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. 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 $\mathsf{T}_{\mathsf{mix}}/n$ up to a logarithmic factor. This constitutes, to our knowledge, the first such result 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 explore an automatic and data-dependent tuning method for the window size 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. We use $\oplus$ to denote the concatenation of sequences, e.g., $(X_1, X_2) \oplus (X_4, X_5) = (X_1, X_2, X_4, X_5)$. 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 by, 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

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

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)

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

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