_{Y'}[\mathbb{I}\left\{Y' \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}] \cdot \mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}] \\ &\quad = \mathop{\mathbb{E}}_{Y'}[\mathbb{I}\left\{Y' \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}] \cdot \mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}] - \widetilde{Q}_{k,j} \mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}] \\ &\quad \geq \mathop{\mathbb{E}}_{Y'}[\mathbb{I}\left\{Y' \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}] \cdot \mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}] - \widetilde{Q}_{k,j} \end{aligned}$$

with the last inequality holding because of the inclusion $\mathbb{E}_Y[\mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}] \in [0,1]$.

By an identical argument to the above, we have

$$U_3 \geq \mathop{\mathbb{E}}_{X^n}\left[\mathop{\mathbb{E}}_{Y'}[\mathbb{I}\left\{Y' \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}] \cdot \mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{Y \notin \boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}\right\}]\right] - \mathop{\mathbb{E}}_{X^n}[\widetilde{Q}_{j,k}] - 4\epsilon. \tag{43}$$

Putting Eqs. (41), (42) and (43) together with the definition of U_4 and performing the requisite cancellations, we have

$$\mathbb{E}[U_{j,k}] = U_1 - U_2 - U_3 + U_4 \leq \mathop{\mathbb{E}}_{X^n}[\widetilde{Q}_{j,k}] + \mathop{\mathbb{E}}_{X^n}[\widetilde{Q}_{k,j}] + 16\epsilon,$$

as claimed. □

7.3 Proof of Theorem 2

Since the best constant predictor of a random variable is its expectation, we have

$$\begin{aligned} \operatorname{var}(M_\pi(X^n)) &\leq \mathbb{E}(M_\pi(X^n) - \mathbb{E}\,\widehat{M}_{\textsc{WingIt}}(\tau))^2 \\ &= \mathbb{E}\left((M_\pi(X^n) - \widehat{M}_{\textsc{WingIt}}(\tau)) + (\widehat{M}_{\textsc{WingIt}}(\tau) - \mathbb{E}\,\widehat{M}_{\textsc{WingIt}}(\tau))\right)^2 \\ &\overset{\text{(i)}}{\leq} 2\mathsf{MSE}(\widehat{M}_{\textsc{WingIt}}(\tau), M_\pi) + 2\operatorname{var}(\widehat{M}_{\textsc{WingIt}}(\tau)), \end{aligned} \tag{44}$$

where step (i) follows by using $(a+b)^2 \leq 2a^2 + 2b^2$. Throughout the rest of this proof, we will choose $\tau = \mathsf{t}_{\mathsf{mix}}\left((\mathsf{T}_{\mathsf{mix}}/n) \wedge 1/4\right) \lesssim \mathsf{T}_{\mathsf{mix}} \cdot \log(1 + n/\mathsf{T}_{\mathsf{mix}})$ (see Eq. (6)).

Bounding MSE: Applying Theorem 1 yields the direct bound

$$\mathsf{MSE}(\widehat{M}_{\textsc{WingIt}}(\tau), M_\pi(X^n)) \lesssim \tau/n \lesssim \frac{\mathsf{T}_{\mathsf{mix}}}{n} \cdot \log(1 + n/\mathsf{T}_{\mathsf{mix}}).$$

It remains to bound the variance term on the RHS of Ineq. (44).

Bounding variance of estimator: To bound the variance, we make use of the fact that the estimator $\widehat{M}_{\textsc{WingIt}}(\tau)$ satisfies a bounded differences property (Doob, 1940; McDiarmid, 1989) with respect to the variables $(X_1, \dots, X_n)$. This allows us to obtain a sub-Gaussian concentration inequality, which in turn is used to bound variance. We state this result as a lemma that may be of independent interest.

To set up some notation, let $X^{-i} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$ denote the sequence without its i-th entry, and let $X^{(i)}(x) = (X_1, \dots, X_{i-1}, x, X_{i+1}, \dots, X_n)$ denote the sequence with x at the i-th position and X^{-i} in the remaining positions. The maximum difference witnessed by a function $f : \mathcal{X}^n \to \mathbb{R}$ at its i-th index is given by

$$\delta_i(f) := \max_{X^{-i} \in \mathcal{X}^{n-1}} \delta_i(f; X^{-i}), \qquad \text{where} \qquad \delta_i(f; X^{-i}) := \max_{x, x' \in \mathcal{X}} |f(X^{(i)}(x)) - f(X^{(i)}(x'))|.$$

For a positive (nonrandom) scalar b_i, the function f is said to satisfy a b_i bounded differences inequality at index i if $\delta_i(f) \le b_i$.

Lemma 7 *Define the shorthand $\widehat{M}^{\tau}_{\textsc{WingIt}} := \widehat{M}_{\textsc{WingIt}}(\tau)$ for convenience. For every $\tau \in [n]$, the function $x^n \mapsto \widehat{M}^{\tau}_{\textsc{WingIt}}(x^n)$ satisfies a $\frac{4\tau}{n}$ bounded differences property on all indices, in that $\max_{i \in [n]} \delta_i(\widehat{M}^{\tau}_{\textsc{WingIt}}) \le \frac{4\tau}{n}$.*

We prove Lemma 7 shortly. For the moment, applying it in conjunction with Corollary 2.10 and Remark 2.11 of Paulin (2015), we obtain that for all $t \ge 0$,

$$\Pr\left\{ |\widehat{M}_{\textsc{WingIt}}(\tau) - \mathbb{E}[\widehat{M}_{\textsc{WingIt}}(\tau)]| \ge t \right\} \le 2 \exp\left(-c \cdot \frac{n t^2}{\mathsf{T}_{\mathsf{mix}} \tau} \right)$$

for some universal constant $c > 0$. Integrating the tail bound and using $\mathbb{E}[Z] = \int_0^\infty \Pr(Z \ge z) dz$ for the non-negative random variable $Z = |\widehat{M}_{\textsc{WingIt}}(\tau) - \mathbb{E}[\widehat{M}_{\textsc{WingIt}}(\tau)]|^2$ yields that for any $\tau \in [n]$, we have $\mathrm{var}(\widehat{M}_{\textsc{WingIt}}(\tau)) \lesssim \frac{\tau \cdot \mathsf{T}_{\mathsf{mix}}}{n}$.

Using Ineq. (44), setting $\tau \asymp \mathsf{T}_{\mathsf{mix}} \log(1 + n/\mathsf{T}_{\mathsf{mix}})$, and putting together the bounds on MSE and variance yields

$$\mathrm{var}(M_\pi(X^n)) \le C \cdot \frac{{\mathsf{T}_{\mathsf{mix}}}^2}{n} \cdot \log(1 + n/\mathsf{T}_{\mathsf{mix}}) \wedge 1,$$

as desired. □

7.3.1 Proof of Lemma 7

We will show that $\delta_i(\widehat{M}^{\tau}_{\textsc{WingIt}}; X^{-i}) \le \frac{4\tau}{n}$ for a fixed $i \in [n]$ and any $X^{-i} \in \mathcal{X}^{n-1}$, which directly implies the desired result. Consider the sequences $X^{(i)}(x)$ and $X^{(i)}(x')$. Recall from Eq. (16) that we defined $\widehat{M}^{\tau}_{\textsc{WingIt}}(X^n) = \frac{1}{n} \sum_{i'=1}^n \widehat{M}^{(i')}_{\tau}(X^n)$, where $\widehat{M}^{(i')}_{\tau}(X^n) = \mathbb{I}\left\{ X_{i'} \notin \boldsymbol{X}_{\mathcal{I}_{i'}} \right\}$. From this, applying the triangle inequality yields

$$\delta_i(\widehat{M}^{\tau}_{\textsc{WingIt}}; X^{-i}) \le \max_{x, x' \in \mathcal{X}} \frac{\sum_{i'=1}^n \mathbb{I}\left\{ \widehat{M}^{(i')}_{\tau}(X^{(i)}(x)) \ne \widehat{M}^{(i')}_{\tau}(X^{(i)}(x')) \right\}}{n}.$$

Thus, it suffices to show that for any $x, x' \in \mathcal{X}$, the total number of indices $i' \in [n]$ for which $\widehat{M}_\tau^{(i')}$ changes when we switch X_i from $x \to x'$, i.e. the quantity

$$V(x, x') := \sum_{i'=1}^{n} \mathbb{I}\left\{\widehat{M}_\tau^{(i')}(X^{(i)}(x)) \neq \widehat{M}_\tau^{(i')}(X^{(i)}(x'))\right\},$$

is less than or equal to 4τ.

Let $Z_{i'} = \mathbb{I}\left\{\widehat{M}_\tau^{(i')}(X^{(i)}(x)) \neq \widehat{M}_\tau^{(i')}(X^{(i)}(x'))\right\}$. On the one hand, if $X_{i'} \notin \{x, x'\}$, then $Z_{i'} = 0$ because changing X_i from x to x' has no impact on the indicator $\mathbb{I}\left\{X_{i'} \notin \boldsymbol{X}_{\mathcal{I}_{i'}}\right\}$. On the other hand, if there are at least $2\tau + 1$ occurrences of x in the sequence X^n, and if $X_{i'} = x$, then $\mathbb{I}\left\{X_{i'} \notin \boldsymbol{X}_{\mathcal{I}_{i'}}\right\} = 0$ for any i' because $\mathcal{D}_{i'}$ has at most $2\tau - 1$ elements and some x remains in $\boldsymbol{X}_{\mathcal{I}_{i'}}$. So, $Z_{i'} = 0$. An analogous argument holds if there are at least $2\tau + 1$ occurrences of x' in the sequence X^n.

Thus, the number of indices i' for which $Z_{i'} = 1$ can be at most $2\tau + 2\tau = 4\tau$. This proves the bound $V(x, x') \leq 4\tau$ for any pair (x, x'), as claimed. □

Remark 8 *The upper bound in Lemma 7 is tight up to a factor* 4*. To see this, fix* $x \neq x' \in \mathcal{X}$*, the index* $i = 2\tau + 1$ *and form the sequence* $X^n = (X_1, \ldots, X_n)$ *by setting*

$$X_{i'} = \begin{cases} x & \text{if } i' \in \{1, \ldots, \tau\} \\ x' & \text{otherwise.} \end{cases}$$

Then, for all the indices $i' \in \{1, \ldots, \tau\}$ *we have* $\widehat{M}_\tau^{(i')}(X^{(i)}(x)) = 0$ *but* $\widehat{M}_\tau^{(i')}(X^{(i)}(x')) = 1$*. Moreover, for all other indices* $i' > \tau$ *we have* $\widehat{M}_\tau^{(i')}(X^{(i)}(x)) = \widehat{M}_\tau^{(i')}(X^{(i)}(x')) = 0$*. This directly implies that*

$$\delta_i(\widehat{M}_{\textsc{WingIt}}(\tau); X^{-i}) = \left|\frac{\tau}{n} - 0\right| = \frac{\tau}{n}.$$

7.4 Proof of Theorem 3

The structure of this proof parallels the proof of Theorem 1; the reader is advised to read that proof first. It is also useful to recall the notation for index sets that was defined in Section 7.1.

Owing to Eq (32), it suffices to establish Proposition 5(b). As in the case before, our argument will apply to bound the MSE of the estimator $\widehat{M}_{\textsc{WingIt},\leq\zeta}(\tau; \ell)$ for any $\ell = 0, \ldots, 2\tau - 1$, so we concentrate on establishing that for an absolute positive constant C and $\tau \geq \mathsf{t}_{\mathsf{mix}}\left((\mathsf{T}_{\mathsf{mix}}/n) \wedge 1/4\right)$:

$$\mathbb{E}\left(\widehat{M}_{\textsc{WingIt},\leq\zeta}(\tau; 0) - M_\pi\right)^2 \leq C \cdot \frac{(\zeta + 1)\tau}{n} \wedge 1. \tag{45}$$

We next proceed via a series of steps that resembles the proof of Theorem 1. Recall the notation $N_x(\boldsymbol{X}_P)$ (and $N_x(X_P)$) that we defined in Section 5, denoting the number of occurrences of x in the subset $\boldsymbol{X}_P$ (and subsequence X_P). Viewing X^n as fixed for the moment

and writing out our estimator $\widehat{M}_{\text{WINGIT},\leq\zeta}(\tau;0) := \frac{1}{n_0}\sum_{j=1}^{n_0}\widehat{M}^{(2\tau j)}_{\tau,\leq\zeta}$, a parallel argument to Eq. (36) yields that $\frac{1}{2}|\widehat{M}_{\text{WINGIT},\leq\zeta}(\tau;0) - M_{\pi,\leq\zeta}(X^n)|^2$ is equal to

$$\frac{1}{2}\cdot\left|\frac{1}{n_0}\sum_{j=1}^{n_0}\mathbb{I}\left\{N_{X_{2\tau j}}(\boldsymbol{X}_{\mathcal{H}_j})\leq\zeta\right\} - \mathop{\mathbb{E}}_{\substack{Y\sim\pi\\ Y\perp\!\!\!\perp X^n}}\mathbb{I}\left\{N_Y(\boldsymbol{X}_{[n]})\leq\zeta\right\}\right|^2$$
$$\leq \underbrace{\left|\frac{1}{n_0}\sum_{j=1}^{n_0}\left(\mathop{\mathbb{E}}_{\substack{Y\sim\pi\\ Y\perp\!\!\!\perp X^n}}\mathbb{I}\left\{N_Y(\boldsymbol{X}_{\mathcal{H}_j})\leq\zeta\right\} - \mathop{\mathbb{E}}_{\substack{Y\sim\pi\\ Y\perp\!\!\!\perp X^n}}\mathbb{I}\left\{N_Y(\boldsymbol{X}_{[n]})\leq\zeta\right\}\right)\right|^2}_{T_1'}$$
$$+\underbrace{\left|\frac{1}{n_0}\sum_{j=1}^{n_0}\left(\mathbb{I}\left\{N_{X_{2\tau j}}(\boldsymbol{X}_{\mathcal{H}_j})\leq\zeta\right\} - \mathop{\mathbb{E}}_{\substack{Y\sim\pi\\ Y\perp\!\!\!\perp X^n}}\mathbb{I}\left\{N_Y(\boldsymbol{X}_{\mathcal{H}_j})\leq\zeta\right\}\right)\right|^2}_{T_2'}.$$

As in the proof of Theorem 1, we upper bound $\mathbb{E}[T_1']$ and $\mathbb{E}[T_2']$.

7.4.1 BOUNDING $\mathbb{E}[T_1']$

As in the proof of Theorem 1, T_1' resembles a conditional squared bias term. For each $j\in[n_0]$, we now define the random variable $P_j' := \mathbb{I}\left\{N_Y(\boldsymbol{X}_{\mathcal{H}_j})\leq\zeta\right\} - \mathbb{I}\left\{N_Y(\boldsymbol{X}_{[n]})\leq\zeta\right\}$. It is easy to see that $T_1' = \frac{1}{n_0^2}\left(\mathbb{E}_Y\sum_{j=1}^{n_0}P_j'\right)^2$. We will bound the term $\sum_{j=1}^{n_0}P_j'$ pointwise. Applying Lemma 12, we have $P_j' \leq \mathbb{I}\left\{Y\in\boldsymbol{X}_{\mathcal{B}_j}\right\}\cdot\mathbb{I}\left\{N_Y(\boldsymbol{X}_{\mathcal{H}_j})\leq\zeta\right\}$. As also argued in Section 7.2.1, 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 $\mathbb{I}\left\{Y\in\boldsymbol{X}_{\mathcal{B}_j}\right\}\cdot\mathbb{I}\left\{N_Y(\boldsymbol{X}_{\mathcal{H}_j})\leq\zeta\right\} = 1$. This means that Y occurs at least once in $\mathcal{B}_j$, but its number of occurrences outside of $\mathcal{B}_j$ is at most ζ. Then, we must have

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

Putting these together yields $\sum_{j=1}^{n_0}P_j'\leq\zeta+1$ pointwise. Ultimately, this yields

$$T_1' = \frac{1}{n_0^2}\left(\mathop{\mathbb{E}}_{Y}\sum_{j=1}^{n_0}P_j'\right)^2 \leq \left(\frac{\zeta+1}{n_0}\right)^2,$$

and so $\mathbb{E}[T_1]\lesssim\left(\frac{(\zeta+1)\tau}{n}\right)^2\wedge 1$.

7.4.2 Bounding $\mathbb{E}[T_2']$

As in the proof of Theorem 1, T_2' resembles a conditional variance term. We now define, for all $j \in [n_0]$, the random variables

$$Z_j' := \mathbb{I}\left\{N_{X_{2j\tau}}(\boldsymbol{X}_{\mathcal{H}_j}) \leq \zeta\right\} - \mathop{\mathbb{E}}_{\substack{Y \sim \pi \\ Y \perp\!\!\!\perp X^n}} \mathbb{I}\left\{N_Y(\boldsymbol{X}_{\mathcal{H}_j}) \leq \zeta\right\}.$$

Then, we have

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

where the inequality follows since $Z_j' \in [-1, 1]$ for all $j \in [n_0]$. Therefore, it suffices to bound the cross terms when $j \neq k$. For each $j, k \in [n_0]$ with $j \neq k$, define the random variables

$$Q_{j,k}' = \mathbb{I}\left\{N_Y(\boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}) \leq \zeta\right\} - \mathbb{I}\left\{N_Y(\boldsymbol{X}_{\mathcal{H}_k}) \leq \zeta\right\}. \tag{46}$$

The following lemma, which is analogous to Lemma 6, relates the expectation of the cross terms to the expectations of these random variables.

Lemma 9 *Suppose $\tau \geq \mathsf{t}_{\mathsf{mix}}(\epsilon)$. Then, for each $j \neq k$, we have*

$$\mathbb{E}[Z_j' Z_k'] \leq \frac{5}{2}\,\mathbb{E}[Q_{j,k}'] + \frac{5}{2}\,\mathbb{E}[Q_{k,j}'] + 16\epsilon,$$

where the random variables $\{Q_{j,k}'\}$ are defined as in Eq. (46)*.*

We take Lemma 9 as given for the moment and prove it in Section 7.4.3. We now use it to bound $\mathbb{E}[T_2']$. Applying Lemma 12, we have $Q_{j,k}' \leq \mathbb{I}\left\{Y \in \boldsymbol{X}_{\mathcal{H}_k \setminus \mathcal{H}_j}\right\} \cdot \mathbb{I}\left\{N_Y(\boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}) \leq \zeta\right\}$. Now, consider some fixed $k \in [n_0]$. As described in Section 7.2.2, since the sets $\{\mathcal{B}_j\}_{j=1}^{n_0}$ are non-overlapping, we have $\mathcal{H}_k \setminus \mathcal{H}_j = \mathcal{B}_j$, and

$$\mathcal{H}_j \cap \mathcal{H}_k \supset \bigsqcup_{j' \in [n_0] \setminus \{j,k\}} \mathcal{B}_{j'}.$$

If for some $j \neq k$, we have

$$\mathbb{I}\left\{Y \in \boldsymbol{X}_{\mathcal{H}_k \setminus \mathcal{H}_j}\right\} \cdot \mathbb{I}\left\{N_Y(\boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}) \leq \zeta\right\} = 1,$$

it means that Y occurs at least once in the block $\mathcal{B}_j$, but at most ζ times in the set $\mathcal{H}_j \cap \mathcal{H}_k$. This implies that

$$\begin{aligned}\sum_{j' \in [n_0] \setminus \{j,k\}} \mathbb{I}\left\{Y \in \boldsymbol{X}_{\mathcal{H}_k \setminus \mathcal{H}_{j'}}\right\} \cdot \mathbb{I}\left\{N_Y(\boldsymbol{X}_{\mathcal{H}_{j'} \cap \mathcal{H}_k}) \leq \zeta\right\} &\leq \sum_{j' \in [n_0] \setminus \{j,k\}} \mathbb{I}\left\{Y \in \boldsymbol{X}_{\mathcal{B}_{j'}}\right\} \\ &\leq N_Y(\boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}) \leq \zeta.\end{aligned}$$

Consequently, we have

$$\sum_{j\in[n_0]\setminus k} Q'_{j,k} \le \zeta + 1.$$

Applying Lemma 9 and using the linearity of expectation then yields

$$\sum_{j=1}^{n_0}\sum_{\substack{k=1\\k\neq j}}^{n_0} \mathbb{E}[Z'_j Z'_k] \le 5(\zeta+1)n_0 + 16n_0^2\epsilon.$$

Consequently, we have $\mathbb{E}[T'_2] \lesssim \frac{\zeta+1}{n_0} + \epsilon$. Substituting $\epsilon = \frac{\mathsf{T}_{\mathsf{mix}}}{n}$ and noting that $\tau \ge \mathsf{T}_{\mathsf{mix}}$ by assumption, we obtain $\mathbb{E}[T'_2] \le C \cdot \frac{\tau(\zeta+1)}{n} \wedge 1$.

Combining our bounds on $\mathbb{E}[T'_1]$ and $\mathbb{E}[T'_2]$ completes the proof of Theorem 3. It remains to prove Lemma 9.

7.4.3 Proof of Lemma 9

The structure of this proof closely resembles the proof of Lemma 6. Define

$$\overline{Q}_{j,k} := \mathop{\mathbb{E}}_{Y}[Q'_{j,k}] = \mathop{\mathbb{E}}_{Y}\left[\mathbb{I}\left\{N_Y(\boldsymbol{X}_{\mathcal{H}_j\cap\mathcal{H}_k}) \le \zeta\right\} - \mathbb{I}\left\{N_Y(\boldsymbol{X}_{\mathcal{H}_k}) \le \zeta\right\}\right] \tag{47}$$

for convenience. Note that by Lemma 12, we have

$$\overline{Q}_{j,k} \le \mathop{\mathbb{E}}_{Y}\left[\mathbb{I}\left\{Y \in \boldsymbol{X}_{\mathcal{H}_k\setminus\mathcal{H}_j}\right\} \cdot \mathbb{I}\left\{N_Y(\boldsymbol{X}_{\mathcal{H}_j\cap\mathcal{H}_k}) \le \zeta\right\}\right] \le 1,$$

and moreover $Q'_{j,k} \ge 0$ pointwise so $\overline{Q}_{j,k} \ge 0$. Therefore, $\overline{Q}_{j,k} \in [0,1]$. We then have the decomposition

$$\begin{aligned}
&Z'_j Z'_k \\
&= \left(\mathbb{I}\left\{N_{X_{2\tau j}}(\boldsymbol{X}_{\mathcal{H}_j}) \le \zeta\right\} - \mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{N_Y(\boldsymbol{X}_{\mathcal{H}_j\cap\mathcal{H}_k}) \le \zeta\right\}] + \overline{Q}_{k,j}\right) \cdot \\
&\qquad\qquad \left(\mathbb{I}\left\{N_{X_{2\tau k}}(\boldsymbol{X}_{\mathcal{H}_k}) \le \zeta\right\} - \mathop{\mathbb{E}}_{Y'}[\mathbb{I}\left\{N_{Y'}(\boldsymbol{X}_{\mathcal{H}_j\cap\mathcal{H}_k}) \le \zeta\right\}] + \overline{Q}_{j,k}\right) \\
&\le \underbrace{\left(\mathbb{I}\left\{N_{X_{2\tau j}}(\boldsymbol{X}_{\mathcal{H}_j}) \le \zeta\right\} - \mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{N_Y(\boldsymbol{X}_{\mathcal{H}_j\cap\mathcal{H}_k}) \le \zeta\right\}]\right) \cdot \left(\mathbb{I}\left\{N_{X_{2\tau k}}(\boldsymbol{X}_{\mathcal{H}_k}) \le \zeta\right\} - \mathop{\mathbb{E}}_{Y'}[\mathbb{I}\left\{N_{Y'}(\boldsymbol{X}_{\mathcal{H}_j\cap\mathcal{H}_k}) \le \zeta\right\}]\right)}_{U'_{j,k}} \\
&\qquad\qquad + \overline{Q}_{j,k} + \overline{Q}_{k,j} + \overline{Q}_{j,k}\cdot\overline{Q}_{k,j} \\
&\overset{\text{(i)}}{\le} U'_{j,k} + \frac{3}{2}(\overline{Q}_{j,k} + \overline{Q}_{k,j}).
\end{aligned}$$

Here step (i) follows due to the following algebraic inequalities: Since each $\overline{Q}_{j,k} \in [0,1]$, we have $\overline{Q}_{j,k}\cdot\overline{Q}_{k,j} \le \sqrt{\overline{Q}_{j,k}\cdot\overline{Q}_{k,j}} \le \frac{1}{2}(\overline{Q}_{j,k} + \overline{Q}_{k,j})$.

It remains to establish that $\mathbb{E}[U'_{j,k}] \leq \mathbb{E}_{X^n}[\overline{Q}_{j,k}] + \mathbb{E}_{X^n}[\overline{Q}_{k,j}] + 16\epsilon$. We have the further decomposition

$$\begin{aligned}
&\mathbb{E}[U'_{j,k}] \\
&= \underbrace{\mathbb{E}[\mathbb{I}\left\{N_{X_{2\tau j}}(\boldsymbol{X}_{\mathcal{H}_j}) \leq \zeta\right\} \cdot \mathbb{I}\left\{N_{X_{2\tau k}}(\boldsymbol{X}_{\mathcal{H}_k}) \leq \zeta\right\}]}_{U'_1} - \underbrace{\mathop{\mathbb{E}}_{X^n}\left[\mathbb{I}\left\{N_{X_{2\tau j}}(\boldsymbol{X}_{\mathcal{H}_j}) \leq \zeta\right\} \cdot \mathop{\mathbb{E}}_{Y'}[\mathbb{I}\left\{N_{Y'}(\boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}) \leq \zeta\right\}]\right]}_{U'_2} \\
&- \underbrace{\mathop{\mathbb{E}}_{X^n}\left[\mathbb{I}\left\{N_{X_{2\tau k}}(\boldsymbol{X}_{\mathcal{H}_k}) \leq \zeta\right\} \cdot \mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{N_{Y}(\boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}) \leq \zeta\right\}]\right]}_{U'_3} \\
&+ \underbrace{\mathop{\mathbb{E}}_{X^n}\left[\mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{N_{Y}(\boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}) \leq \zeta\right\}] \cdot \mathop{\mathbb{E}}_{Y'}[\mathbb{I}\left\{N_{Y'}(\boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}) \leq \zeta\right\}]\right]}_{U'_4} \qquad (48)
\end{aligned}$$

We now bound each of the above terms in turn.

First, we bound U'_1 as

$$\begin{aligned}
U'_1 &\leq \mathbb{E}[\mathbb{I}\left\{N_{X_{2\tau j}}(\boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}) \leq \zeta\right\} \cdot \mathbb{I}\left\{N_{X_{2\tau k}}(\boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}) \leq \zeta\right\}] \\
&\overset{\text{(i)}}{\leq} \mathbb{E}[\mathbb{I}\left\{N_{Y'}(\boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}) \leq \zeta\right\} \cdot \mathbb{I}\left\{N_{Y}(\boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}) \leq \zeta\right\}] + 8\epsilon \\
&\overset{\text{(ii)}}{=} \mathop{\mathbb{E}}_{X^n}\left\{\mathop{\mathbb{E}}_{Y'}[\mathbb{I}\left\{N_{Y'}(\boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}) \leq \zeta\right\}] \cdot \mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{N_{Y}(\boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}) \leq \zeta\right\}]\right\} + 8\epsilon, \qquad (49)
\end{aligned}$$

where step (i) uses Lemma 14 (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

$$f'(X_{2j\tau}; \boldsymbol{X}_{\mathcal{H}_j}) := \mathbb{I}\left\{N_{X_{2\tau j}}(\boldsymbol{X}_{\mathcal{H}_j}) \leq \zeta\right\} \cdot \mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{N_{Y}(\boldsymbol{X}_{\mathcal{H}_k \cap \mathcal{H}_j}) \leq \zeta\right\}],$$

which is bounded in the range $[0,1]$. Since $\tau \geq \mathsf{t}_{\mathsf{mix}}(\epsilon)$, we may now apply Lemma 13 (for the choice $i = 2\tau j$) to obtain $|\,\mathbb{E}[f'(X_{2\tau j}; \boldsymbol{X}_{\mathcal{H}_j})] - \mathbb{E}[f'(Y'; \boldsymbol{X}_{\mathcal{H}_j})]| \leq 4\epsilon$. Thus,

$$\begin{aligned}
U'_2 &\geq \mathop{\mathbb{E}}_{X^n}\left[\mathop{\mathbb{E}}_{Y'}[\mathbb{I}\left\{N_{Y'}(\boldsymbol{X}_{\mathcal{H}_j}) \leq \zeta\right\}] \cdot \mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{N_{Y}(\boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}) \leq \zeta\right\}]\right] - 4\epsilon \\
&\geq \mathop{\mathbb{E}}_{X^n}\left[\mathop{\mathbb{E}}_{Y'}[\mathbb{I}\left\{N_{Y'}(\boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}) \leq \zeta\right\}] \cdot \mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{N_{Y}(\boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}) \leq \zeta\right\}]\right] - \mathop{\mathbb{E}}_{X^n}[\overline{Q}_{k,j}] - 4\epsilon. \qquad (50)
\end{aligned}$$

By an identical argument to the above, we have

$$U'_3 \geq \mathop{\mathbb{E}}_{X^n}\left[\mathop{\mathbb{E}}_{Y'}[\mathbb{I}\left\{N_{Y'}(\boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}) \leq \zeta\right\}] \cdot \mathop{\mathbb{E}}_{Y}[\mathbb{I}\left\{N_{Y}(\boldsymbol{X}_{\mathcal{H}_j \cap \mathcal{H}_k}) \leq \zeta\right\}]\right] - \mathop{\mathbb{E}}_{X^n}[\overline{Q}_{j,k}] - 4\epsilon. \qquad (51)$$

Putting Eqs. (49), (50) and (51) together with the definition of U'_4 and performing the requisite cancellations, we have

$$\mathbb{E}[U'_{j,k}] = U'_1 - U'_2 - U'_3 + U'_4 \leq \mathop{\mathbb{E}}_{X^n}[\overline{Q}_{j,k}] + \mathop{\mathbb{E}}_{X^n}[\overline{Q}_{k,j}] + 16\epsilon.$$

This completes the proof of the lemma and so the proof of Theorem 3. $\square$

8. Discussion

We presented the WingIt estimator for estimating the stationary mass missing from a Markovian sequence. While the vanilla Good–Turing estimator can suffer constant bias in the Markovian setting, our estimator achieves (near) minimax optimal mean-squared error over mixing Markov chains. It can also be computed with a linear-time algorithm, and performs favorably in our experiments, even in language text applications in which the Markovian assumption is clearly violated. We also presented a variant of WingIt for estimating the small-count probability in a Markov sequence and established mean squared error bounds for this task.

Our work leaves open several important and intriguing questions aside from the conjectured improvement of Theorem 2. First, while Theorem 1 provides a complete picture—up to a logarithmic factor—for stationary missing mass estimation from the point of view of MSE, it would be interesting to complement this result with a concentration inequality. Such a concentration result could, for instance, be used to provide a provable guarantee on the validation procedure that we outlined in Section 6. Second, we reiterate that our estimator is only optimal up to a logarithmic factor in $n/\mathsf{T}_{\mathsf{mix}}$, and removing this factor to match the minimax lower bound—possibly by designing an alternative estimator—is an interesting open problem.

Third, we believe that the Markov property may not be central to our main results, and that Theorem 1 could be extended to more general α-mixing sequences (Rosenblatt, 1956). This extension would capture, for instance, other classes of interesting temporal processes such as some hidden Markov models. Fourth, a related point is that the assumption (5) of geometric ergodicity itself is central to the design and analysis of our estimator; designing estimators that do not require ergodicity—perhaps just irreducibility (Fried, 2023)—would be of great interest and likely require new ideas.

Finally, it would be interesting to estimate other functionals of the Markov chain other than the stationary missing mass and solve related estimation problems such as competitive distribution estimation of the stationary measure. Our extensions to estimating the mass of elements occurring at most ζ times in Section 5 might be a useful starting point as in the i.i.d. case (Drukh and Mansour, 2005; Acharya et al., 2013), but several questions remain, such as obtaining a bound on the error of estimating all such quantities uniformly over $\zeta \in \{0, 1, \ldots, n\}$.

Acknowledgments

This work was supported in part by National Science Foundation grants CCF-2107455, DMS-2210734, CCF-2239151 and IIS-2212182, and by research awards/gifts from Adobe, Amazon, Google and MathWorks. AP thanks Wenlong Mou for helpful discussions. We are also thankful to the anonymous reviewers, whose comments improved the scope and presentation of the manuscript.

References

Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Optimal probability estimation with applications to prediction and classification. In *Conference*

on Learning Theory, pages 764–796, 2013.

Jayadev Acharya, Yelun Bao, Yuheng Kang, and Ziteng Sun. Improved bounds for minimax risk of estimating missing mass. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 326–330. IEEE, 2018.

Fadhel Ayed, Marco Battiston, Federico Camerlenghi, and Stefano Favaro. On consistent and rate optimal estimation of the missing mass. In *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*, volume 57, pages 1476–1494. Institut Henri Poincaré, 2021.

Anna Ben-Hamou, Stéphane Boucheron, and Mesrob I. Ohannessian. Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli*, 23(1):249 – 287, 2017. doi: 10.3150/15-BEJ743. URL `https://doi.org/10.3150/15-BEJ743`.

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

Daniel Berend and Aryeh 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`.

Prafulla Chandra, Aditya Pradeep, and Andrew Thangaraj. Improved tail bounds for missing mass and confidence intervals for Good–Turing estimator. In *2019 National Conference on Communications (NCC)*, pages 1–6. IEEE, 2019.

Prafulla Chandra, Andrew Thangaraj, and Nived Rajaraman. How good is Good–Turing for Markov samples? *arXiv preprint arXiv:2102.01938*, 2021.

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

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

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

Robert K Colwell, Anne Chao, Nicholas J Gotelli, Shang-Yi Lin, Chang Xuan Mao, Robin L Chazdon, and John T Longino. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, 5(1):3–21, 2012.

Charles Dickens. *A Tale of Two Cities*. Project Gutenberg, Urbana, IL, USA, 1994. URL `gutenberg.org/ebooks/98`.

Joseph L Doob. Regularity properties of certain families of chance variables. *Transactions of the American Mathematical Society*, 47(3):455–486, 1940.

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

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

Stefano Favaro, Bernardo Nipoti, and Yee Whye Teh. Rediscovery of Good–Turing estimators via Bayesian nonparametrics. *Biometrics*, 72(1):136–145, 2016.

Ronald A Fisher, A Steven Corbet, and Carrington 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.

Sela Fried. On the α-lazy version of Markov chains in estimation and testing problems. *Statistical Inference for Stochastic Processes*, 26(2):413–435, 2023.

William Gale and Kenneth Church. What is wrong with adding one? In *Corpus-based research into language*, pages 189–198. Brill, 1994.

William A Gale and Geoffrey Sampson. Good–Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2(3):217–237, 1995.

William A Gale, Kenneth W Church, and David Yarowsky. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415–439, 1992.

Alberto Gandolfi and Chelluri CA Sastri. Nonparametric estimations about species not observed in a random sample. *Milan Journal of Mathematics*, 72:81–105, 2004.

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

Michael Grabchak and Zhiyi Zhang. Asymptotic properties of Turing's formula in relative error. *Machine Learning*, 106:1771–1785, 2017.

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

Yi Hao, Alon Orlitsky, and Venkatadheeraj Pichapati. On learning Markov chains. *Advances in Neural Information Processing Systems*, 31, 2018.

Daniel Hsu, Aryeh Kontorovich, David A. Levin, Yuval Peres, Csaba Szepesvári, and Geoffrey Wolfer. Mixing time estimation in reversible Markov chains from a single sample path. *The Annals of Applied Probability*, 29(4):2439 – 2480, 2019. doi: 10.1214/18-AAP1457. URL `https://doi.org/10.1214/18-AAP1457`.

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

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

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

Pierre-Simon Laplace. *Essai philosophique sur les probabilités.* 1814.

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

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

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

David McAllester and Robert E Schapire. Learning theory and language modeling. In *Seventeenth International Joint Conference on Artificial Intelligence*, 2001.

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

Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141 (1):148–188, 1989.

Elchanan Mossel and Mesrob I Ohannessian. On the impossibility of learning the missing mass. *Entropy*, 21(1):28, 2019.

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

James Norris, Yuval Peres, and Alex Zhai. Surprise probabilities in Markov chains. *Combinatorics, Probability and Computing*, 26(4):603–627, 2017.

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

Roberto Oliveira. Mixing and hitting times for finite Markov chains. *Electronic Journal of Probability*, 17(none):1 – 12, 2012. doi: 10.1214/EJP.v17-2274. URL `https://doi.org/10.1214/EJP.v17-2274`.

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

Alon Orlitsky, Narayana P Santhanam, and Junan Zhang. Always Good–Turing: Asymptotically optimal probability estimation. *Science*, 302(5644):427–431, 2003.

Amichai Painsky. Convergence guarantees for the Good–Turing estimator. *Journal of Machine Learning Research*, 23(279):1–37, 2022.

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

Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20(none):1 – 32, 2015. doi: 10.1214/EJP.v20-4039. URL `https://doi.org/10.1214/EJP.v20-4039`.

Yuval Peres and Perla Sousi. Mixing times are hitting times of large sets. *Journal of Theoretical Probability*, 28(2):488–519, 2015.

Nikhilesh Rajaraman, Andrew Thangaraj, and Ananda Theertha Suresh. Minimax risk for missing mass estimation. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 3025–3029. IEEE, 2017.

Murray Rosenblatt. A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences*, 42(1):43–47, 1956.

Tsung-Jen Shen, Anne Chao, and Chih-Feng Lin. Predicting the number of new species in further taxonomic sampling. *Ecology*, 84(3):798–804, 2003.

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

Fei Song and W Bruce Croft. A general language model for information retrieval. In *Proceedings of the Eighth International Conference on Information and Knowledge Management*, pages 316–321, 1999.

Andrew Thangaraj, Ashwin Pananjady, and Vidya Muthukumar. Missing Mass of Markov Chains, 2024. URL `https://github.com/andrewthan/Missing-Mass`.

Vincent Q Vu, Bin Yu, and Robert E Kass. Coverage-adjusted entropy estimation. *Statistics in Medicine*, 26(21):4039–4060, 2007.

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

Appendix A. Technical lemmas

In this section, we collect technical lemmas that were stated and used in the main paper. We first collect lemmas that were used to formalize basic calculations for the Good–Turing estimator, and next lemmas that were used in the proofs of the main results (Theorems 1 and 2).

A.1 Elementary lemmas

Our first lemma shows a tight characterization of the mixing time $\mathsf{T}_{\mathsf{mix}} = \mathsf{t}_{\mathsf{mix}}(1/4)$ for the class of sticky Markov chains, defined in Eq. (10).

Lemma 10 *Suppose $|\mathcal{X}| \geq 2$ and $p \in (0, 1/2]$. For any sticky Markov chain as defined in Eq. (10), we have*

$$\frac{1}{2p} \leq \mathsf{T}_{\mathsf{mix}} \leq \frac{2}{p}. \tag{52}$$

Proof We proceed by exactly calculating the total variation distance $\max_{x\in\mathcal{X}} \|e_x^\top \boldsymbol{P}^t - \pi^\top\|_{\mathsf{TV}}$. For any starting state $x \in \mathcal{X}$ we would reach the stationary distribution π in a number of steps that is a geometric random variable, i.e. $\tau = \mathrm{Geom}(p)$. This means that $\Pr\{\tau \geq t\} = (1-p)^t$, directly implying that

$$\begin{aligned}\max_{x\in\mathcal{X}} \|e_x^\top \boldsymbol{P}^t - \pi^\top\|_{\mathsf{TV}} &= \max_{x\in\mathcal{X}} \frac{1}{2}\|(1-p)^t \cdot (e_x - \pi)\|_1 \\ &= \frac{(1-p)^t}{2} \cdot \max_{x\in\mathcal{X}} \|e_x - \pi\|_1.\end{aligned}$$

Since $(1-p)^t \cdot \max_{x\in\mathcal{X}} \|e_x - \pi\|_{\mathsf{TV}}$ is monotonically decreasing in t, we have

$$\mathsf{T}_{\mathsf{mix}} = \mathsf{t}_{\mathsf{mix}}(1/4) = \frac{\log(2 \cdot \max_{x\in\mathcal{X}} \|e_x - \pi\|_1)}{\log(1/(1-p))}.$$

But

$$\|e_x - \pi\|_1 = (1-\pi_x) + \sum_{y\in\mathcal{X}\setminus\{x\}} \pi_y = 2 - 2\pi_x,$$

and if $|\mathcal{X}| \geq 2$, then we have $1 \leq \max_{x\in\mathcal{X}} \|e_x - \pi\|_1 \leq 2$. Furthermore, if $p \in (0, 1/2]$, then $p \leq \log(1/(1-p)) \leq p \log 4$. Putting together the pieces, we obtain the sandwich bound

$$\frac{1}{2p} \leq \mathsf{T}_{\mathsf{mix}} \leq \frac{2}{p},$$

as desired. ■

For any set $P \subseteq [n]$, recall that $\boldsymbol{X}_P = \{X_k\}_{k\in P}$ denotes the set of random variables in X^n restricted to the index set P. The following lemma is a deterministic statement regarding indicator random variables.

Lemma 11 *Consider the sequence X^n and any random variable Y defined on the space $\mathcal{X}$. Let $P \subseteq Q \subseteq [n]$ denote two index sets, and let $R := Q \setminus P$. We have*

$$\mathbb{I}\{Y \notin \boldsymbol{X}_P\} - \mathbb{I}\{Y \notin \boldsymbol{X}_Q\} = \mathbb{I}\{Y \in \boldsymbol{X}_R\} \cdot \mathbb{I}\{Y \notin \boldsymbol{X}_P\}.$$

Proof Since $P \subseteq Q$, we have that $\boldsymbol{X}_P \subseteq \boldsymbol{X}_Q$. Consequently, if $Y \notin \boldsymbol{X}_Q$, then Y cannot be included in the subset $\boldsymbol{X}_P$. Therefore, $\mathbb{I}\{Y \notin \boldsymbol{X}_P\} - \mathbb{I}\{Y \notin \boldsymbol{X}_Q\} = 1$ if and only if $Y \notin \boldsymbol{X}_P$ *and* $Y \in \boldsymbol{X}_Q$. Since $R = Q \setminus P$, this is equivalent to saying that $Y \in \boldsymbol{X}_R$ and $Y \notin \boldsymbol{X}_P$. Thus, we have shown that

$$\mathbb{I}\{Y \notin \boldsymbol{X}_P\} - \mathbb{I}\{Y \notin \boldsymbol{X}_Q\} = \mathbb{I}\{Y \in \boldsymbol{X}_R\} \cdot \mathbb{I}\{Y \notin \boldsymbol{X}_P\},$$

as claimed. ■

We also define an extension of Lemma 11 to the slightly more complicated indicator random variables involving the count of an element in index sets.

Lemma 12 *Consider the sequence X^n and any random variable Y defined on the space $\mathcal{X}$. Let $P \subseteq Q \subseteq [n]$ denote two index sets, and let $R := Q \setminus P$. Then, for any $\zeta \geq 0$, we have*

$$\mathbb{I}\{N_Y(\boldsymbol{X}_P) \leq \zeta\} - \mathbb{I}\{N_Y(\boldsymbol{X}_Q) \leq \zeta\} \leq \mathbb{I}\{Y \in \boldsymbol{X}_R\} \cdot \mathbb{I}\{N_Y(\boldsymbol{X}_P) \leq \zeta\}.$$

Proof Since $P \subseteq Q$, we have that $\boldsymbol{X}_P \subseteq \boldsymbol{X}_Q$. Then, if Y occurs less than ζ times in $\boldsymbol{X}_Q$, i.e. $N_Y(\boldsymbol{X}_Q) \leq \zeta$, then Y must occur less than ζ times in the subset $\boldsymbol{X}_P$, i.e. $N_Y(\boldsymbol{X}_P) \leq \zeta$. Consequently, $\mathbb{I}\{N_Y(\boldsymbol{X}_P) \leq \zeta\} - \mathbb{I}\{N_Y(\boldsymbol{X}_Q) \leq \zeta\} = 1$ if and only if the number of occurrences of Y in $\boldsymbol{X}_P$ is less than or equal to ζ, i.e. $N_Y(\boldsymbol{X}_P) \leq \zeta$; but the number of occurrences of Y in $\boldsymbol{X}_Q$ is greater than ζ, i.e. $N_Y(\boldsymbol{X}_Q) > \zeta$. Further, we have $N_Y(\boldsymbol{X}_P) \leq \zeta$ and $N_Y(\boldsymbol{X}_Q) > \zeta$ *only if* $N_Y(\boldsymbol{X}_P) \leq \zeta$ *and* there exists at least one occurrence of Y in $\boldsymbol{X}_R$, i.e. $N_Y(\boldsymbol{X}_R) \geq 1$. This gives us

$$\mathbb{I}\{N_Y(\boldsymbol{X}_P) \leq \zeta\} - \mathbb{I}\{N_Y(\boldsymbol{X}_Q) \leq \zeta\} \leq \mathbb{I}\{Y \in \boldsymbol{X}_R\} \cdot \mathbb{I}\{N_Y(\boldsymbol{X}_P) \leq \zeta\}.$$

■

A.2 Lemmas on surrogate processes

We next present two important consequences of mixing. In all the lemmas below, let $(X_1, \ldots, X_n)$ denote a Markov chain with unique stationary distribution π and $X_1 \sim \pi$. Let $\mathsf{t}_{\mathsf{mix}}(\epsilon)$ denote its mixing time in the sense of Eq. (5), with $\epsilon \in (0, 1/2]$.

Lemma 13 *Fix a positive scalar $\epsilon \leq 1/2$, and let $\tau \geq \mathsf{t}_{\mathsf{mix}}(\epsilon)$ be an integer. For each $i \in [n]$, define the stochastic processes*

$$Z_i = (X_1, X_2, \ldots, X_{i-\tau}, X_i, X_{i+\tau}, X_{i+\tau+1}, \ldots, X_n), \tag{53}$$
$$Z_i' = (X_1, X_2, \ldots, X_{i-\tau}, X_i', X_{i+\tau}, X_{i+\tau+1}, \ldots, X_n), \tag{54}$$

where $X_i' \sim \pi$ is drawn independently of everything else. Then $\mathsf{d}_{\mathsf{TV}}(Z_i, Z_i') \leq 4\epsilon$.

Consequently, for any function $f : \mathcal{X}^{n-(2\tau-2)} \to [0, 1]$, we have

$$|\,\mathbb{E}[f(Z_i) - f(Z_i')]| \leq 4\epsilon.$$

Proof Let $A_i = (X_{i-\tau}, X_i, X_{i+\tau})$ and $A_i' = (X_{i-\tau}, X_i', X_{i+\tau})$. By the Markov property, we have $\mathsf{d}_{\mathsf{TV}}(Z_i, Z_i') = \mathsf{d}_{\mathsf{TV}}(A_i, A_i')$. We now define the notation $p^{(t)}(y|x) = \Pr\{X_{i+t} = y | X_i = x\}$ for any $t \geq 1$ and any $x, y \in \mathcal{X}$. Owing to the time invariant nature of the process, the distributions of these triples can be written explicitly as

$$\Pr\{A_i = (x, y, z)\} = \Pr\{(X_{i-\tau}, X_i, X_{i+\tau}) = (x, y, z)\} = \pi_x \cdot p^{(\tau)}(y|x) \cdot p^{(\tau)}(z|y) \text{ and}$$
$$\Pr\{A_i' = (x, y, z)\} = \Pr\{(X_{i-\tau}, X_i', X_{i+\tau}) = (x, y, z)\} = \pi_x \cdot \pi_y \cdot p^{(2\tau)}(z|x).$$

For each tuple of indices $(x, y, z) \in \mathcal{X} \times \mathcal{X} \times \mathcal{X}$ and any choice $t \geq 1$, define

$$\delta_{x,y}^{(t)} := \pi_y - p^{(t)}(y|x) \text{ and}$$
$$\overline{\delta}_{x,y,z}^{(t)} := p^{(2t)}(z|x) - p^{(t)}(z|y).$$

Below, we use the shorthand $\delta_{x,y} := \delta^{(\tau)}_{x,y}$ and $\overline{\delta}_{x,y,z} := \delta^{(\tau)}_{x,y,z}$ for convenience. Owing to our total variation mixing assumption (5) and the choice $\tau \geq \mathsf{t}_{\mathsf{mix}}(\epsilon)$, we have that the ℓ_1-norm of each of these errors is bounded for any $t \geq \tau$ as:

$$\max_{x \in \mathcal{X}} \sum_{y \in \mathcal{X}} |\delta^{(t)}_{x,y}| \leq 2\epsilon \text{ and} \tag{55a}$$

$$\max_{x,y \in \mathcal{X}} \sum_{z \in \mathcal{X}} |\overline{\delta}^{(t)}_{x,y,z}| \leq 2\epsilon + 2\epsilon^2. \tag{55b}$$

With this shorthand notation, we may define

$$\begin{aligned}
&\Pr\{A_i = (x,y,z)\} = \pi_x \cdot p^{(\tau)}(y|x) \cdot p^{(\tau)}(z|y) \text{ and} \\
&\Pr\{A'_i = (x,y,z)\} = \pi_x \cdot (p^{(\tau)}(y|x) + \delta_{x,y}) \cdot (p^{(\tau)}(z|y) + \overline{\delta}_{x,y,z})
\end{aligned}$$

and we can write the desired total variation explicitly as

$$\begin{aligned}
&\mathsf{d}_{\mathsf{TV}}(A_i, A'_i) \\
&= \frac{1}{2} \sum_{x,y,z \in \mathcal{X}} |\pi_x \cdot p^{(\tau)}(y|x) \cdot p^{(\tau)}(z|y) - \pi_x \cdot (p^{(\tau)}(y|x) + \delta_{x,y}) \cdot (p^{(\tau)}(z|y) + \overline{\delta}_{x,y,z})| \\
&= \frac{1}{2} \sum_{x,y,z \in \mathcal{X}} |\pi_x \cdot p^{(\tau)}(y|x) \cdot \overline{\delta}_{x,y,z} + \pi_x \cdot p^{(\tau)}(z|y) \cdot \delta_{x,y} + \pi_x \cdot \delta_{x,y} \cdot \overline{\delta}_{x,y,z}| \\
&\leq \sum_{x,y,z \in \mathcal{X}} \frac{1}{2} |\pi_x \cdot p^{(\tau)}(y|x) \cdot \overline{\delta}_{x,y,z}| + \sum_{x,y,z \in \mathcal{X}} \frac{1}{2} |\pi_x \cdot p^{(\tau)}(z|y) \cdot \delta_{x,y}| + \sum_{x,y,z \in \mathcal{X}} \frac{1}{2} |\pi_x \cdot \delta_{x,y} \cdot \overline{\delta}_{x,y,z}| \\
&\overset{\text{(i)}}{\leq} (\epsilon + \epsilon^2) + \epsilon + (\epsilon + \epsilon^2) = 3\epsilon + 2\epsilon^2 \leq 4\epsilon,
\end{aligned} \tag{56}$$

where we claim that the bound in step (i) holds term by term and the last inequality uses the fact that $\epsilon \leq 1/2$. It remains to prove step (i). The first term can be bounded as

$$\begin{aligned}
\sum_{x,y,z \in \mathcal{X}} |\pi_x \cdot p^{(\tau)}(y|x) \cdot \overline{\delta}_{x,y,z}| = \sum_{x,y \in \mathcal{X}} \pi_x \cdot p^{(\tau)}(y|x) \sum_{z \in \mathcal{X}} |\overline{\delta}_{x,y,z}| &\leq \sum_{x,y \in \mathcal{X}} \pi_x \cdot p^{(\tau)}(y|x) \cdot (2\epsilon + 2\epsilon^2) \\
&= 2\epsilon + 2\epsilon^2.
\end{aligned}$$

The remaining terms can be bounded using similar logic, so we omit the steps for brevity.

The consequence of the TV bound for expectations of bounded functions follows by the definition of total variation distance. ■

Lemma 14 *Fix a positive scalar $\epsilon \leq 1/2$, and let $\tau \geq \mathsf{t}_{\mathsf{mix}}(\epsilon)$ be an integer. For each $i_1 < i_2 \in [n]$ with $i_2 - i_1 \geq 2\tau$, define the stochastic sub-processes*

$$Z_{i_1,i_2} = (X_1, X_2, \ldots, X_{i_1-\tau}, X_{i_1}, X_{i_1+\tau}, \ldots, X_{i_2-\tau}, X_{i_2}, X_{i_2+\tau}, \ldots, X_n), \tag{57}$$

$$Z'_{i_1,i_2} = (X_1, X_2, \ldots, X_{i_1-\tau}, X'_{i_1}, X_{i_1+\tau}, \ldots, X_{i_2-\tau}, X'_{i_2}, X_{i_2+\tau}, \ldots, X_n), \tag{58}$$

where $X'_{i_1}, X'_{i_2} \sim \pi$ are drawn independently of each other and of everything else. Then we have $\mathsf{d}_{\mathsf{TV}}(Z_{i_1,i_2}, Z'_{i_1,i_2}) \leq 8\epsilon$.

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

$$|\,\mathbb{E}[f(Z_{i,j}) - f(Z'_{i,j})]| \le 8\epsilon.$$

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

As in the proof of Lemma 13, define the sub-processes

$$A_{i_1,i_2} = (X_{i_1-\tau}, X_{i_1}, X_{i_1+\tau}, X_{i_2-\tau}, X_{i_2}, X_{i_2+\tau}),$$
$$A'_{i_1,i_2} = (X_{i_1-\tau}, X'_{i_1}, X_{i_1+\tau}, X_{i_2-\tau}, X'_{i_2}, X_{i_2+\tau}).$$

(Note that in the special case where $i_2 - i_1 = 2\tau$, we have $i_1 + \tau = i_2 - \tau$ and so the above definition remains valid — the sub-process in this case contains a duplicated random variable $X_{i_1+\tau} = X_{i_2-\tau}$.)

We also define an intermediate sub-process $\widetilde{A}_{i_1,i_2} = (X_{i_1-\tau}, X'_{i_1}, X_{i_1+\tau}, X_{i_2-\tau}, X_{i_2}, X_{i_2+\tau})$ for convenience. Then, simplifying the expression for total variation over the entire Markov chain and noting that the stochastic processes $Z_{i_1,i_2}, Z'_{i_1,i_2}$ follow identical transition laws over the indices $\{1,\dots,i_1-\tau\} \cup \{i_1+\tau+1,\dots,i_2-\tau\} \cup \{i_2+\tau+1,\dots,n\}$, we have

$$\mathsf{d_{TV}}(Z_{i_1,i_2}, Z'_{i_1,i_2}) = \mathsf{d_{TV}}(A_{i_1,i_2}, A'_{i_1,i_2}) \overset{\text{(i)}}{\le} \mathsf{d_{TV}}(A_{i_1,i_2}, \widetilde{A}_{i_1,i_2}) + \mathsf{d_{TV}}(\widetilde{A}_{i_1,i_2}, A'_{i_1,i_2}),$$

where step (i) follows by the triangle inequality. We proceed to upper bound each of the terms $\mathsf{d_{TV}}(A_{i_1,i_2}, \widetilde{A}_{i_1,i_2})$ and $\mathsf{d_{TV}}(\widetilde{A}_{i_1,i_2}, A'_{i_1,i_2})$ by 4ϵ each, using a similar argument to the proof of Lemma 13. We denote $\rho_{y_2,w_1} = \Pr[X_{i_2-\tau} = y_2 | X_{i_1+\tau} = w_1]$ as shorthand, noting that for each $w_1 \in \mathcal{X}$, we have $\sum_{y_2 \in \mathcal{X}} \rho_{y_2,w_1} = 1$. (Note that in the special case where $i_2 - i_1 = 2\tau$, we have $X_{i_1+\tau} = X_{i_2-\tau}$ and this conditional distribution takes on the special form $\rho_{y_2,w_1} = \mathbb{I}\{y_2 = w_1\}$.)

We begin with the first term $\mathsf{d_{TV}}(A_{i_1,i_2}, \widetilde{A}_{i_1,i_2})$ and characterize the distributions of $A_{i_1,i_2}, \widetilde{A}_{i_1,i_2}$ as below:

$$\begin{aligned}
&\Pr\{A_{i_1,i_2} = (y_1,z_1,w_1,y_2,z_2,w_2)\} \\
&\qquad = \pi_{y_1} \cdot p^{(\tau)}(z_1|y_1) \cdot p^{(\tau)}(w_1|z_1) \cdot \rho_{y_2,w_1} \cdot p^{(\tau)}(z_2|y_2) \cdot p^{(\tau)}(w_2|z_2) \\
&\Pr\{\widetilde{A}_{i_1,i_2} = (y_1,z_1,w_1,y_2,z_2,w_2)\} \\
&\qquad = \pi_{y_1} \cdot \pi_{z_1} \cdot p^{(2\tau)}(w_1|y_1) \cdot \rho_{y_2,w_1} \cdot p^{(\tau)}(z_2|y_2) \cdot p^{(\tau)}(w_2|z_2).
\end{aligned}$$

Recalling the shorthand notation $\delta_{x,y}, \overline{\delta}_{x,y,z}$ defined in the proof of Lemma 13, we can write the above as

$$\begin{aligned}
&\Pr\{A_{i_1,i_2} = (y_1,z_1,w_1,y_2,z_2,w_2)\} \\
&\quad = \pi_{y_1} \cdot p^{(\tau)}(z_1|y_1) \cdot p^{(\tau)}(w_1|z_1) \cdot \rho_{y_2,w_1} \cdot p^{(\tau)}(z_2|y_2) \cdot p^{(\tau)}(w_2|z_2) \\
&\Pr\{\widetilde{A}_{i_1,i_2} = (y_1,z_1,w_1,y_2,z_2,w_2)\} \\
&\quad = \pi_{y_1} \cdot (p^{(\tau)}(z_1|y_1) + \delta_{y_1,z_1}) \cdot (p^{(\tau)}(w_1|z_1) + \overline{\delta}_{y_1,z_1,w_1}) \cdot \rho_{y_2,w_1} \cdot p^{(\tau)}(z_2|y_2) \cdot p^{(\tau)}(w_2|z_2).
\end{aligned}$$

Next, we note that $\rho_{y_2,w_1} \cdot p^{(\tau)}(z_2|y_2) \cdot p^{(\tau)}(w_2|z_2) = \Pr[X_{i_2-\tau} = y_2, X_{i_2} = z_2, X_{i_2+\tau} = w_2 | X_{i_1+\tau} = w_1]$ which is a conditional probability distribution that is identical for the

stochastic processes $\widetilde{A}_{i_1,i_2}$ and A_{i_1,i_2}. Therefore, we have $\sum_{y_2,z_2,w_2\in\mathcal{X}} \rho_{y_2,w_1}\cdot p^{(\tau)}(z_2|y_2)\cdot p^{(\tau)}(w_2|z_2) = 1$ for any value of $w_1 \in \mathcal{X}$. This yields

$$\begin{aligned}
&\mathsf{d_{TV}}(A_{i_1,i_2}, \widetilde{A}_{i_1,i_2}) \\
&= \frac{1}{2} \sum_{\substack{y_1,z_1,w_1\in\mathcal{X}\\ y_2,z_2,w_2\in\mathcal{X}}} \Big| \pi_{y_1}\cdot p^{(\tau)}(z_1|y_1)\cdot p^{(\tau)}(w_1|z_1)\cdot \rho_{y_2,w_1}\cdot p^{(\tau)}(z_2|y_2)\cdot p^{(\tau)}(w_2|z_2) \\
&\qquad - \pi_{y_1}\cdot (p^{(\tau)}(z_1|y_1)+\delta_{y_1,z_1})\cdot (p^{(\tau)}(w_1|z_1)+\overline{\delta}_{y_1,z_1,w_1})\cdot \rho_{y_2,w_1}\cdot p^{(\tau)}(z_2|y_2)\cdot p^{(\tau)}(w_2|z_2)\Big| \\
&= \sum_{y_1,z_1,w_1\in\mathcal{X}} \frac{1}{2}\Big|\pi_{y_1}p^{(\tau)}(z_1|y_1)p^{(\tau)}(w_1|z_1) - \pi_{y_1}(p^{(\tau)}(z_1|y_1)+\delta_{y_1,z_1})(p^{(\tau)}(w_1|z_1)+\overline{\delta}_{y_1,z_1,w_1})\Big|.
\end{aligned}$$

We have thus arrived at an expression for $\mathsf{d_{TV}}(A_{i_1,i_2}, \widetilde{A}_{i_1,i_2})$ that is identical to Equation (56), which is upper bounded by 4ϵ. Therefore, we have $\mathsf{d_{TV}}(A_{i_1,i_2}, \widetilde{A}_{i_1,i_2}) \le 4\epsilon$.

We now use a similar technique to bound the other term $\mathsf{d_{TV}}(\widetilde{A}_{i_1,i_2}, A'_{i_1,i_2})$. In particular, we have

$$\begin{aligned}
&\Pr\{\widetilde{A}_{i_1,i_2} = (y_1,z_1,w_1,y_2,z_2,w_2)\} \\
&= \pi_{y_1}\cdot\pi_{z_1}\cdot p^{(2\tau)}(w_1|y_1)\cdot\rho_{y_2,w_1}\cdot p^{(\tau)}(z_2|y_2)\cdot p^{(\tau)}(w_2|z_2) \\
&\Pr\{A'_{i_1,i_2} = (y_1,z_1,w_1,y_2,z_2,w_2)\} \\
&= \pi_{y_1}\cdot\pi_{z_1}\cdot p^{(2\tau)}(w_1|y_1)\cdot\rho_{y_2,w_1}\cdot (p^{(\tau)}(z_2|y_2)+\delta_{y_2,z_2})\cdot (p^{(\tau)}(w_2|z_2)+\overline{\delta}_{y_2,z_2,w_2}).
\end{aligned}$$

This time, we note that $\pi_{y_1}\cdot\pi_{z_1}\cdot p^{(2\tau)}(w_1|y_1)\cdot\rho_{y_2,w_1} = \Pr[X_{i_1-\tau}=y_1, X'_{i_1}=z_1, X_{i_1+\tau}=w_1, X_{i_2-\tau}=y_2]$ and therefore $\sum_{y_1,z_1,w_1\in\mathcal{X}}\pi_{y_1}\cdot\pi_{z_1}\cdot p^{(2\tau)}(w_1|y_1)\cdot\rho_{y_2,w_1} = \Pr[X_{i_2-\tau}=y_2] = \pi_{y_2}$. Using a similar series of steps to the preceding calculation, we obtain

$$\begin{aligned}
&\mathsf{d_{TV}}(\widetilde{A}_{i_1,i_2}, A'_{i_1,i_2}) \\
&= \sum_{y_2,z_2,w_2\in\mathcal{X}} \frac{1}{2}\Big|\pi_{y_2}p^{(\tau)}(z_2|y_2)p^{(\tau)}(w_2|z_2) - \pi_{y_2}(p^{(\tau)}(z_2|y_2)+\delta_{y_2,z_2})(p^{(\tau)}(w_2|z_2)+\overline{\delta}_{y_2,z_2,w_2})\Big| \\
&\le 4\epsilon
\end{aligned}$$

by an identical argument to Eq. (56). Putting these together yields $\mathsf{d_{TV}}(Z_{i_1,i_2}, Z'_{i_1,i_2}) \le 8\epsilon$. ■

Appendix B. Intuition for data-dependent tuning of window size τ

In this section, we provide some simple intuition to justify the data-dependent tuning procedure for the window size τ that we described in Section 6.1. Assuming that[8] $n \gg \mathsf{T_{mix}}$, we have that $Z^{(1)}$ and $Z^{(2)}$ are near-independent since they are significantly separated within the sequence. Thus, conditioned on $Z^{(1)}$, the sequence $Z^{(2)}$ should be thought of as an independent Markov chain started at the stationary distribution π. Consequently, the random variable $\widetilde{M}(Z^{(1)})$ ought to be close to the estimand $M_\pi(Z^{(1)})$, and this can be formalized

8. If $n \lesssim \mathsf{T_{mix}}$, it is impossible to obtain consistent estimation anyway, at least in a minimax sense.

via a bounded differences inequality for mixing Markov chains (Paulin, 2015). Indeed, if independence between $Z^{(1)}$ and $Z^{(2)}$ held exactly, then it is straightforward to show that with high probability over the randomness in $Z^{(2)}$, we have

$$|\widetilde{M}(Z^{(1)}) - M_\pi(Z^{(1)})|^2 \lesssim \frac{\mathsf{T}_{\mathsf{mix}} \log(n/\mathsf{T}_{\mathsf{mix}})}{n}. \tag{59}$$

Now Theorem 1 guarantees that for some $\tau_0 \asymp \mathsf{T}_{\mathsf{mix}} \log(n/\mathsf{T}_{\mathsf{mix}})$, we must have the inequality[9] $\left|\widehat{M}_{\textsc{WingIt}}(Z^{(1)}; \tau_0) - M_\pi(Z^{(1)})\right|^2 \lesssim \frac{\tau}{n}$. Combining this observation with Ineq. (59) and noting that $\tau_0 \asymp \mathsf{T}_{\mathsf{mix}} \log(n/\mathsf{T}_{\mathsf{mix}})$, we have

$$\begin{aligned}\left|\widehat{M}_{\textsc{WingIt}}(Z^{(1)}; \tau_0) - \widetilde{M}(Z^{(1)})\right|^2 &\leq 2\left|\widehat{M}_{\textsc{WingIt}}(Z^{(1)}; \tau_0) - M_\pi(Z^{(1)})\right|^2 + 2\left|M_\pi(Z^{(1)}) - \widetilde{M}(Z^{(1)})\right|^2 \\ &\lesssim \frac{\tau_0}{n}.\end{aligned}$$

Thus, we see that Ineq. (28) is a reasonable validation criterion since it is satisfied for some choice of window size at most τ_0, for a suitable choice of constant C_{tune} on the RHS. Conversely, if Ineq. (28) holds for some smaller window size $\tau = \widehat{\tau} \leq \tau_0$, then combining this with Ineq. (59) yields

$$\begin{aligned}\frac{1}{2}\left|\widehat{M}_{\textsc{WingIt}}(Z^{(1)}; \widehat{\tau}) - M_\pi(Z^{(1)})\right|^2 &\leq \left|\widehat{M}_{\textsc{WingIt}}(Z^{(1)}; \widehat{\tau}) - \widetilde{M}(Z^{(1)})\right|^2 + \left|M_\pi(Z^{(1)}) - \widetilde{M}(Z^{(1)})\right|^2 \\ &\lesssim \frac{\widehat{\tau}}{n} + \frac{\mathsf{T}_{\mathsf{mix}} \log(n/\mathsf{T}_{\mathsf{mix}})}{n} \\ &\leq \frac{\tau_0}{n} + \frac{\mathsf{T}_{\mathsf{mix}} \log(n/\mathsf{T}_{\mathsf{mix}})}{n} \lesssim \frac{\mathsf{T}_{\mathsf{mix}} \log(n/\mathsf{T}_{\mathsf{mix}})}{n}.\end{aligned}$$

Putting together the pieces, we see that our validation procedure is reasonable since (a) It is satisfied by the window size τ_0 prescribed by Theorem 1, and (b) It always produces a good value of tuning window size $\widehat{\tau}$, in that this choice of window size leads to the optimal rate of estimation of the functional $M_\pi(Z^{(1)})$.

It is important to note that the above sketch does not constitute a rigorous argument. In order to make it rigorous, one would have to formally establish Eq. (59) and also a version of Theorem 1 that holds with high probability, both of which are interesting directions for future work.

9. Note that this step of the argument is heuristic, since Theorem 1 only gives such a guarantee in expectation.