Non-Bayesian Parametric Missing-Mass Estimation

Shir Cohen, Tirza Routtenberg, *Senior Member, IEEE*, and Lang Tong, *Fellow, IEEE*,

***Abstract*—We consider the classical problem of missing-mass estimation, which deals with estimating the total probability of unseen elements in a sample. The missing-mass estimation problem has various applications in machine learning, statistics, language processing, ecology, sensor networks, and others. The naive, constrained maximum likelihood (CML) estimator is inappropriate for this problem since it tends to overestimate the probability of the observed elements. Similarly, the conventional constrained Cramér-Rao bound (CCRB), which is a lower bound on the mean-squared-error (MSE) of unbiased estimators, does not provide a relevant bound on the performance for this problem. In this paper, we introduce a frequentist, non-Bayesian parametric model of the problem of missing-mass estimation. We introduce the concept of missing-mass unbiasedness by using the Lehmann unbiasedness definition. We derive a non-Bayesian CCRB-type lower bound on the missing-mass MSE (mmMSE), named the missing-mass CCRB (mmCCRB), based on the missing-mass unbiasedness. The missing-mass unbiasedness and the proposed mmCCRB can be used to evaluate the performance of existing estimators. Based on the new mmCCRB, we propose a new method to improve existing estimators by an iterative missing-mass Fisher scoring method. Finally, we demonstrate via numerical simulations that the proposed mmCCRB is a valid and informative lower bound on the mmMSE of state-of-the-art estimators for this problem: the CML, the Good-Turing, and Laplace estimators. We also show that the performance of the Laplace estimator is improved by using the new Fisher-scoring method.**

***Index Terms*—Non-Bayesian estimation, Good-Turing estimator, probability of missing mass, constrained Cramér-Rao bound, Lehmann unbiasedness**

I. Introduction

Given N samples from a population of elements belonging to different types with unknown proportions, how should one estimate the total probability of unseen types? This is a classical problem in statistics, commonly referred to as the missing-mass estimation problem [1, 2]. Missing-mass estimation has gained significant interest in various applications, such as ecological studies [3], sensor networks [4, 5], machine learning, and statistics. In the context of language processing, for example, estimation of new and existing words in text has applications such as language modeling, spelling correction, and word-sense disambiguation [6, 7]. Missing-mass estimation is especially important for applied problems where the sampling procedure is expensive, and the need for acquiring more data is determined by the possibility of observing new unobserved elements.

It is well known that the naive, constrained maximum likelihood (CML) estimator of the probability, i.e. the empirical probability, is ineffective if there are insufficient samples [8, 9]. In particular, the CML estimator assigns a zero probability to unseen events. Similarly, the mean-squared-error (MSE) is not suitable for the problem of estimation of the missing mass. As a result, the Cramér-Rao bound (CRB) on the MSE is inappropriate. Various estimators of the missing mass have been suggested over the years [2, 8, 10-18]. Despite their practical popularity, no objective evaluation or optimality results for these estimators have been established [8]. In particular, there is no comprehensive non-Bayesian estimation theory for estimating the missing mass and new performance bounds are required.

A. Summary of results

In this paper, we consider the problem of estimating the missing mass, where it is assumed that we observe samples that are drawn uniformly from an unknown, stationary distribution. First, we introduce a non-Bayesian parametric formulation of this estimation problem. We introduce a suitable cost function, the missing-mass squared-error, and derive the associated Lehmann unbiasedness for this cost. We develop a new non-Bayesian constrained Cramér-Rao bound (CCRB), the missing-mass CCRB (mmCCRB), which is a lower bound on the missing-mass MSE (mmMSE) of any estimator with a specific Lehmann bias. The new bound is obtained by using linear parametric constraints on the probability space and the Lehmann unbiasedness. We investigate the properties of the mmCCRB and some special cases of this bound. Based on the new mmCCRB, we propose a new method to improve existing estimators by an iterative missing-mass Fisher scoring method. The new bound is examined in simulations and compared with the performance of state-of-the-art estimators: CML, Good-Turing and Laplace estimators. We also show that performance of the Laplace estimator is improved by using the new Fisher-scoring method.

B. Related works

The Good-Turing probability estimator [2], which was invented to decipher the Enigma code during World War II, its extensions by smoothing techniques [11, 12], and the Laplace estimator [10, 13, 14], have been shown to be useful for the estimation of the probability of unseen elements [8, 15], and have been implemented in many practical applications. On the theoretical side, theoretical interpretations of the Good-Turing estimator have been proposed [19], and its performance in terms of attenuation has been established in [8]. For example, using a uniform prior, Good gave a derivation of the Good-Turing estimator from a Bayesian point of view [2]. Some works discussed the properties of the Good-Turing estimator, including its bias [2, 20], confidence intervals and convergence rate [21], and (un)consistency [22], as well as lower and upper bounds on the expected missing mass [23], and concentration

This work is partially supported by the ISRAEL SCIENCE FOUNDATION (ISF), grant No. 1173/16. The work of Shir Cohen was supported under a grant from the Ministry of Science and Technology of Israel.

S. Cohen and T. Routtenberg are with the School of Electrical and Computer Engineering Ben-Gurion University of the Negev Beer-Sheva 84105, Israel, e-mail: shiru@post.bgu.ac.il, tirzar@bgu.ac.il. L. Tong is with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853 USA (e-mail:lt35@cornell.edu).

inequalities for the missing mass [24]. The performance of the Good-Turing estimator is analyzed using the theory of large deviations in [4]. The sequence attenuation of the estimator was proposed as a performance measure in [8]. Upper bounds on the MSE of missing-mass estimation problems have been proposed in [25].

However, there is no comprehensive non-Bayesian estimation theory for estimating the missing mass. This theory is crucial for the system design, error analysis, and quality assessments of existing estimation methods and for the development of new non-Bayesian estimators. In particular, the CCRB [26-29], which is associated with the CML estimator, is unsuited as a bound on the performance of Good-Turing estimators outside the asymptotic region, while it provides a lower bound on the MSE of any χ-unbiased estimator [28-30]. Our recent works on non-Bayesian estimation after selection [31-33] suggest that conditional schemes, in which the performance criterion depends on the observed data, require different CRB-type bounds.

C. Organization and notation

The remainder of the paper is organized as follows: Section II presents the non-Bayesian parametric model of missing-mass estimation under a multinomial model, including the conventional CCRB and constrained unbiasedness for this model and the appropriate cost function. In Section III, we derive a new CCRB-type lower bound on the mmMSE. Numerical simulations are presented in Section IV. Finally, our conclusions can be found in Section V.

In the rest of this paper, vectors are denoted by boldface lowercase letters and matrices by boldface uppercase letters. The notations $\mathbf{1}_{\{A\}}$ and $\mathbf{I}$ denote the indicator function of an event A and the identity matrix, respectively. The vectors $\mathbf{1}$ and $\mathbf{0}$ are column vectors of ones and zeros, respectively, and $\mathbf{e}_m$ is the mth column of the identity matrix, all with appropriate dimensions. The matrix $\mathrm{diag}(\mathbf{a})$ denotes the diagonal matrix with vector $\mathbf{a}$ on the diagonal. The mth element of the vector $\mathbf{a}$, the (m,q)th element of the matrix $\mathbf{A}$, and the $(m_1 : m_2 \times q_1 : q_2)$ submatrix of $\mathbf{A}$ are denoted by a_m, $\mathbf{A}_{m,q}$, and $\mathbf{A}_{m_1:m_2,q_1:q_2}$, respectively. The trace of a matrix $\mathbf{A} \in \mathbb{R}^{M\times M}$ is defined as $\mathrm{trace}(\mathbf{A}) = \sum_{m=1}^{M} \mathbf{A}_{m,m}$. The gradient of a vector function, $\mathbf{c}$, of $\boldsymbol{\theta}$, $\nabla_{\boldsymbol{\theta}}\mathbf{c}$, is a matrix in $\mathbb{R}^{K\times M}$, with the (k,m)th element equal to $\frac{\partial c_k}{\partial \theta_m}$, where $\mathbf{c} = [c_1,\ldots,c_K]^T$ and $\boldsymbol{\theta} = [\theta_1,\ldots,\theta_M]^T$. For a scalar function c, we denote $\nabla_{\boldsymbol{\theta}}^T c \triangleq (\nabla_{\boldsymbol{\theta}} c)^T$, and $\nabla_{\boldsymbol{\theta}}^2 c \triangleq \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}}^T c$. The notations $\mathrm{E}_{\boldsymbol{\theta}}[\cdot]$ and $\mathrm{E}_{\boldsymbol{\theta}}[\cdot|A]$ represent the expectation and conditional expectation operators, parametrized by a deterministic vector, $\boldsymbol{\theta}$, and given the event A. For a set X, $|X|$ represents its cardinality.

II. Non-Bayesian Estimation of the Missing Mass

In this section we present the problem of estimating the missing mass as a non-Bayesain parameter estimation. In Subsection II-A we describe the observation model and the relevant probability functions. In Subsection II-B we develop the χ-unbiasedness and the CCRB for estimating the unknown probability mass function under this model. Finally, in Subsection II-C we formulate the missing-mass estimation problem and present the missing-mass squared-error cost function, which is used in this paper.

A. Non-Bayesian model

Assume that there is a set of M symbols, $\mathcal{S} = \{s_1,\ldots,s_M\}$, where $M \geq 1$ is assumed to be known. The elements in $\mathcal{S}$ may represent, for example, species in the jungle [8], words in a dictionary [6, 7], or operating sensors [4, 5]. The true probability of observing symbol s_m is denoted by θ_m, $\forall m = 1,\ldots,M$, where $\theta_m \neq 0$ for all $1 \leq m \leq M$ and $\sum_{m=1}^{M} \theta_m = 1$. Thus, $\boldsymbol{\theta} \triangleq [\theta_1,\ldots,\theta_M]^T$ is a probability mass function (pmf) vector over the discrete and finite set of symbols, $\mathcal{S}$. As a result, $\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}$, where

$$\Omega_{\boldsymbol{\theta}} \triangleq \left\{ \boldsymbol{\theta} \in [0,1]^M \,|\, f(\boldsymbol{\theta}) = 0 \right\}, \tag{1}$$

in which

$$f(\boldsymbol{\theta}) \triangleq \sum_{m=1}^{M} \theta_m - 1. \tag{2}$$

Under the independent and identically distributed (i.i.d.) multinomial model [2], it is assumed that there are N i.i.d. samples, $\{x_n\}_{n=1}^{N}$, drawn according to the pmf described by the unknown vector, $\boldsymbol{\theta}$. We consider the problem of estimating the missing mass of the unobserved symbols, which is a function of $\boldsymbol{\theta}$. It can be verified that the pmf of the random observation vector, $\mathbf{x} \triangleq [x_1,\ldots,x_N]^T \in \mathcal{S}^N$, is a binomial distribution:

$$p(\mathbf{x};\boldsymbol{\theta}) = \prod_{m=1}^{M} \theta_m^{C_{N,m}(\mathbf{x})},\ \mathbf{x} \in \mathcal{S}^N, \tag{3}$$

where

$$C_{N,m}(\mathbf{x}) \triangleq \sum_{n=1}^{N} \mathbf{1}_{\{x_n = s_m\}},\ m = 1,\ldots,M, \tag{4}$$

is the number of times that the mth element was observed out of the N samples. Therefore, the vector $[C_{N,0}(\mathbf{x}),\ldots,C_{N,M}(\mathbf{x})]^T$ has a multinomial distribution with parameters N and $\boldsymbol{\theta}$. The pmf in (3) can also be written as

$$p(\mathbf{x};\boldsymbol{\theta}) = \prod_{m=1}^{M} \theta_m^{\sum_{r=0}^{N} r \mathbf{1}_{\{s_m \in G_{N,r}(\mathbf{x})\}}}, \tag{5}$$

where $G_{N,r}(\mathbf{x})$ is the group of elements that appear exactly r times in the N-length observation vector, $\mathbf{x}$. That is, if $s_m \in G_{N,r}(\mathbf{x})$ then $C_{N,m}(\mathbf{x}) = r$. In particular, the set

$$G_{N,0}(\mathbf{x}) \triangleq \{m : m = 1,\ldots,M, s_m \neq x_n,\ \forall n = 0,\ldots,N\} \tag{6}$$

is the set of elements which do not appear in the observation vector, $\mathbf{x}$, i.e. the missing mass, with $C_{N,m}(\mathbf{x}) = 0$.

Let us define the subspace of all observation vectors that do not include s_m as

$$\mathcal{A}_m \triangleq \{\mathbf{x} \in \mathcal{S}^N : m \in G_{N,0}(\mathbf{x})\},\ m = 1,\ldots,M. \tag{7}$$

For a given number of measurements, N, the probability of the mth element being unobserved in these measurements is

$$\Pr(\mathbf{x} \in \mathcal{A}_m;\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}\left[\mathbf{1}_{\{m \in G_{N,0}(\mathbf{x})\}}\right] = (1-\theta_m)^N, \tag{8}$$

$\forall m = 1,\ldots,M$. By using Bayes rule it can be seen that

$$p(\mathbf{x}|\mathbf{x} \in \mathcal{A}_m;\boldsymbol{\theta}) = \begin{cases} \frac{p(\mathbf{x};\boldsymbol{\theta})}{\Pr(\mathbf{x} \in \mathcal{A}_m;\boldsymbol{\theta})} & \text{if } \mathbf{x} \in \mathcal{A}_m \\ 0 & \text{otherwise} \end{cases}, \tag{9}$$

$m = 1, \ldots, M$. By substituting (3) and (8) in (9) we obtain

$$p(\mathbf{x}|\mathbf{x} \in \mathcal{A}_m; \boldsymbol{\theta}) = \begin{cases} \frac{\prod_{l=1}^{M} \theta_l^{C_{N,l}(\mathbf{x})}}{(1-\theta_m)^N} & \text{if } \mathbf{x} \in \mathcal{A}_m \\ 0 & \text{otherwise} \end{cases}, \quad (10)$$

$m = 1, \ldots, M$.

We denote by $\hat{\boldsymbol{\theta}} : \mathcal{S}^N \rightarrow \Omega_{\boldsymbol{\theta}}$ an arbitrary estimator of the pmf vector, $\boldsymbol{\theta}$, based on the observation vector, $\mathbf{x}$. The CML estimator of $\boldsymbol{\theta}$ under the parametric constraint $f(\boldsymbol{\theta}) = 0$, where $f(\cdot)$ is defined in (2), is given by

$$\hat{\theta}_m^{\text{CML}} = \frac{C_{N,m}(\mathbf{x})}{N}, \; m = 1, \ldots, M. \quad (11)$$

In particular, the CML estimator assigns zero probability for unseen elements, i.e. for the missing mass. Some alternative estimators are presented in Subsection III-D.

In order to clarify our notations, we give the following example.

Example 1: Let us assume that the observation vector is $\mathbf{x} = [a, a, c, c, c, e, e, f, h, h]^T$ with $N = 10$ and $\mathcal{S} = \{a, b, c, d, e, f, g, h\}$. Then, according to (4), the histogram values are $C_{10,1}(\mathbf{x}) = 2$, $C_{10,3}(\mathbf{x}) = 3$, $C_{10,5}(\mathbf{x}) = 2$, $C_{10,6}(\mathbf{x}) = 1$, and $C_{10,7}(\mathbf{x}) = 2$. The different subsets of elements are $G_{10,0}(\mathbf{x}) = \{b, d, g\}$, $G_{10,1}(\mathbf{x}) = \{f\}$, $G_{10,2}(\mathbf{x}) = \{a, e, h\}$, $G_{10,3}(\mathbf{x}) = \{c\}$ and $G_{10,0} = \{b, d, g\}$. The CML estimator in this case is $\hat{\boldsymbol{\theta}}^{\text{CML}} = [0.2, 0, 0.3, 0, 0.2, 0.1, 0, 0.2]^T$. The missing mass is the total probability mass of the elements in $G_{10,0}$.

In general constrained parameter estimation, the null-space matrix, orthogonal to the constraints, plays an important role [26, 27]. In particular, for the considered model the unknown parameter vector, $\boldsymbol{\theta}$, is restricted to satisfy the parametric constraint

$$f(\boldsymbol{\theta}) = 0, \quad (12)$$

where $f(\cdot)$ is defined in (2). The gradient of $f(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ is

$$\mathbf{F} \stackrel{\triangle}{=} \nabla_{\boldsymbol{\theta}}^T f(\boldsymbol{\theta}) = \mathbf{1}_M^T. \quad (13)$$

In addition, there exists a null-space matrix $\mathbf{U} \in \mathbb{R}^{M \times (M-1)}$ such that

$$\mathbf{F}\mathbf{U} = \mathbf{1}_M^T \mathbf{U} = \mathbf{0}^T, \; \mathbf{U}^T\mathbf{U} = \mathbf{I}. \quad (14)$$

In particular, it can be verified that $\mathbf{U}^T\mathbf{y} = \mathbf{0}_{M-1}$ *iff* $\mathbf{y} = c\mathbf{1}_M$, where $c \in \mathbb{R}$ is an arbitrary constant.

B. CCRB and constrained unbiasedness

In this subsection we develop the conventional CCRB and the unbiasedness condition for estimating $\boldsymbol{\theta}$ under the considered model. The CCRB [26, 27] provides a lower bound on the MSE of any locally χ-unbiased estimator [28-30], which is a weaker requirement than ordinary mean unbiasedness, and is defined as follows.

Definition 1: An estimator $\hat{\boldsymbol{\theta}} : \mathcal{S}^N \rightarrow \Omega_{\boldsymbol{\theta}}$ is said to be a locally χ-unbiased estimator in the neighborhood of $\tilde{\boldsymbol{\theta}} \in \Omega_{\boldsymbol{\theta}}$ if it satisfies

$$\mathbf{U}^T \mathrm{E}_{\tilde{\boldsymbol{\theta}}}[\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}] = \mathbf{0}_{M-1} \quad (15)$$

and

$$\left\{\nabla_{\boldsymbol{\theta}}^T \mathrm{E}_{\boldsymbol{\theta}}[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}]\right\}\Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} \mathbf{U} = \mathbf{0}_{M \times (M-1)}, \quad (16)$$

where $\mathbf{U}$ is defined in (14).

It should be noted that in this paper the notation $\tilde{\boldsymbol{\theta}}$ represents a specific value (or "local" value) of the unknown parameter vector in $\Omega_{\boldsymbol{\theta}}$, while $\boldsymbol{\theta}$ is used as a general parameter in the different functions. For the CML estimator in (11) we obtain that

$$\mathrm{E}_{\boldsymbol{\theta}}\left[\hat{\theta}_m^{\text{CML}} - \theta_m\right] = \mathrm{E}_{\boldsymbol{\theta}}\left[\frac{C_{N,m}(\mathbf{x})}{N} - \theta_m\right] = 0, \quad (17)$$

for all $m = 1, \ldots, M$ and for any $\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}$, where the last equality follows from the mean of a variable with a multinomial distribution. Thus, (17) implies that the CML estimator satisfies Definition 1 and it is a locally χ-unbiased estimator for any $\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}$, and, thus, it is a uniformly χ-unbiased estimator. In addition, it was shown in [28] that for linear parametric constraints, which is the case for the considered model with the constraint in (12), the χ-unbiasedness coincides with the C-unbiasedness in the Lehmann sense [34] w.r.t. the squared-error cost function.

The CCRB on the MSE of any unbiased estimator in the sense of Definition 1 at $\tilde{\boldsymbol{\theta}} \in \Omega_{\boldsymbol{\theta}}$ is given by [27-30]

$$\mathrm{E}_{\tilde{\boldsymbol{\theta}}}\left[(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})^T\right] \succeq \mathbf{U}(\mathbf{U}^T\mathbf{J}(\tilde{\boldsymbol{\theta}})\mathbf{U})^{-1}\mathbf{U}^T, \quad (18)$$

where the conventional Fisher information matrix (FIM) is defined as

$$\mathbf{J}(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}\left[\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^T \log p(\mathbf{x}; \boldsymbol{\theta})\right], \; \boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}. \quad (19)$$

Taking the logarithm of (3) yields the following log-likelihood function

$$\log p(\mathbf{x}; \boldsymbol{\theta}) = \sum_{m=1}^{M} C_{N,m}(\mathbf{x}) \log \theta_m, \; \mathbf{x} \in \mathcal{S}^N. \quad (20)$$

By substituting the derivative of (20) w.r.t. $\boldsymbol{\theta}$ in (19), we obtain that the (l, m) element of the FIM is given by

$$\begin{aligned} [\mathbf{J}(\boldsymbol{\theta})]_{l,m} &= \frac{1}{\theta_l \theta_m} \mathrm{E}_{\boldsymbol{\theta}}\left[C_{N,l}(\mathbf{x}) C_{N,m}(\mathbf{x})\right] \\ &= \begin{cases} N(N-1) & m \neq l \\ N^2 + N\frac{1-\theta_m}{\theta_m} & m = l \end{cases}, \end{aligned} \quad (21)$$

$\forall m, l = 1, \ldots, M$, where $C_{N,m}(\mathbf{x})$, $m = 1, \ldots, M$, are defined in (4) and the last equality holds by using known results on the moments of the multinomial distributed variables [35]. By using the elements in (21), the FIM for the probability estimation model can be written in a matrix form as

$$\mathbf{J}(\boldsymbol{\theta}) = N(N-1)\mathbf{1}\mathbf{1}^T + N\mathrm{diag}^{-1}(\boldsymbol{\theta}), \quad (22)$$

where $\mathrm{diag}^{-1}(\boldsymbol{\theta}) = (\mathrm{diag}(\boldsymbol{\theta}))^{-1}$. It should be noted that $\mathbf{J}(\boldsymbol{\theta})$ from (22) is a well-defined, non-singular matrix, since we assume that $\theta_m \neq 0$, $\forall m = 1, \ldots, M$. By substituting (22) in (18), one obtains the following closed-form CCRB on the MSE under the constraint in (12):

$$\begin{aligned} \mathrm{E}_{\tilde{\boldsymbol{\theta}}}\left[(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})^T\right] &\succeq \frac{1}{N(N-1)}\mathbf{U}(\mathbf{U}^T\mathbf{1}\mathbf{1}^T\mathbf{U})^{-1}\mathbf{U}^T \\ &\quad + \frac{1}{N}\mathbf{U}\left(\mathbf{U}^T\mathrm{diag}^{-1}(\tilde{\boldsymbol{\theta}})\mathbf{U}\right)^{-1}\mathbf{U}^T \\ &= \frac{1}{N}\mathbf{U}\left(\mathbf{U}^T\mathrm{diag}^{-1}(\tilde{\boldsymbol{\theta}})\mathbf{U}\right)^{-1}\mathbf{U}^T, \end{aligned} \quad (23)$$

where the last equality is obtained by substituting (14), which implies $\mathbf{1}_M^T\mathbf{U}=\mathbf{0}_{M-1}$. By applying the trace operator on the CCRB from (23) and using $\mathbf{U}^T\mathbf{U}=\mathbf{I}$ from (14) we obtain the bound on the trace MSE:

$$\sum_{m=1}^{M}\mathrm{E}_{\tilde{\boldsymbol{\theta}}}\left[(\hat{\theta}_m-\tilde{\theta}_m)^2\right]\geq B^{\mathrm{CCRB}}(\tilde{\boldsymbol{\theta}}), \tag{24}$$

where

$$B^{\mathrm{CCRB}}(\boldsymbol{\theta})\triangleq\frac{1}{N}\mathrm{trace}\left(\left(\mathbf{U}^T\mathrm{diag}^{-1}(\boldsymbol{\theta})\mathbf{U}\right)^{-1}\right). \tag{25}$$

However, the CCRB in (25), which is a lower bound on the MSE of χ-unbiased estimators, does not provide a relevant bound on the performance for the missing-mass estimation problem. This is similar to the mismatch of the naive CML estimator from (11), which tends to overestimate the probability of the observed elements. In the following section, we develop a new CCRB-type bound on the missing-mass estimation.

C. Performance criterion: missing-mass MSE risk

The missing mass, namely the total probability mass of the outcomes not observed in the samples in $\mathbf{x}$, is defined as

$$p_0(\mathbf{x},\boldsymbol{\theta})=\sum_{m=1}^{M}\theta_m\mathbf{1}_{\{m\in G_{N,0}(\mathbf{x})\}}. \tag{26}$$

It can be seen that the missing mass in (26) is a function of both the pmf vector, $\boldsymbol{\theta}$, and the observation vector, $\mathbf{x}$. Thus, some papers in the literature [36, 37] treat the estimation problem as the estimation of the *hybrid* parameter, $p_0(\mathbf{x},\boldsymbol{\theta})$, which allegedly has both random and deterministic parts. However, since the random observation vector, $\mathbf{x}$, is given, the true unknown part in the missing mass $p_0(\mathbf{x},\boldsymbol{\theta})$ is only the deterministic vector $\boldsymbol{\theta}$. Therefore, in this work we adopt the non-Bayesian approach for the estimation of *deterministic* parameters. Moreover, since all the elements of the pmf vector, $\boldsymbol{\theta}$, are unknown, we treat this estimation problem as the estimation of the parameters of interest that include the probabilities of unseen events and refer to the other (seen) parameters in $\boldsymbol{\theta}$ as nuisance parameters [38, 39].

The estimation approach is based on choosing an appropriate cost function. Taking practical considerations into account, a cost function should capture the relevant errors meaningfully and, at the same time, be easily computed. Direct calculation of the MSE of $p_0(\mathbf{x},\boldsymbol{\theta})$ from (26) involves computing the expectation based on a sum of 2^M possible events (represent the binary options that m is within/without $G_{N,0}(\mathbf{x})$, for any $\mathbf{x}$ and any $m=1,\ldots,M$) and, thus, is infeasible and leads to an intractable bound, unbiasedness conditions, and estimators. Therefore, we propose an alternative cost as follows. Since we are interested in estimating the parameters that belong to the missing mass, only estimation errors of elements with the indices that are in $G_{N,0}(\mathbf{x})$ from (6) should be penalized. Therefore, in this paper, we use the following missing-mass squared-error cost function:

$$C(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})\triangleq\sum_{m=1}^{M}(\hat{\theta}_m-\theta_m)^2\mathbf{1}_{\{m\in G_{N,0}(\mathbf{x})\}}, \tag{27}$$

for any estimator $\hat{\boldsymbol{\theta}}=[\hat{\theta}_1,\ldots,\hat{\theta}_M]^T$ of the pmf vector, $\boldsymbol{\theta}=[\theta_1,\ldots,\theta_M]^T$. The associated mmMSE risk, which is the expected value of (27), is

$$\begin{aligned}\mathrm{E}_{\boldsymbol{\theta}}\left[C(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})\right]&=\sum_{m=1}^{M}\mathrm{E}_{\boldsymbol{\theta}}\left[(\hat{\theta}_m-\theta_m)^2\mathbf{1}_{\{m\in G_{N,0}(\mathbf{x})\}}\right]\\&=\sum_{m=1}^{M}\mathrm{E}_{\boldsymbol{\theta}}\left[(\hat{\theta}_m-\theta_m)^2|\mathbf{x}\in\mathcal{A}_m\right]\Pr(\mathbf{x}\in\mathcal{A}_m;\boldsymbol{\theta}),\end{aligned} \tag{28}$$

where the last equality is obtained by using the law of total probability and the conditional distribution from (9). The use of the indicator functions implies that the error of the mth parameter, $\hat{\theta}_m-\theta_m$, affects the mmMSE only for observations $\mathbf{x}$ such that s_m has not been observed. Thus, it can be seen that the mmMSE is the sum of the MSE of the parameters of interest, that is, the probabilities of unseen events.

III. Missing-Mass Constrained Cramér-Rao (mmCCRB) Bound

In this section, a CCRB-type lower bound is derived. In Subsection III-A we develop the uniform and local unbiasedness in the Lehmann sense under the missing-mass squared-error cost function and under the probability-space parametric constraints. In Subsection III-B, we derive the proposed bound, which is a lower bound on the mmMSE and is a function of the Lehmann bias of the estimators. For the sake of generality, the unbiasedness and the mmCCRB are first derived for a general observation-model distribution $p(\mathbf{x};\boldsymbol{\theta})$. Thus, the missing-mass unbiasedness in Subsection III-A and the mmCCRB in Subsection III-B can be used for various variations of the missing-mass estimation problem, such as estimating an unknown Markov chain from its sample [24, 40, 41]. Then, in Subsection III-C, we develop the closed-form mmCCRB for the classical i.i.d model, given in (3). Finally, in Subsection III-D we present some special cases of the mmCCRB.

A. Lehmann unbiasedness

The mean-unbiasedness constraint is commonly used in non-Bayesian parameter estimation [42]. Lehmann [34] proposed a generalization of the unbiasedness concept, which is based on the considered cost function, as follows.

Definition 2: An estimator $\hat{\boldsymbol{\theta}}:\mathcal{S}^N\rightarrow\mathbb{R}^M$ is an unbiased estimator of $\boldsymbol{\theta}$ in the Lehmann sense [34] w.r.t. a given cost function, $C(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})$, if

$$\mathrm{E}_{\boldsymbol{\theta}}[C(\hat{\boldsymbol{\theta}},\boldsymbol{\eta})]\geq\mathrm{E}_{\boldsymbol{\theta}}[C(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})],\ \forall\boldsymbol{\eta},\boldsymbol{\theta}\in\Omega_{\boldsymbol{\theta}}, \tag{29}$$

where $\Omega_{\boldsymbol{\theta}}$ is the parameter space.

The Lehmann unbiasedness definition implies that an estimator is unbiased if, on average, it is "closer" to the true parameter $\boldsymbol{\theta}$ than to any other value in the parameter space (here, denoted by an arbitrary vector, $\boldsymbol{\eta}$). The measure of closeness is determined by the considered cost function, $C(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})$. Examples for Lehmann unbiasedness with different cost functions and under parametric constraints can be found in [28, 29, 31-34, 43]. The following proposition states the Lehmann unbiasedness for the estimation problem of missing-mass probability. To this

end, we define the elements of the missing-mass bias vector, $\mathbf{b}_{N,0}(\boldsymbol{\theta}) \in \mathbb{R}^M$, as follows:

$$[\mathbf{b}_{N,0}(\boldsymbol{\theta})]_m \triangleq \mathrm{E}_{\boldsymbol{\theta}}\left[(\hat{\theta}_m - \theta_m)\mathbf{1}_{\{m\in G_{N,0}(\mathbf{x})\}}\right] = \mathrm{E}_{\boldsymbol{\theta}}\left[\hat{\theta}_m - \theta_m | \mathbf{x}\in\mathcal{A}_m\right]\Pr(\mathbf{x}\in\mathcal{A}_m;\boldsymbol{\theta}), \quad (30)$$

$\forall m = 1,\ldots,M$, where the last equality is obtained by using the law of total probability.

Proposition 1: An estimator $\hat{\boldsymbol{\theta}} : \mathcal{S}^N \rightarrow \Omega_{\boldsymbol{\theta}}$ is said to be a *uniformly* Lehmann-unbiased estimator of $\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}$ w.r.t. the missing-mass squared-error cost function from (27) if

$$\mathbf{U}^T\mathbf{b}_{N,0}(\boldsymbol{\theta}) = \mathbf{0}_{M-1},\ \forall\boldsymbol{\theta}\in\Omega_{\boldsymbol{\theta}}, \quad (31)$$

where $\mathbf{U}$ and $\mathbf{b}_{N,0}(\boldsymbol{\theta})$ are defined in (14) and (30), respectively.

Proof: The proof appears in Appendix A. ∎

The CRB is a *local* bound, meaning that it determines the achievable performance at a particular value of $\boldsymbol{\theta}$, denoted here by $\tilde{\boldsymbol{\theta}}$, based on the statistics at its neighborhood. Similar to the local χ-unbiasedness in Definition 1, we can define the *local* missing-mass unbiasedness as follows.

Definition 3: An estimator $\hat{\boldsymbol{\theta}} : \mathcal{S}^N \rightarrow \Omega_{\boldsymbol{\theta}}$ is said to be a *locally* Lehmann-unbiased estimator [34] in the neighborhood of $\tilde{\boldsymbol{\theta}} \in \Omega_{\boldsymbol{\theta}}$ w.r.t. the missing-mass squared-error cost function from (27) if it satisfies

$$\mathbf{U}^T\mathbf{b}_{N,0}(\tilde{\boldsymbol{\theta}}) = \mathbf{0}_{M-1} \quad (32)$$

and

$$\left\{\nabla_{\boldsymbol{\theta}}^T\mathbf{b}_{N,0}(\boldsymbol{\theta})\right\}\big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}\mathbf{U} = \mathbf{0}_{M\times(M-1)}. \quad (33)$$

It should be noted that the condition in (31) requires a *uniform* unbiasedness, for any $\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}$, while the conditions in (32) and (33) are *local* conditions that are required to be satisfied only at the specific $\boldsymbol{\theta}$, denoted here by $\tilde{\boldsymbol{\theta}}$, for which the bound is developed. Both the local and uniform missing-mass unbiasedness definitions restrict only the values that belong to the set of unseen symbols, i.e. elements that belong to the set $G_{N,0}(\mathbf{x})$ from (6), in $\mathcal{S}$ to be unbiased. In addition, by comparing Definition 1 and Definition 3, it can be seen that the differences between the local χ-unbiasedness and the local missing-mass unbiasedness follows from the difference of the cost functions, where both definitions use the null-space matrix, $\mathbf{U}$, which is due to the parametric constraint.

The uniform missing-mass unbiasedness from (31) can be interpreted as follows. From the definition of $\mathbf{U}$ in (14), it can be verified that $\mathbf{U}^T\mathbf{y} = \mathbf{0}_{M-1}$ *iff* $\mathbf{y} = c\mathbf{1}_M$, where $c \in \mathbb{R}$ is an arbitrary constant. Thus, the condition in (31) implies that for a uniformly Lehmann unbiased estimator, the missing-mass bias vector satisfies

$$\mathbf{b}_{N,0}(\boldsymbol{\theta}) = \beta_{N,0}(\boldsymbol{\theta})\mathbf{1}_M,\ \forall\boldsymbol{\theta}\in\Omega_{\boldsymbol{\theta}}, \quad (34)$$

where $\beta_{N,0}(\boldsymbol{\theta}) \in \mathbb{R}$ is a constant. The condition in (34) is that the missing-mass mth bias is identical for any m. This property recalls the notion of natural estimators [19], since it assigns the same bias requirements to all symbols appearing with the same probability.

B. mmCCRB

Lower bounds on the mmMSE are useful for performance analysis and system design. In this subsection, a constrained Cramér-Rao-type lower bound on the mmMSE from (28) is derived. The new bound is based on the missing-mass bias in the Lehmann sense, as defined in Subsection III-A.

Let us define the following *missing-mass Fisher information matrix (mmFIM)* :

$$\mathbf{J}^{(0)}(\boldsymbol{\theta}) \triangleq \mathrm{E}_{\boldsymbol{\theta}}\left[\boldsymbol{\Delta}(\mathbf{x},\boldsymbol{\theta})\boldsymbol{\Delta}^T(\mathbf{x},\boldsymbol{\theta})\right], \quad (35)$$

where the mth column of the matrix $\boldsymbol{\Delta}(\mathbf{x},\boldsymbol{\theta}) \in \mathbb{R}^{M\times M}$ is defined as

$$\boldsymbol{\Delta}_{1:M,m}(\mathbf{x},\boldsymbol{\theta}) \triangleq \nabla_{\boldsymbol{\theta}}\log p(\mathbf{x}|\mathbf{x}\in\mathcal{A}_m;\boldsymbol{\theta})\mathbf{1}_{\{\mathbf{x}\in\mathcal{A}_m\}}, \quad (36)$$

$m = 1,\ldots,M$. In addition, we define the auxiliary matrix $\mathbf{S}(\boldsymbol{\theta}) \in \mathbb{R}^{M\times M}$, in which the mth row is defined as

$$\begin{aligned}&\mathbf{S}_{m,1:M}(\boldsymbol{\theta})\\ &\triangleq \left(\nabla_{\boldsymbol{\theta}}^T\left\{\frac{[\mathbf{b}_{N,0}(\boldsymbol{\theta})]_m}{\Pr(\mathbf{x}\in\mathcal{A}_m;\boldsymbol{\theta})}\right\} + \mathbf{e}_m^T\right)\Pr(\mathbf{x}\in\mathcal{A}_m;\boldsymbol{\theta})\\ &= \nabla_{\boldsymbol{\theta}}^T\left\{\mathrm{E}_{\boldsymbol{\theta}}\left[\hat{\theta}_m|\mathbf{x}\in\mathcal{A}_m\right]\right\}\Pr(\mathbf{x}\in\mathcal{A}_m;\boldsymbol{\theta}), \end{aligned} \quad (37)$$

where the last equality is obtained by substituting (30) and using $\nabla_{\boldsymbol{\theta}}^T\theta_m = \mathbf{e}_m$. The auxiliary matrix, $\mathbf{S}(\boldsymbol{\theta})$, involves the gradient of the missing-mass bias, which makes it intractable for many estimators. The following lemma presents a tractable form of the auxiliary matrix.

Lemma 1: The mth row of the auxiliary matrix $\mathbf{S}(\boldsymbol{\theta})$ from (37) can be calculated as

$$\mathbf{S}_{m,1:M}(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}\left[\hat{\theta}_m(\mathbf{x})\mathbf{v}^T(\mathbf{x},\boldsymbol{\theta})\mathbf{1}_{\{\mathbf{x}\in\mathcal{A}_m\}}\right] - \frac{N}{1-\theta_m}\mathbf{e}_m^T\mathrm{E}_{\boldsymbol{\theta}}\left[\hat{\theta}_m(\mathbf{x})\mathbf{1}_{\{\mathbf{x}\in\mathcal{A}_m\}}\right], \quad (38)$$

where

$$\mathbf{v}(\mathbf{x},\boldsymbol{\theta}) \triangleq \left[\frac{C_{N,1}(\mathbf{x})}{\theta_1},\ldots,\frac{C_{N,M}(\mathbf{x})}{\theta_M}\right]^T. \quad (39)$$

Proof: The proof appears in Appendix B. ∎

We define the following regularity condition:

C.1) The likelihood gradient vector, $\boldsymbol{\Delta}_{1:M,m}(\mathbf{x},\boldsymbol{\theta})$, defined in (36), exists and is finite $\forall\boldsymbol{\theta}\in\Omega_{\boldsymbol{\theta}}$ and $\forall m = 1,\ldots,M$. That is, the matrix, $\mathbf{U}^T\mathbf{J}^{(0)}(\boldsymbol{\theta})\mathbf{U}$, is a well-defined, non-singular, and non-zero matrix for any $\boldsymbol{\theta}\in\Omega_{\boldsymbol{\theta}}$.

Theorem 1: Let the regularity condition C.1 be satisfied and $\hat{\boldsymbol{\theta}}$ be an estimator of $\boldsymbol{\theta}\in\Omega_{\boldsymbol{\theta}}$ with a local missing-mass bias vector in the neighborhood of $\tilde{\boldsymbol{\theta}}\in\Omega_{\boldsymbol{\theta}}$ given by $\mathbf{b}_{N,0}(\tilde{\boldsymbol{\theta}})$, as defined in (30). Then, the mmMSE from (28) satisfies

$$\mathrm{E}_{\tilde{\boldsymbol{\theta}}}\left[C(\hat{\boldsymbol{\theta}},\tilde{\boldsymbol{\theta}})\right] \geq B^{\mathrm{mmCCRB}}(\tilde{\boldsymbol{\theta}}), \quad (40)$$

where the mmCCRB evaluated at the local point, $\tilde{\boldsymbol{\theta}}$, is

$$B^{\mathrm{mmCCRB}}(\tilde{\boldsymbol{\theta}}) \triangleq \mathrm{trace}\left(\mathbf{S}^T(\tilde{\boldsymbol{\theta}})\mathbf{U}(\mathbf{U}^T\mathbf{J}^{(0)}(\tilde{\boldsymbol{\theta}})\mathbf{U})^{-1}\mathbf{U}^T\mathbf{S}(\tilde{\boldsymbol{\theta}})\right) + \sum_{m=1}^{M}\frac{\left[\mathbf{b}_{N,0}(\tilde{\boldsymbol{\theta}})\right]_m^2}{\Pr(\mathbf{x}\in\mathcal{A}_m;\tilde{\boldsymbol{\theta}})}. \quad (41)$$

Moreover, equality is achieved in (41) if

$$\hat{\theta}_m - \tilde{\theta}_m = \frac{[b_{N,0}(\tilde{\boldsymbol{\theta}})]_m}{\Pr(\mathbf{x} \in \mathcal{A}_m; \tilde{\boldsymbol{\theta}})} + \left[\mathbf{S}^T(\tilde{\boldsymbol{\theta}})\mathbf{U}(\mathbf{U}^T\mathbf{J}^{(0)}(\tilde{\boldsymbol{\theta}})\mathbf{U})^{-1}\mathbf{U}^T\boldsymbol{\Delta}(\mathbf{x}, \boldsymbol{\theta})\right]_{m,m}, \quad (42)$$

$\forall m = 1, \ldots, M$, such that $m \in G_{N,0}(\mathbf{x})$.

Proof: The proof appears in Appendix C. ∎

It can be seen that the requirement for the equality condition in (42) for the achievability of the mmCCRB only determines the values of the missing-mass estimation errors. In addition, it can be seen that the estimator defined in (42) assigns, in the general case, a different value for each element in the missing mass. That is, it is not necessarily a natural estimator [19], in the sense that elements that appeared the same number of times will not necessarily get the same estimated probability. Moreover, the estimator defined by (42) is a function of the (local) unknown parameter vector, $\tilde{\boldsymbol{\theta}}$, in the general case. Only if it is independent of $\tilde{\boldsymbol{\theta}}$, then it is an efficient estimator with mmMSE equals to the mmCCRB.

Similar to the Fisher scoring method, this equality condition can be used to obtain an iterative estimation procedure. In this case, the estimator at the kth iteration, $\hat{\boldsymbol{\theta}}^{(k)}$, is obtained by substituting the estimator from the previous iteration, $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^{(k-1)}$, in (42) to obtain

$$\hat{\theta}_m^{(k)} - \hat{\theta}_m^{(k-1)} = \psi \left\{ \frac{[b_{N,0}(\tilde{\boldsymbol{\theta}})]_m}{\Pr(\mathbf{x} \in \mathcal{A}_m; \tilde{\boldsymbol{\theta}})} + \left[\mathbf{S}^T(\tilde{\boldsymbol{\theta}})\mathbf{U}(\mathbf{U}^T\mathbf{J}^{(0)}(\tilde{\boldsymbol{\theta}})\mathbf{U})^{-1}\mathbf{U}^T\boldsymbol{\Delta}(\mathbf{x}, \tilde{\boldsymbol{\theta}})\right]_{m,m} \right\} \Bigg|_{\tilde{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}^{(k-1)}}, \quad (43)$$

$\forall m = 1, \ldots, M$, such that $m \in G_{N,0}(\mathbf{x})$, $m \geq 1$, where ψ is the step size. The initial estimator, $\hat{\boldsymbol{\theta}}^{(0)}$, can be chosen to be any existing estimator, such as the CML, Good-Turing, or add-constant estimators, described in Subsections III-D2, III-D3, and III-D4, respectively.

C. mmCCRB for the i.i.d. model

The mmCCRB in Theorem 1 is a lower bound on the mmMSE from (28), which has been developed for the general observation model, $p(\mathbf{x}; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}$. In this subsection we develop the closed-form expression of the mmCCRB for the i.i.d. model, as described by (3). The following Lemma describes the closed-form mmFIM for this case.

Lemma 2: The mmFIM from (35) for the model described in Subsection II-A with the observation pmf in (3) is:

$$\mathbf{J}^{(0)}(\boldsymbol{\theta}) = \sum_{m=1}^{M} \frac{N(N-1)}{(1-\theta_m)^2}(1-\theta_m)^N \mathbf{1}\mathbf{1}^T + \sum_{m=1}^{M} \frac{N}{(1-\theta_m)^2}\left(\mathbf{e}_m\mathbf{1}^T + \mathbf{1}\mathbf{e}_m^T\right) + N\mathbf{D}(\boldsymbol{\theta}), \quad (44)$$

where $\mathbf{D}(\boldsymbol{\theta})$ is a $M \times M$ diagonal matrix with the following elements on its diagonal:

$$[\mathbf{D}(\boldsymbol{\theta})]_{m,m} = \sum_{l=1, l\neq m}^{M} \frac{1}{\theta_l(1-\theta_m)} - \frac{1}{(1-\theta_m)^2}, \quad (45)$$

$m = 1, \ldots, M$.

Proof: The proof appears in Appendix D. ∎

It should be noted that by using (44) and the null-space property of the matrix $\mathbf{U}$ from (14), $\mathbf{1}_M^T\mathbf{U} = \mathbf{0}^T$, we obtain

$$\mathbf{U}^T\mathbf{J}^{(0)}(\boldsymbol{\theta})\mathbf{U} = N\mathbf{U}^T\mathbf{D}(\boldsymbol{\theta})\mathbf{U}. \quad (46)$$

By substituting (46) in (41), we obtain the mmCCRB for the classical model:

$$B^{\text{mmCCRB}}(\tilde{\boldsymbol{\theta}}) = \frac{1}{N}\text{trace}\left(\mathbf{S}^T(\tilde{\boldsymbol{\theta}})\mathbf{U}(\mathbf{U}^T\mathbf{D}(\tilde{\boldsymbol{\theta}})\mathbf{U})^{-1}\mathbf{U}^T\mathbf{S}(\tilde{\boldsymbol{\theta}})\right) + \sum_{m=1}^{M} \frac{\left[\mathbf{b}_{N,0}(\tilde{\boldsymbol{\theta}})\right]_m^2}{\Pr(\mathbf{x} \in \mathcal{A}_m; \tilde{\boldsymbol{\theta}})}. \quad (47)$$

By substituting (46) in (42), we obtain the equality condition of the mmCCRB for the classical model.

D. Special cases

In this subsection we develop some important special cases of the mmCCRB for the i.i.d. model from Subsection III-C.

1) mmCCRB on the mmMSE of missing-mass unbiased estimators: For the sake of simplicity of derivation, we assume in this subsection that $\mathbf{b}_{N,0}(\boldsymbol{\theta}) = \mathbf{0}$. According to Proposition 1, this condition is a *sufficient* condition for the Lehmann unbiasedness in (32) and (16). For this case and the classical model described by (3), it can be verified that $\mathbf{S}(\boldsymbol{\theta})$ from (38) is a diagonal matrix with the diagonal elements

$$[\mathbf{S}(\boldsymbol{\theta})]_{m,m} = \Pr(\mathbf{x} \in \mathcal{A}_m; \boldsymbol{\theta}), \ m = 1, \ldots, M. \quad (48)$$

By substituting $\mathbf{b}_{N,0}(\boldsymbol{\theta}) = \mathbf{0}$ and (48) in (47), we obtain that the mmCCRB on the mmMSE of a missing-mass unbiased estimator is

$$B^{\text{mmCCRB}}(\tilde{\boldsymbol{\theta}}) = \frac{1}{N}\text{trace}\left(\mathbf{U}^T\mathbf{P}^T(\tilde{\boldsymbol{\theta}})\mathbf{U}(\mathbf{U}^T\mathbf{D}(\tilde{\boldsymbol{\theta}})\mathbf{U})^{-1}\right), \quad (49)$$

where $\mathbf{P}(\boldsymbol{\theta})$ is a diagonal $M \times M$ matrix with the following elements on its diagonal:

$$[\mathbf{P}(\boldsymbol{\theta})]_{m,m} \stackrel{\triangle}{=} \left(\Pr(\mathbf{x} \in \mathcal{A}_m; \boldsymbol{\theta})\right)^2, \ m = 1, \ldots, M. \quad (50)$$

By substituting (46) and (50) in (49), and using the trace operator properties, one obtains that for this case

$$B^{\text{mmCCRB}}(\tilde{\boldsymbol{\theta}}) = \frac{1}{N}\sum_{m=1}^{M}(1-\tilde{\theta}_m)^{2N}\left[\mathbf{U}(\mathbf{U}^T\mathbf{D}(\tilde{\boldsymbol{\theta}})\mathbf{U})^{-1}\mathbf{U}^T\right]_{m,m}. \quad (51)$$

2) mmCCRB on the mmMSE of the CML estimator: The CML estimator from (11) assigns a zero probability to unseen events, that is

$$\hat{\theta}_m^{\text{CML}} = 0, \ \forall m \in G_{N,0}(\mathbf{x}). \quad (52)$$

By substituting (52) in (30), one obtains that the missing-mass bias of the CML estimator satisfies

$$[\mathbf{b}_{N,0}^{\text{CML}}(\boldsymbol{\theta})]_m = \mathrm{E}_{\boldsymbol{\theta}}[\hat{\theta}_m^{\text{CML}} - \theta_m | \mathbf{x} \in \mathcal{A}_m] \Pr(\mathbf{x} \in \mathcal{A}_m; \boldsymbol{\theta}) = -\theta_m \Pr(\mathbf{x} \in \mathcal{A}_m; \boldsymbol{\theta}), \quad (53)$$

for any $m = 1, \ldots, M$. By substituting (53) in (37), we obtain that for the CML estimator, the auxiliary matrix satisfies $\mathbf{S}(\boldsymbol{\theta}) = \mathbf{0}_{M\times M}$. By substituting this result and (53) in (47),

we obtain that the mmCCRB on the mmMSE of the CML estimator (or any other estimator with the same bias function as in (53)) is

$$B^{\text{mmCCRB}}(\tilde{\boldsymbol{\theta}}) = \sum_{m=1}^{M} \tilde{\theta}_m^2 \Pr(\mathbf{x} \in \mathcal{A}_m; \boldsymbol{\theta}). \tag{54}$$

On the other hand, by substituting (52) in (28), it can be seen that the mmMSE of the CML estimator evaluated at the local point, $\tilde{\boldsymbol{\theta}}$, is

$$\mathrm{E}_{\tilde{\boldsymbol{\theta}}}\left[C(\hat{\boldsymbol{\theta}}^{\text{CML}}, \tilde{\boldsymbol{\theta}})\right] = \sum_{m=1}^{M} \theta_m^2 \Pr(\mathbf{x} \in \mathcal{A}_m; \tilde{\boldsymbol{\theta}}). \tag{55}$$

Thus, in this case, the proposed mmCCRB coincides with the mmMSE of the CML estimator. Therefore, we can conclude that there is no other estimator with the same missing-mass bias as that of the CML estimator, given in (53), that achieves a lower mmMSE than the CML estimator. It should be noted that since we used the mmCCRB from Theorem 1, which was developed for the general observation model, $p(\mathbf{x};\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}$, this result also holds for a non-i.i.d. sampling with a general structure of $p(\mathbf{x};\boldsymbol{\theta})$ [24, 40, 41].

3) mmCCRB on the mmMSE of Good-Turing estimator: The Good-Turing estimator [2] of the missing mass from (26) is defined as the fraction of symbols occurring exactly once in the observed samples divided by the length of the observation vector. It is well known that smoothing of the Good-Turing estimator is needed in order to obtain reasonable results [2, 8]. Here we use a smooth modified version of the Good-Turing estimator, proposed in [8]. This modified Good-Turing estimator is given by

$$p_0(\mathbf{x}, \hat{\boldsymbol{\theta}}) = \sum_{m=1}^{M} \hat{\theta}_m^{\text{GT}} \mathbf{1}_{\{m \in G_{N,0}(\mathbf{x})\}} = \frac{\varphi(|G_{N,0}(\mathbf{x})|)}{N'}, \tag{56}$$

where $\varphi(t) = \max\{t, 1\},\ t \geq 0$, $|G_{N,0}(\mathbf{x})|$ is the number of elements that appear exactly once in the N-length observation vector, $\mathbf{x}$, and N' is a normalization factor. Thus, the Good-Turing estimator of a specific element in $G_{N,0}(\mathbf{x})$ is [44]

$$\hat{\theta}_m^{\text{GT}} = \frac{\varphi(|G_{N,0}(\mathbf{x})|)}{N'|G_{N,0}(\mathbf{x})|},\ m \in G_{N,0}(\mathbf{x}), \tag{57}$$

for any $\mathbf{x}$ such that $G_{N,0}(\mathbf{x}) \neq \emptyset$. By substituting (57) in (30), one obtains that the missing-mass bias of the Good-Turing estimator satisfies

$$\left[\mathbf{b}_{N,0}^{\text{GT}}(\boldsymbol{\theta})\right]_m = \mathrm{E}_{\boldsymbol{\theta}}\left[\frac{\varphi(|G_{N,0}(\mathbf{x})|)}{N'|G_{N,0}(\mathbf{x})|} - \theta_m | \mathbf{x} \in \mathcal{A}_m\right] \Pr(\mathbf{x} \in \mathcal{A}_m; \boldsymbol{\theta}). \tag{58}$$

While a closed-form expression of the missing-mass bias of the Good-Turing estimator in (58) is intractable, the proposed bound can be used by calculating the auxiliary matrix for this case. In particular, by substituting (57) in (38), one obtains

$$\begin{aligned}&\mathbf{S}_{m,1:M}(\boldsymbol{\theta}) \\ &= \mathrm{E}_{\boldsymbol{\theta}}\left[\frac{\varphi(|G_{N,0}(\mathbf{x})|)}{N'|G_{N,0}(\mathbf{x})|}\mathbf{v}^T(\mathbf{x},\boldsymbol{\theta})\mathbf{1}_{\{m \in G_{N,0}(\mathbf{x})\}}\right] \\ &- \sum_{m=1}^{M} \frac{1}{1-\theta_m}\mathbf{e}_m^T \mathrm{E}_{\boldsymbol{\theta}}\left[\frac{\varphi(|G_{N,0}(\mathbf{x})|)}{|G_{N,0}(\mathbf{x})|}\mathbf{1}_{\{m \in G_{N,0}(\mathbf{x})\}}\right].\end{aligned} \tag{59}$$

Then, by substituting (59) in (47), we obtain the associated mmCCRB, which can be approximated by low-complexity methods. In particular, a Monte Carlo approach can be developed to approximate the expectation in (59), in a similar manner to the empirical FIM approximation described in [45].

4) mmCCRB on the mmMSE of add-constant estimator: The add-constant estimator of the missing mass from (26) is defined as [11]

$$p_0(\mathbf{x}, \hat{\boldsymbol{\theta}}) = \frac{c}{N + c(M - |G_{N,0}(\mathbf{x})| + 1)}, \tag{60}$$

for a positive constant c. The add-constant estimator has been applied and studied extensively and has been shown to have some optimality properties [8]. For $c = 1$, we obtain the special case of the Laplace estimator [10]. By dividing (60) by $|G_{N,0}(\mathbf{x})|$, under the assumption that $G_{N,0}(\mathbf{x}) \neq \emptyset$, the add-constant estimator of a specific element in $G_{N,0}(\mathbf{x})$ is [44]

$$\hat{\theta}_m^{\text{add-c}} = \frac{c}{|G_{N,0}(\mathbf{x})|(N + c(M - |G_{N,0}(\mathbf{x})| + 1))}, \tag{61}$$

$\forall m \in G_{N,0}(\mathbf{x})$. By substituting (61) in (30), one obtains the missing-mass bias of the add-constant estimator. While a closed-form expression of the missing-mass bias of add-constant estimator is intractable, the proposed mmCCRB can be used by calculating the auxiliary matrix for this case. That is, similar to the derivation of (59), we can substitute (61) in (38) and then substitute the result in (47), to obtain the mmCCRB for this case. Then, the mmCCRB can be approximated by low-complexity methods.

5) Uniform distribution: For the special case where $\boldsymbol{\theta} = \frac{1}{M}\mathbf{1}_M$, the diagonal elements of the matrix $\mathbf{D}(\boldsymbol{\theta})$ from (45) are given by

$$[\mathbf{D}(\boldsymbol{\theta})]_{m,m} = -\frac{1}{(1-\frac{1}{M})^2} + \frac{M-1}{\frac{1}{M}(1-\frac{1}{M})} = \frac{M^4 - 2M^3}{(M-1)^2}, \tag{62}$$

$m = 1, \ldots, M$. Similarly, for this case (50) is reduced to

$$\mathbf{P}(\boldsymbol{\theta}) = \left(\frac{M-1}{M}\right)^{2N} \mathbf{I}_M. \tag{63}$$

By substituting (14), (62), and (63) in (51), it can be verified that for this case the mmCCRB missing-mass unbiased estimator is given by

$$B^{\text{mmCCRB}}(\boldsymbol{\theta}) = \frac{1}{N}\left(\frac{M-1}{M}\right)^{2N} \frac{(M-1)^3}{M^4 - 2M^3}, \tag{64}$$

where we used (14) and the cyclic property of the trace. We can see that the mmCCRB for the uniform pmf from (64) decreases as the number of samples, N, increases, since we have more information. It can be verified that this bound increases as M increases if $N+2-\sqrt{N^2+2} \leq M \leq N+2+\sqrt{N^2+2}$, and decreases as M increases otherwise.

On the other hand, for this case of uniform pmf the CCRB in (25) on the trace MSE is reduced to

$$B^{\text{CCRB}}(\boldsymbol{\theta}) = \frac{1}{N}. \tag{65}$$

Thus, the CCRB is independent of the number of elements, M, in contrast to the missing-mass CCRB in (64), and, thus, is less informative for the problem of missing-mass estimation.

IV. Simulation

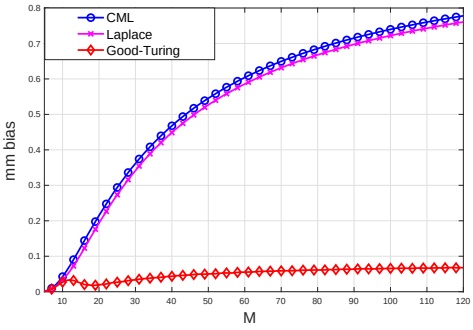
In this section, we evaluate the following bounds:

- The CCRB from (25), which is a lower bound on the MSE and is presented here in order to compare its asymptotic behaviour with the proposed mmMSE bounds.
- Three versions of the biased mmCCRB from (47): 1) with the bias of the CML estimator, as given in (54); 2) with the empirical bias of the Good-Turing estimator, as described in Subsection III-D3; and 3) with the empirical bias of the Laplace estimator, as described in Subsection III-D4, with the constant $c = 1$.
- The mmCCRB on missing-mass unbiased estimators from (51).

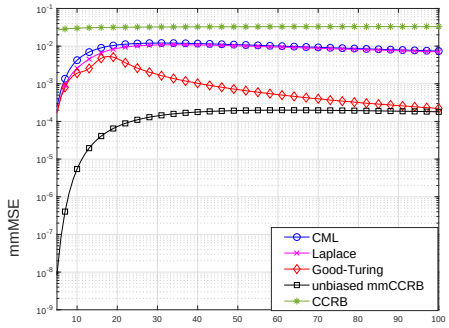
It can be verified that in all the simulations the regularity condition C.1 is satisfied. We compare these bounds with the performance of the following estimators of the missing mass: 1) the CML estimator, as described in (52); 2) the Good-Turing estimator of the missing mass from (57); and 3) the Laplace estimator [10, 11], which is the add-constant estimator from (61) with $c = 1$. The performance of these estimators is evaluated using $10,000$ Monte-Carlo simulations that are used to evaluate the mmMSE, $\sum_{m=1}^{M} \mathrm{E}_{\boldsymbol{\theta}}\left[(\hat{\theta}_m - \theta_m)^2 \mathbf{1}_{\{m \in G_{N,0}(\mathbf{x})\}}\right]$, and the missing-mass total bias, $\sum_{m=1}^{M} \left[\mathbf{b}_{N,0}(\boldsymbol{\theta})\right]_m$, as defined in (28) and (30), respectively.

A. Example 1: Uniform distribution

In the first experiment we examine the case of a uniform pmf with equally-likely elements, i.e. where $\boldsymbol{\theta} = \frac{1}{M}\mathbf{1}_M$, as described in Subsection III-D5. The performance is evaluated for different values of M and for $N = 30$. In Figs. 1a and 1b we present the missing-mass bias and the mmMSE, respectively, of the different estimators versus the number of elements, M, for $N = 30$. The CCRB and the mmCCRB of unbiased estimators are also presented in Fig. 1b. It can be seen that for this case, the Good-Turing estimator outperforms the two other estimators in both missing-mass bias and mmMSE terms, and that the differences between the performance of the CML and the Laplace estimators is insignificant. In addition, it can be seen in Fig. 1b that the mmmSE of the Good-Turing estimator asymptotically achieves the proposed mmCCRB on the mmMSE of *unbiased* estimators.



(a)



(b)

Fig. 1: Example 1 (uniform pmf): The performance of the CML, Good-Turing, and Laplace estimators versus the number of elements, M, in terms of missing-mass bias (a) and the mmMSE (b). In (b) we also present the CCRB and the proposed mmCCRB on missing-mass *unbiased* estimators.

B. Example 2: Zipf distribution

In the second experiment, we consider a Zipf's law distribution, $\theta_m = \frac{m^{-s}}{\sum_{k=1}^{M} k^{-s}}$ for all $m = 1, \ldots, M$, where s is the skewness parameter and M is the alphabet size. The Zipf's law distribution is a commonly-used heavy-tailed distribution that is widely used in physical and social sciences, linguistics, economics, and other fields. The bias and mmMSE of the CML, Good-Turing, and Laplace estimators are presented in Figs. 2a and 2b, respectively, versus the number of samples, N, for $M = 15$. In addition, in Fig. 2b we also present the CCRB and the mmCCRB for unbiased estimators. It can be seen that the CML estimator has the largest missing-mass bias and the largest mmMSE. Additionally, the Good-Turing estimator has the smallest missing-mass bias for all N. In Fig. 2b we can see that the mmCCRB on the mmMSE of missing-mass unbiased estimators is much lower than the actual mmMSE of the estimators. Since these estimators are *biased* in the Lehmann sense, in Fig. 2b, we compare the mmMSE of the estimators with the associated *biased* mmCCRB associated with the estimators. It can be seen in the leftest figure in Fig. 2b that the biased mmCCRB with the CML bias coincides with the mmMSE of the CML estimator, as shown analytically in (54)-(55). Similarly, the biased mmCCRB associated with the Laplace estimator coincides with the mmMSE of the Laplace estimator. However, for the Good-Turing estimator in the rightest figure of Fig. 2b, there is a gap between the associated mmCCRB and the mmMSE. As the sample size, N, increases, the mmMSE of the Good-Turing estimator achieves the mmCCRB. In Figs. 3a and 3b, we compare the missing-mass bias and the mmMSE of the Laplace estimator and the proposed Fisher-scoring estimators that are obtained after $1-5$ iterations, as appear in (43) with $\psi = \frac{1}{N}$, and are initialized by the Laplace estimator. Thus, these estimators are based on the proposed mmCCRB, which is evaluated numerically, as described in Subsection III-D4. It can be seen that the proposed Fisher-scoring method improves the performance the missing-mass bias and the mmMSE of the Laplace estimator. In addition, the proposed method is consistent in the sense that by using more iterations we obtain better estimators. Thus, the proposed mmCCRB is also a way to improve existing estimators of the missing mass.

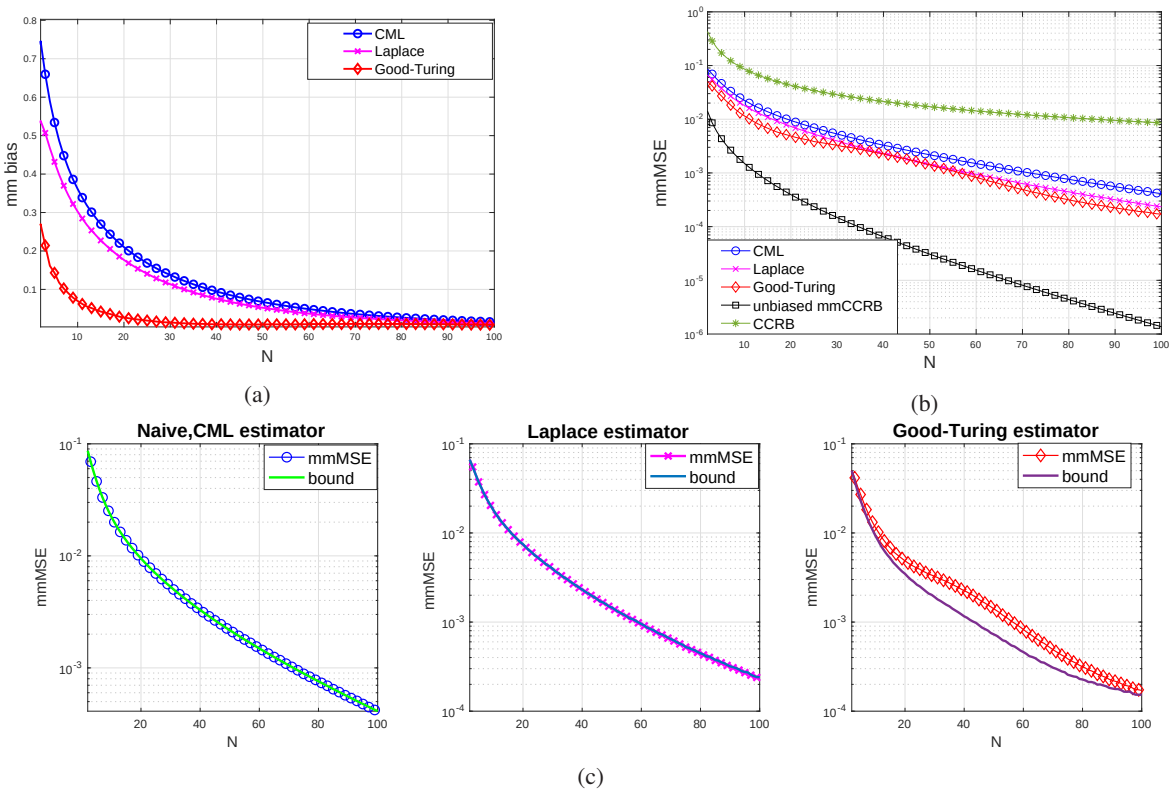


Fig. 2: Example 2 (Zipf distribution): The performance of the CML, Good-Turing, and Laplace estimators versus the number of samples, N, in terms of missing-mass bias (a) and the mmMSE compared with the CCRB and the mmCCRBs (b), (c).

V. Conclusion

In this paper, we consider the problem of estimation of the missing mass. First, we develop the CCRB for this problem, which provides a lower bound on the MSE of any χ-unbiased estimator. However, similar to the naive CML estimator, which overestimates the probability of the observed elements, the CCRB does not provide a relevant bound for the missing-mass estimation problem. Hence, we define the mmMSE risk function that only penalizes the estimation errors of elements that belong to the missing mass. A novel unbiasedness restriction, denoted by missing-mass unbiasedness, is proposed. The missing-mass unbiasedness is based on Lehmann's concept of unbiasedness that takes into account the chosen cost function and the relevant parameter space. We develop a new CRB-type bound for this problem, the mmCCRB, which is a lower bound on the mmMSE of locally missing-mass unbiased estimators. In addition, the biased mmCCRB is developed. By using the mmCCRB on the mmMSE of the CML estimator, we show that the CML estimator has the smallest mmMSE among all estimators that have the same missing-mass bias as the CML estimator. In the simulations, we show that the Good-Turing estimator has the lowest mmMSE when the distribution is uniform and the number of elements is large.

Appendix A
Proof of Proposition (1)

In this appendix, we develop the missing-mass Lehmann unbiasedness. By substituting the missing-mass squared-error cost function from (27) and $\Omega_{\boldsymbol{\theta}}$ from (1) in (29), one obtains that the Lehmann-unbiasedness condition for the missing-mass estimation problem is given by

$$\sum_{m=1}^{M} \mathrm{E}_{\boldsymbol{\theta}}\left[(\hat{\theta}_m-\eta_m)^2\mathbf{1}_{\{m\in G_{N,0}(\mathbf{x})\}}\right] \geq \sum_{m=1}^{M} \mathrm{E}_{\boldsymbol{\theta}}\left[(\hat{\theta}_m-\theta_m)^2\mathbf{1}_{\{m\in G_{N,0}(\mathbf{x})\}}\right], \quad (66)$$

$\forall \boldsymbol{\theta},\boldsymbol{\eta}\in\Omega_{\boldsymbol{\theta}}$. By using the definition of the constrained set in (1) and since $\mathbf{U}$ from (14) is the null-space matrix of this constrained set, then, for a given $\boldsymbol{\theta}\in\Omega_{\boldsymbol{\theta}}$, any $\boldsymbol{\eta}\in\Omega_{\boldsymbol{\theta}}$ can be written as (see, e.g. Section 4.2.4 in [46])

$$\boldsymbol{\eta}=\boldsymbol{\theta}+\mathbf{U}\mathbf{w}, \quad (67)$$

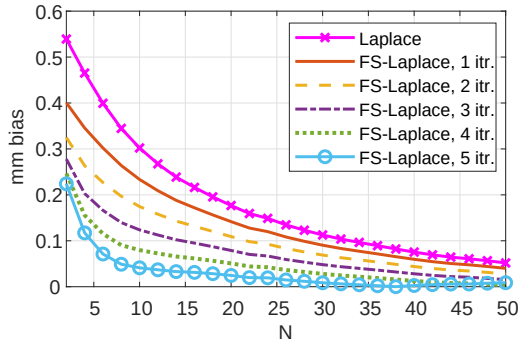
where $\mathbf{w}\in\mathbb{R}^{M-1}$ is an arbitrary vector. By substituting (67) in (66), we obtain

$$\sum_{m=1}^{M} \mathrm{E}_{\boldsymbol{\theta}}\left[(\hat{\theta}_m-\theta_m-\mathbf{e}_m^T\mathbf{U}\mathbf{w}))^2\mathbf{1}_{\{m\in G_{N,0}(\mathbf{x})\}}\right] \geq \sum_{m=1}^{M} \mathrm{E}_{\boldsymbol{\theta}}\left[(\hat{\theta}_m-\theta_m)^2\mathbf{1}_{\{m\in G_{N,0}(\mathbf{x})\}}\right], \quad (68)$$

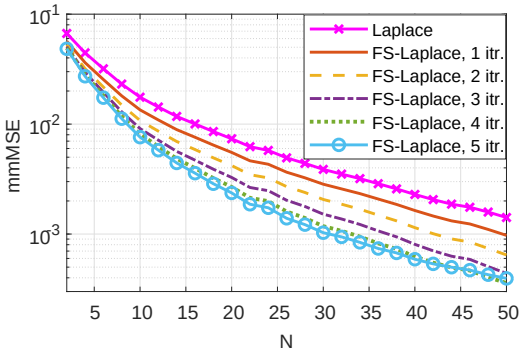
$\forall\boldsymbol{\theta}\in\Omega_{\boldsymbol{\theta}},\mathbf{w}\in\mathbb{R}^{M-1}$. By using (8), the unbiasedness condition from (68) can be rewritten as:

$$\sum_{m=1}^{M}(\mathbf{e}_m^T\mathbf{U}\mathbf{w})^2\Pr(\mathbf{x}\in\mathcal{A}_m;\boldsymbol{\theta}) \geq 2\sum_{m=1}^{M}\mathrm{E}_{\boldsymbol{\theta}}\left[(\hat{\theta}_m-\mathbf{e}_m^T\boldsymbol{\theta})\mathbf{1}_{\{m\in G_{N,0}(\mathbf{x})\}}\right]\mathbf{e}_m^T\mathbf{U}\mathbf{w}, \quad (69)$$

$\forall\boldsymbol{\theta}\in\Omega_{\boldsymbol{\theta}},\mathbf{w}\in\mathbb{R}^{M-1}$. Since the condition in (69) should be satisfied for any $\mathbf{w}\in\mathbb{R}^{M-1}$, it should be satisfied in particular for both $\mathbf{w}=\epsilon\mathbf{e}_k$ and $\mathbf{w}=-\epsilon\mathbf{e}_k$, where $\epsilon>0$. By summing the separate substitution of $\mathbf{w}=\pm\epsilon\mathbf{e}_k$ (that is, the result of



(a)



(b)

Fig. 3: Example 2 (Zipf distribution): The performance of Laplace estimator and the proposed Fisher-scoring estimators after 1-5 iterations versus the number of samples, N, in terms of missing-mass bias (a) and the mmMSE (b) compared with the proposed mmCCRB.

substituting $\mathbf{w} = \epsilon \mathbf{e}_k$ into (69) and the result of substituting $\mathbf{w} = \epsilon \mathbf{e}_k$ into the same equation) we obtain the following *necessary* condition for (69) to hold:

$$\sum_{m=1}^{M} \mathrm{E}_{\boldsymbol{\theta}} \left[(\hat{\theta}_m - \theta_m) \mathbf{1}_{\{m \in G_{N,0}(\mathbf{x})\}} \right] \mathbf{e}_m^T \mathbf{U} = \mathbf{0}_{M-1}^T, \quad (70)$$

$\forall \boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}$. Since the l.h.s. of (69) is a quadratic term, it can be verified that (70) is also a *sufficient* condition for unbiasedness in this case. Therefore, by applying the transpose operator on (70), one obtains that the missing-mass unbiasedness in (31) is the Lehmann unbiasedness under the missing-mass squared error cost function.

Appendix B
Proof of Lemma 1

By using the definition of conditional expectation and (10), we obtain that

$$\begin{aligned}
&\nabla_{\boldsymbol{\theta}}^T \left\{ \mathrm{E}_{\boldsymbol{\theta}} \left[\hat{\theta}_m | \mathbf{x} \in \mathcal{A}_m \right] \right\} \\
&= \nabla_{\boldsymbol{\theta}}^T \left\{ \sum\nolimits_{\boldsymbol{\alpha} \in \mathcal{A}_m} \hat{\theta}_m(\boldsymbol{\alpha}) \Pr\left(\mathbf{x} = \boldsymbol{\alpha} | \mathbf{x} \in \mathcal{A}_m; \boldsymbol{\theta} \right) \right\} \\
&= \nabla_{\boldsymbol{\theta}}^T \left\{ \sum\nolimits_{\boldsymbol{\alpha} \in \mathcal{A}_m} \hat{\theta}_m(\boldsymbol{\alpha}) \frac{\prod_{l=1}^{M} \theta_l^{C_{N,l}(\boldsymbol{\alpha})}}{(1-\theta_m)^N} \right\} \\
&= \sum\nolimits_{\boldsymbol{\alpha} \in \mathcal{A}_m} \hat{\theta}_m(\boldsymbol{\alpha}) \nabla_{\boldsymbol{\theta}}^T \frac{\prod_{l=1}^{M} \theta_l^{C_{N,l}(\boldsymbol{\alpha})}}{(1-\theta_m)^N}, \quad (71)
\end{aligned}$$

where $\mathcal{A}_m$ is defined in (7), $\boldsymbol{\alpha}$ is a realization of the random vector $\mathbf{x}$, and the last equality is obtained by replacing the order of the derivative and the sum. Applying the product rule on (71), results in

$$\begin{aligned}
&\nabla_{\boldsymbol{\theta}}^T \left\{ \mathrm{E}_{\boldsymbol{\theta}} \left[\hat{\theta}_m | \mathbf{x} \in \mathcal{A}_m \right] \right\} \\
&= \sum_{\boldsymbol{\alpha} \in \mathcal{A}_m} \hat{\theta}_m(\boldsymbol{\alpha}) \mathbf{v}^T(\boldsymbol{\alpha}, \boldsymbol{\theta}) \frac{\prod_{l=1}^{M} \theta_l^{C_{N,l}(\boldsymbol{\alpha})}}{(1-\theta_m)^N} \\
&\quad - \mathbf{e}_m^T \frac{N}{(1-\theta_m)} \sum_{\boldsymbol{\alpha} \in \mathcal{A}_m} \hat{\theta}_m(\boldsymbol{\alpha}) \frac{\prod_{l=1}^{M} \theta_l^{C_{N,l}(\boldsymbol{\alpha})}}{(1-\theta_m)^N} \\
&\quad = \mathrm{E}_{\boldsymbol{\theta}} \left[\hat{\theta}_m(\mathbf{x}) \mathbf{v}^T(\mathbf{x}, \boldsymbol{\theta}) | \mathbf{x} \in \mathcal{A}_m \right] \\
&\quad - \mathbf{e}_m^T \frac{N}{(1-\theta_m)} \mathrm{E}_{\boldsymbol{\theta}} \left[\hat{\theta}_m(\mathbf{x}) | \mathbf{x} \in \mathcal{A}_m \right], \quad (72)
\end{aligned}$$

where the random vector $\mathbf{v}(\mathbf{x}, \boldsymbol{\theta})$ is defined in (39) and the last equality is obtained by substituting the definition of the conditional pmf from (10). By substituting (72) in (37) we obtain (38).

Appendix C
Proof of Theorem 1

In this appendix we develop the new mmCCRB from Theorem 1. The proof is divided into: 1) the development of Lemma 3 in Subsection C-A; 2) the main development of the bound based on the covariance inequality in Subsection C-B; and 3) derivation of the equality condition in Subsection C-C. To this end, we define $\boldsymbol{\Gamma}(\mathbf{x}, \boldsymbol{\theta})$ as a diagonal $M \times M$ matrix with the following elements on its diagonal:

$$[\boldsymbol{\Gamma}(\mathbf{x}, \boldsymbol{\theta})]_{m,m} \triangleq \epsilon_m(\boldsymbol{\theta}) \mathbf{1}_{\{m \in G_{N,0}(\mathbf{x})\}}, \ m = 1 \ldots, M, \quad (73)$$

where

$$\epsilon_m(\boldsymbol{\theta}) \triangleq \hat{\theta}_m - \mathrm{E}_{\boldsymbol{\theta}} \left[\hat{\theta}_m | \mathbf{x} \in \mathcal{A}_m \right], \ m = 1, \ldots, M. \quad (74)$$

A. Lemma 3

In this subsection we prove the following Lemma.

Lemma 3:

$$\mathrm{E}_{\boldsymbol{\theta}} \left[\boldsymbol{\Gamma}(\mathbf{x}, \boldsymbol{\theta}) \boldsymbol{\Delta}^T(\mathbf{x}, \boldsymbol{\theta}) \right] = \mathbf{S}(\boldsymbol{\theta}), \quad (75)$$

where $\mathbf{S}(\boldsymbol{\theta})$ and $\boldsymbol{\Gamma}(\mathbf{x}, \boldsymbol{\theta})$ are defined in (37) and (73), respectively.

Proof: By substituting (36), (73), and (74) in (75) one obtains that the mth row of $\mathrm{E}_{\boldsymbol{\theta}} \left[\boldsymbol{\Gamma}(\mathbf{x}, \boldsymbol{\theta}) \boldsymbol{\Delta}^T(\mathbf{x}, \boldsymbol{\theta}) \right]$ on the r.h.s. of (75) satisfies

$$\begin{aligned}
&\mathrm{E}_{\boldsymbol{\theta}} \left[\epsilon_m(\boldsymbol{\theta}) \boldsymbol{\Delta}_{1:M,m}^T(\mathbf{x}, \boldsymbol{\theta}) \right] = \Pr(\mathbf{x} \in \mathcal{A}_m; \boldsymbol{\theta}) \\
&\quad \times \sum\nolimits_{\boldsymbol{\alpha} \in \mathcal{A}_m} \epsilon_m(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^T \Pr(\mathbf{x} = \boldsymbol{\alpha} | \mathbf{x} \in \mathcal{A}_m; \boldsymbol{\theta}), \quad (76)
\end{aligned}$$

$m = 1, \ldots, M$, where we use the fact that

$$\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x} | \mathbf{x} \in \mathcal{A}_m; \boldsymbol{\theta}) = \frac{\nabla_{\boldsymbol{\theta}} p(\mathbf{x} | \mathbf{x} \in \mathcal{A}_m; \boldsymbol{\theta})}{p(\mathbf{x} | \mathbf{x} \in \mathcal{A}_m; \boldsymbol{\theta})},$$

in which $\mathcal{A}_m$ is defined in (7). Then, by applying the product rule on the r.h.s. of (76), we obtain

$$\begin{aligned}
&\sum_{\boldsymbol{\alpha}\in\mathcal{A}_m} \epsilon_m(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}^T \Pr(\mathbf{x}=\boldsymbol{\alpha}|\mathbf{x}\in\mathcal{A}_m;\boldsymbol{\theta}) \\
&= \nabla_{\boldsymbol{\theta}}^T \left\{ \sum_{\boldsymbol{\alpha}\in\mathcal{A}_m} \epsilon_m(\boldsymbol{\theta}) \Pr\left(\mathbf{x}=\boldsymbol{\alpha}|\mathbf{x}\in\mathcal{A}_m;\boldsymbol{\theta}\right) \right\} \\
&\quad - \sum_{\boldsymbol{\alpha}\in\mathcal{A}_m} \nabla_{\boldsymbol{\theta}}^T\{\epsilon_m(\boldsymbol{\theta})\} \Pr(\mathbf{x}=\boldsymbol{\alpha}|\mathbf{x}\in\mathcal{A}_m(\mathbf{x});\boldsymbol{\theta}) \\
&= \nabla_{\boldsymbol{\theta}}^T \left\{ \mathrm{E}_{\boldsymbol{\theta}}\left[\epsilon_m(\boldsymbol{\theta})|\mathbf{x}\in\mathcal{A}_m\right] \right\} \\
&\quad - \sum_{\boldsymbol{\alpha}\in\mathcal{A}_m} \nabla_{\boldsymbol{\theta}}^T\{\epsilon_m(\boldsymbol{\theta})\} \Pr(\mathbf{x}=\boldsymbol{\alpha}|\mathbf{x}\in\mathcal{A}_m(\mathbf{x});\boldsymbol{\theta}). \quad (77)
\end{aligned}$$

Computing the conditional expectation of (74), given the event that $m \in G_{N,0}(\mathbf{x})$, results in

$$\mathrm{E}_{\boldsymbol{\theta}}\left[\epsilon_m(\boldsymbol{\theta})|\mathbf{x}\in\mathcal{A}_m\right] = 0,\ m=1,\ldots,M. \quad (78)$$

In addition, computing the gradient of (74) results in

$$\nabla_{\boldsymbol{\theta}}^T \epsilon_m(\boldsymbol{\theta}) = -\nabla_{\boldsymbol{\theta}}^T \mathrm{E}_{\boldsymbol{\theta}}\left[\hat{\theta}_m|\mathbf{x}\in\mathcal{A}_m\right], \quad (79)$$

$m=1,\ldots,M$. By using the conditional expectation definition, and then substituting (78) and (79) in (77), one obtains

$$\begin{aligned}
&\sum_{\boldsymbol{\alpha}\in\mathcal{A}_m} \epsilon_m(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}^T \Pr(\mathbf{x}=\boldsymbol{\alpha}|\mathbf{x}\in\mathcal{A}_m(\mathbf{x});\boldsymbol{\theta}) \\
&= \nabla_{\boldsymbol{\theta}}^T \left\{ \mathrm{E}_{\boldsymbol{\theta}}\left[\hat{\theta}_m|\mathbf{x}\in\mathcal{A}_m\right] \right\} \sum_{\boldsymbol{\alpha}\in\mathcal{A}_m} \Pr(\mathbf{x}=\boldsymbol{\alpha}|\mathbf{x}\in\mathcal{A}_m;\boldsymbol{\theta}) \\
&= \nabla_{\boldsymbol{\theta}}^T \left\{ \mathrm{E}_{\boldsymbol{\theta}}\left[\hat{\theta}_m|\mathbf{x}\in\mathcal{A}_m\right] \right\}, \quad (80)
\end{aligned}$$

where the last equality stems from the fact that for a conditional pmf we have $\sum_{\boldsymbol{\alpha}\in\mathcal{A}_m} \Pr(\mathbf{x}=\boldsymbol{\alpha}|m\in G_{N,0}(\mathbf{x});\boldsymbol{\theta}) = 1$. By substituting the definition of the missing-mass bias vector, $\mathbf{b}_{N,0}(\boldsymbol{\theta})$, from (30) in (80) and using the fact that $\nabla_{\boldsymbol{\theta}}^T \theta_m = \mathbf{e}_m^T$, one obtains

$$\begin{aligned}
&\sum_{\boldsymbol{\alpha}\in\mathcal{A}_m} \epsilon_m(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}^T \Pr(\mathbf{x}=\boldsymbol{\alpha}|\mathbf{x}\in\mathcal{A}_m(\mathbf{x});\boldsymbol{\theta}) \\
&\qquad = \nabla_{\boldsymbol{\theta}}^T \left\{ \frac{[\mathbf{b}_{N,0}(\boldsymbol{\theta})]_m}{\Pr(\mathbf{x}\in\mathcal{A}_m;\boldsymbol{\theta})} \right\} + \mathbf{e}_m^T, \quad (81)
\end{aligned}$$

$m=1,\ldots,M$. Substitution of (37) in (81) and then substituting the result in (76), results in (75). ∎

B. Covariance inequality

The following part of the proof is along the path of the proof from [27] for the CCRB on the MSE in a conventional estimation problem. Let $\mathbf{W}\in\mathbb{R}^{M\times M}$ be an arbitrary matrix and $\tilde{\boldsymbol{\theta}}\in\Omega_{\boldsymbol{\theta}}$ is a local parameter vector. Then,

$$\begin{aligned}
\mathbf{0} &\preceq \mathrm{E}_{\tilde{\boldsymbol{\theta}}}\left[\left(\boldsymbol{\Gamma}(\mathbf{x},\tilde{\boldsymbol{\theta}}) - \mathbf{W}^T\mathbf{U}\mathbf{U}^T\boldsymbol{\Delta}(\mathbf{x},\tilde{\boldsymbol{\theta}}) \right) \right. \\
&\quad \left. \times \left(\boldsymbol{\Gamma}(\mathbf{x},\tilde{\boldsymbol{\theta}}) - \mathbf{W}^T\mathbf{U}\mathbf{U}^T\boldsymbol{\Delta}(\mathbf{x},\tilde{\boldsymbol{\theta}}) \right)^T \right] \\
&= \mathrm{E}_{\tilde{\boldsymbol{\theta}}}\left[\boldsymbol{\Gamma}(\mathbf{x},\tilde{\boldsymbol{\theta}})\boldsymbol{\Gamma}^T(\mathbf{x},\tilde{\boldsymbol{\theta}}) \right] - \mathbf{S}^T(\tilde{\boldsymbol{\theta}})\mathbf{U}\mathbf{U}^T\mathbf{W} \\
&\quad -\mathbf{W}^T\mathbf{U}\mathbf{U}^T\mathbf{S}(\tilde{\boldsymbol{\theta}}) + \mathbf{W}^T\mathbf{U}\mathbf{U}^T\mathbf{J}^{(0)}(\tilde{\boldsymbol{\theta}})\mathbf{U}\mathbf{U}^T\mathbf{W}, \quad (82)
\end{aligned}$$

where we substitute (75) from Lemma 3 and the mmFIM definition from (35). By rearranging (82), we obtain

$$\begin{aligned}
\mathbf{0} &\preceq \mathrm{E}_{\tilde{\boldsymbol{\theta}}}\left[\boldsymbol{\Gamma}(\mathbf{x},\tilde{\boldsymbol{\theta}})\boldsymbol{\Gamma}^T(\mathbf{x},\tilde{\boldsymbol{\theta}}) \right] - \mathbf{S}^T(\tilde{\boldsymbol{\theta}})\mathbf{U}\mathbf{U}^T\mathbf{W} \\
&\quad -\mathbf{W}^T\mathbf{U}\mathbf{U}^T\mathbf{S}(\tilde{\boldsymbol{\theta}}) + \mathbf{W}^T\mathbf{U}\mathbf{U}^T\mathbf{J}^{(0)}(\tilde{\boldsymbol{\theta}})\mathbf{U}\mathbf{U}^T\mathbf{W}. \quad (83)
\end{aligned}$$

By applying the trace operator on (83), it can be verified that the matrix inequality in (83) provides a family of bounds on the mmMSE from (28), which depends on the specific choice of the matrix $\mathbf{W}$. Theorem 1 is obtained by choosing the optimal member from this family, as described in the following.

Since $\mathbf{U}^T\mathbf{J}^{(0)}(\tilde{\boldsymbol{\theta}})\mathbf{U}$ is a non-singular matrix under regularity Condition C.1, then it is shown in [27] that the greatest lower bound, i.e. the supremum of the r.h.s. of (83) over $\mathbf{W}$, is obtained by a matrix $\mathbf{W}$ which satisfies

$$\mathbf{W}^T\mathbf{U} = \mathbf{S}^T(\tilde{\boldsymbol{\theta}})\mathbf{U}\left(\mathbf{U}^T\mathbf{J}^{(0)}(\tilde{\boldsymbol{\theta}})\mathbf{U}\right)^{-1}. \quad (84)$$

By substituting (84) into (83), one obtains

$$\begin{aligned}
&\mathrm{E}_{\tilde{\boldsymbol{\theta}}}\left[\boldsymbol{\Gamma}(\mathbf{x},\tilde{\boldsymbol{\theta}})\boldsymbol{\Gamma}^T(\mathbf{x},\tilde{\boldsymbol{\theta}}) \right] \\
&\qquad \succeq \mathbf{S}^T(\tilde{\boldsymbol{\theta}})\mathbf{U}(\mathbf{U}^T\mathbf{J}^{(0)}(\tilde{\boldsymbol{\theta}})\mathbf{U})^{-1}\mathbf{U}^T\mathbf{S}(\tilde{\boldsymbol{\theta}}). \quad (85)
\end{aligned}$$

By applying the trace operator on (85) and substituting (73), we obtain

$$\begin{aligned}
&\sum_{m=1}^{M} \mathrm{E}_{\tilde{\boldsymbol{\theta}}}\left[\epsilon_m(\tilde{\boldsymbol{\theta}})^2 \mathbf{1}_{\{m\in G_{N,0}(\mathbf{x})\}} \right] \\
&\qquad \geq \mathrm{trace}(\mathbf{S}^T(\tilde{\boldsymbol{\theta}})\mathbf{U}(\mathbf{U}^T\mathbf{J}^{(0)}(\tilde{\boldsymbol{\theta}})\mathbf{U})^{-1}\mathbf{U}^T\mathbf{S}(\tilde{\boldsymbol{\theta}})). \quad (86)
\end{aligned}$$

The following equation relates the mmMSE from (28) and the l.h.s. of (86). From (74), it can be see that

$$\begin{aligned}
&\mathrm{E}_{\tilde{\boldsymbol{\theta}}}\left[(\hat{\theta}_m - \tilde{\theta}_m)^2 | \mathbf{x}\in\mathcal{A}_m \right] \\
&= \mathrm{E}_{\tilde{\boldsymbol{\theta}}}\left[(\epsilon_m(\tilde{\boldsymbol{\theta}}) + \mathrm{E}_{\tilde{\boldsymbol{\theta}}}\left[\hat{\theta}_m|\mathbf{x}\in\mathcal{A}_m\right] - \tilde{\theta}_m)^2 | \mathbf{x}\in\mathcal{A}_m \right] \\
&= \mathrm{E}_{\tilde{\boldsymbol{\theta}}}\left[\epsilon_m^2(\tilde{\boldsymbol{\theta}}) | \mathbf{x}\in\mathcal{A}_m \right] + (\mathrm{E}_{\tilde{\boldsymbol{\theta}}}\left[\hat{\theta}_m|\mathbf{x}\in\mathcal{A}_m\right] - \tilde{\theta}_m)^2 \\
&= \mathrm{E}_{\tilde{\boldsymbol{\theta}}}\left[\epsilon_m^2(\tilde{\boldsymbol{\theta}}) | \mathbf{x}\in\mathcal{A}_m \right] + \left(\frac{\left[\mathbf{b}_{N,0}(\tilde{\boldsymbol{\theta}})\right]_m}{\Pr(\mathbf{x}\in\mathcal{A}_m;\tilde{\boldsymbol{\theta}})} \right)^2, \quad (87)
\end{aligned}$$

where the second equality stems from (78) and the last equality is obtained by substituting (30). By multiplying (87) by $\Pr(\mathbf{x}\in\mathcal{A}_m;\tilde{\boldsymbol{\theta}})$, then summing the result over $m=1,\ldots,M$, and substituting the result in (86), one obtains the mmBCRB on the mmMSE of biased estimator evaluated at the local point, $\tilde{\boldsymbol{\theta}}$, in (40)-(41).

C. Derivation of (42)

In this subsection we develop the equality condition in (42). According to Cauchy-Schwartz inequality properties (which is the basis for the covariance inequality) the equality in (82) for $\mathbf{W}$ that satisfies (84) holds if

$$\begin{aligned}
&\mathrm{E}_{\tilde{\boldsymbol{\theta}}}\left[(\boldsymbol{\Gamma}(\mathbf{x},\tilde{\boldsymbol{\theta}}) - \mathbf{S}^T(\tilde{\boldsymbol{\theta}})\mathbf{U}(\mathbf{U}^T\mathbf{J}^{(0)}(\tilde{\boldsymbol{\theta}})\mathbf{U})^{-1}\mathbf{U}^T\boldsymbol{\Delta}(\mathbf{x},\tilde{\boldsymbol{\theta}})) \right. \\
&\left. \times (\boldsymbol{\Gamma}(\mathbf{x},\tilde{\boldsymbol{\theta}}) - \mathbf{S}^T(\tilde{\boldsymbol{\theta}})\mathbf{U}(\mathbf{U}^T\mathbf{J}^{(0)}(\tilde{\boldsymbol{\theta}})\mathbf{U})^{-1}\mathbf{U}^T\boldsymbol{\Delta}(\mathbf{x},\tilde{\boldsymbol{\theta}}))^T \right] \\
&\qquad = \mathbf{0}. \quad (88)
\end{aligned}$$

The condition in (88) holds if

$$\mathbf{\Gamma}(\mathbf{x},\tilde{\boldsymbol{\theta}}) = \mathbf{S}^T(\tilde{\boldsymbol{\theta}})\mathbf{U}(\mathbf{U}^T\mathbf{J}^{(0)}(\tilde{\boldsymbol{\theta}})\mathbf{U})^{-1}\mathbf{U}^T\mathbf{\Delta}(\mathbf{x},\tilde{\boldsymbol{\theta}}) \tag{89}$$

which, by using (73), implies

$$\epsilon_m(\tilde{\boldsymbol{\theta}}) = \left[\mathbf{S}(\tilde{\boldsymbol{\theta}})^T\mathbf{U}(\mathbf{U}^T\mathbf{J}^{(0)}(\tilde{\boldsymbol{\theta}})\mathbf{U})^{-1}\mathbf{U}^T\mathbf{\Delta}(\mathbf{x},\tilde{\boldsymbol{\theta}})\right]_{m,m}, \tag{90}$$

$\forall m = 1,\ldots,M$ that satisfies $m \in G_{N,0}\mathbf{x}$. By substituting (30) and (74) in (90) we get (42).

Appendix D
Proof of Lemma 2

In this appendix, we develop the closed-form mmFIM, defined in (35), for the observation model from (3). By substituting (10) in (36), one obtains

$$\begin{aligned}&\mathbf{\Delta}_{1:M,m}(\mathbf{x},\boldsymbol{\theta}) = \\ &\nabla_{\boldsymbol{\theta}}\left(\sum_{l=1}^{M} C_{N,l}(\mathbf{x})\log\theta_l - N\log(1-\theta_m)\right)\mathbf{1}_{\{m\in G_{N,0}(\mathbf{x})\}} \\ &= \left(\mathbf{v}^T(\mathbf{x},\boldsymbol{\theta}) - \frac{N}{1-\theta_m}\mathbf{e}_m^T\right)\mathbf{1}_{\{m\in G_{N,0}(\mathbf{x})\}},\end{aligned} \tag{91}$$

where $\mathbf{v}(\mathbf{x},\boldsymbol{\theta})$ is defined in (39). By substituting (91) in the mmFIM definition from (35), we obtain that the (k,l)th element of the mmFIM for our model is given by

$$\begin{aligned}\left[\mathbf{J}^{(0)}(\boldsymbol{\theta})\right]_{k,l} &= \sum_{m=1}^{M}\mathrm{E}_{\boldsymbol{\theta}}\left[[\mathbf{\Delta}(\mathbf{x},\boldsymbol{\theta})]_{k,m}[\mathbf{\Delta}(\mathbf{x},\boldsymbol{\theta})]_{l,m}\right] \\ &= \sum_{m=1}^{M}\mathrm{E}_{\boldsymbol{\theta}}\left[\left(\frac{C_{N,k}(\mathbf{x})}{\theta_k} + \frac{N}{1-\theta_m}\delta_{k,m}\right)\right. \\ &\quad\left.\times\left(\frac{C_{N,l}(\mathbf{x})}{\theta_l} + \frac{N}{1-\theta_m}\delta_{l,m}\right)\mathbf{1}_{\{m\in G_{N,0}(\mathbf{x})\}}\right] \\ &= \sum_{m=1}^{M}\left\{\frac{1}{\theta_k\theta_l}\mathrm{E}_{\boldsymbol{\theta}}\left[C_{N,k}(\mathbf{x})C_{N,l}(\mathbf{x})|m\in G_{N,0}(\mathbf{x})\right]\right. \\ &\quad + \frac{N\delta_{k,m}}{(1-\theta_m)\theta_l}\mathrm{E}_{\boldsymbol{\theta}}\left[C_{N,l}(\mathbf{x})|\mathbf{x}\in\mathcal{A}_m\right] \\ &\quad + \frac{N\delta_{l,m}}{(1-\theta_m)\theta_k}\mathrm{E}_{\boldsymbol{\theta}}\left[C_{N,k}(\mathbf{x})|\mathbf{x}\in\mathcal{A}_m\right] \\ &\quad\left. + \frac{N^2\delta_{k,m}\delta_{l,m}}{(1-\theta_m)^2}\right\}\Pr(\mathbf{x}\in\mathcal{A}_m;\boldsymbol{\theta}),\end{aligned} \tag{92}$$

where the last equality is obtained by using the law of total probability. In the following we compute the conditional expectation terms from (92). First, it can be seen that

$$\begin{aligned}\sum_{\boldsymbol{\alpha}\in\mathcal{A}_m}\Pr(\mathbf{x}=\boldsymbol{\alpha};\boldsymbol{\theta}) &= \Pr(\mathbf{x}\in\mathcal{A}_m;\boldsymbol{\theta}) \\ &= \left(\sum\nolimits_{n=1,n\neq m}^{N}\theta_n\right)^N,\end{aligned} \tag{93}$$

$m = 1,\ldots,M$, where $\mathcal{A}_m$ is defined in (7) and where the probability of the mth element to be unobserved on the r.h.s. of (93) is an alternative way of writing the r.h.s of (8). Then, by using (10), it can be verified that

$$\begin{aligned}&\mathrm{E}_{\boldsymbol{\theta}}\left[C_{N,k}(\mathbf{x})|\mathbf{x}\in\mathcal{A}_m\right] \\ &\quad= \sum_{\boldsymbol{\alpha}\in\mathcal{A}_m} C_{N,k}(\boldsymbol{\alpha})\frac{\prod_{n=1}^{M}\theta_n^{C_{N,n}(\boldsymbol{\alpha})}}{(1-\theta_m)^N} \\ &\quad= \frac{\theta_k}{(1-\theta_m)^N}\sum_{\boldsymbol{\alpha}\in\mathcal{A}_m}\frac{\partial}{\partial\theta_k}\prod_{n=1}^{M}\theta_n^{C_{N,n}(\boldsymbol{\alpha})} \\ &\quad= \frac{\theta_k}{(1-\theta_m)^N}\frac{\partial}{\partial\theta_k}\sum_{\boldsymbol{\alpha}\in\mathcal{A}_m}\Pr(\mathbf{x}=\boldsymbol{\alpha};\boldsymbol{\theta}),\end{aligned} \tag{94}$$

$m,k = 1,\ldots,M$. It should be noted that the last equality in (94) is only valid if we use the probability term, $\Pr(\mathbf{x}=\boldsymbol{\alpha};\boldsymbol{\theta})$, in (93). By substituting (93) in (94), one obtains

$$\begin{aligned}&\mathrm{E}_{\boldsymbol{\theta}}\left[C_{N,k}(\mathbf{x})|\mathbf{x}\in\mathcal{A}_m\right] \\ &\quad= \frac{\theta_k}{(1-\theta_m)^N}\frac{\partial}{\partial\theta_k}\Big(\sum_{n=1,n\neq m}^{N}\theta_n\Big)^N \\ &\quad= \begin{cases}\frac{N\theta_k}{1-\theta_m} & m\neq k \\ 0 & m=k\end{cases}.\end{aligned} \tag{95}$$

Similar to the derivation of (94), by using (10) we obtain

$$\begin{aligned}&\mathrm{E}_{\boldsymbol{\theta}}\left[C_{N,k}(\mathbf{x})C_{N,l}(\mathbf{x})|\mathbf{x}\in\mathcal{A}_m\right] \\ &= \sum_{\boldsymbol{\alpha}\in\mathcal{A}_m} C_{N,k}(\boldsymbol{\alpha})C_{N,l}(\boldsymbol{\alpha})\frac{\prod_{n=1}^{M}\theta_n^{C_{N,n}(\boldsymbol{\alpha})}}{(1-\theta_m)^N} \\ &= \sum_{\boldsymbol{\alpha}\in\mathcal{A}_m}\frac{\theta_k\theta_l}{(1-\theta_m)^N}\frac{\partial^2}{\partial\theta_k\partial\theta_l}\prod_{n=1}^{M}\theta_n^{C_{N,n}(\boldsymbol{\alpha})} \\ &\quad+\frac{\delta_{k,l}}{(1-\theta_m)^N}\sum_{\boldsymbol{\alpha}\in\mathcal{A}_m}C_{N,k}(\boldsymbol{\alpha})\prod_{n=1}^{M}\theta_n^{C_{N,n}(\boldsymbol{\alpha})} \\ &= \frac{\theta_k\theta_l}{(1-\theta_m)^N}\frac{\partial^2}{\partial\theta_k\partial\theta_l}\sum_{\boldsymbol{\alpha}\in\mathcal{A}_m}\Pr(\mathbf{x}=\boldsymbol{\alpha};\boldsymbol{\theta}) \\ &\quad+\frac{N\delta_{k,l}}{1-\theta_m}\mathrm{E}_{\boldsymbol{\theta}}\left[C_{N,k}(\mathbf{x})|\mathbf{x}\in\mathcal{A}_m\right],\end{aligned} \tag{96}$$

where we replace the order of the derivative and the sum, and we use (3). Again, it should be noted that the last equality in (96) is only valid if we use the probability term, $\Pr(\mathbf{x}=\boldsymbol{\alpha};\boldsymbol{\theta})$, in (93). By substituting (93) and (95) in (96), one obtains

$$\begin{aligned}&\mathrm{E}_{\boldsymbol{\theta}}\left[C_{N,k}(\mathbf{x})C_{N,l}(\mathbf{x})|\mathbf{x}\in\mathcal{A}_m\right] = \\ &\frac{\theta_k\theta_l}{(1-\theta_m)^N}\frac{\partial^2}{\partial\theta_k\partial\theta_l}\Big(\sum_{n=1,n\neq m}^{N}\theta_n\Big)^N + \frac{N\theta_k\delta_{k,l}(1-\delta_{m,k})}{1-\theta_m} \\ &\quad= \begin{cases}\frac{N(N-1)\theta_k\theta_l}{(1-\theta_m)^2} & \text{if } m\neq k,\ l\neq m,\ k\neq l \\ \frac{N(N-1)\theta_k^2}{(1-\theta_m)^2}+\frac{N\theta_k}{1-\theta_m} & \text{if } k=l\neq m \\ 0 & \text{otherwise}\end{cases}.\end{aligned} \tag{97}$$

Thus, by substituting (8), (95), and (97) in (92), one obtains that the elements of the mmFIM are given by

$$\begin{aligned}[\mathbf{J}^{(0)}(\boldsymbol{\theta})]_{k,k} &= \sum_{m=1}^{M}\left(\frac{N(N-1)}{(1-\theta_m)^2}+\frac{N}{\theta_k(1-\theta_m)}\right)(1-\theta_m)^N \\ &\quad +\frac{N}{(1-\theta_k)^2}(1-\theta_k)^N \\ &\quad -\frac{N}{\theta_k(1-\theta_k)}(1-\theta_k)^N, \end{aligned} \tag{98}$$

for $k=1,\ldots,M$, and

$$\begin{aligned}[\mathbf{J}^{(0)}(\boldsymbol{\theta})]_{k,l} &= \sum_{m=1}^{M}\frac{N(N-1)}{(1-\theta_m)^2}(1-\theta_m)^N \\ &\quad +\frac{N}{(1-\theta_k)^2}(1-\theta_k)^N+\frac{N}{(1-\theta_l)^2}(1-\theta_l)^N, \end{aligned} \tag{99}$$

for $k,l=1,\ldots,M$, $k\neq l$. By rearranging the elements in (98) and (99), and substituting (8), we obtain the closed-form matrix $\mathbf{J}^{(0)}(\boldsymbol{\theta})$ in its matrix representation in (44).

REFERENCES

[1] H. Robbins, "Estimating the total probability of the unobserved outcomes of an experiment," *The Annals of Mathematical Statistics*, vol. 39, no. 1, pp. 256–257, 1968.

[2] I. J. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, pp. 237–264, 1953.

[3] B. Efron and R. Thisted, "Estimating the number of unseen species: How many words did Shakespeare know?" *Biometrika*, vol. 63, no. 3, pp. 435–447, 12 1976.

[4] C. Budianu and L. Tong, "Good-Turing estimation of the number of operating sensors: a large deviations analysis," in *Proc. ICASSP 2004*, vol. 2, May 2004, pp. 1029–1032.

[5] C. Budianu, S. Ben-David, and L. Tong, "Estimation of the number of operating sensors in large-scale sensor networks with mobile access," *IEEE Trans. Signal Processing*, vol. 54, no. 5, pp. 1703–1715, May 2006.

[6] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Annual Meeting on Association for Computational Linguistics*, 1996, pp. 310–318.

[7] S. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE Trans. acoustics, speech, and signal processing*, vol. 35, no. 3, pp. 400–401, 1987.

[8] A. Orlitsky, N. P. Santhanam, and J. Zhang, "Always Good Turing: Asymptotically optimal probability estimation," *Science*, vol. 302, no. 5644, pp. 427–431, 2003.

[9] G. Valiant and P. Valiant, "Estimating the unseen: An n/log(n)-sample estimator for entropy and support size, shown optimal via new CLTs," in *the 43rd ACM Symposium on Theory of Computing*, 2011, pp. 685–694.

[10] P. S. Laplace, *Pierre-Simon Laplace Philosophical Essay on Probabilities: Translated from the fifth French edition of 1825 With Notes by the Translator*. Springer Science & Business Media, 2012, vol. 13.

[11] A. Nadas, "On Turing's formula for word probabilities," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 33, no. 6, pp. 1414–1416, Dec. 1985.

[12] W. A. Gale and G. Sampson, "Good-Turing frequency estimation without tears," *Journal of quantitative linguistics*, vol. 2, no. 3, pp. 217–237, 1995.

[13] D. Braess and T. Sauer, "Bernstein polynomials and learning theory," *Journal of Approximation Theory*, vol. 128, no. 2, pp. 187 – 206, 2004.

[14] R. Krichevsky and V. Trofimov, "The performance of universal encoding," *IEEE Trans. Information Theory*, vol. 27, no. 2, pp. 199–207, March 1981.

[15] W. Gale and K. Church, "What's wrong with adding one," *Corpus-Based Research into Language: In honour of Jan Aarts*, pp. 189–200, 1994.

[16] K. P. Burnham and W. S. Overton, "Robust estimation of population size when capture probabilities vary among animals," *Ecology*, vol. 60, no. 5, pp. 927–936, 1979.

[17] I. Good and G. Toulmin, "The number of new species, and the increase in population coverage, when a sample is increased," *Biometrika*, vol. 43, no. 1-2, pp. 45–63, 1956.

[18] C. X. Mao and B. G. Lindsay, "Estimating the number of classes," *the Annals of Statistics*, pp. 917–930, 2007.

[19] A. Orlitsky and A. T. Suresh, "Competitive distribution estimation: Why is Good-Turing good," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 2143–2151.

[20] B. H. Juang and S. H. Lo, "On the bias of the Turing-Good estimate of probabilities," *IEEE Trans. Signal Processing*, vol. 42, no. 2, pp. 496–498, Feb. 1994.

[21] D. A. McAllester and R. E. Schapire, "On the convergence rate of Good-Turing estimators," in *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*. Morgan Kaufmann Publishers Inc., 2000, pp. 1–6.

[22] A. B. Wagner, P. Viswanath, and S. R. Kulkarni, "Probability estimation in the rare-events regime," *IEEE Trans. Information Theory*, vol. 57, no. 6, pp. 3207–3229, June 2011.

[23] D. Berend and A. Kontorovich, "The missing mass problem," *Statistics & Probability Letters*, vol. 82, no. 6, pp. 1102 – 1110, 2012.

[24] D. Berend, A. Kontorovich *et al.*, "On the concentration of the missing mass," *Electronic Communications in Probability*, vol. 18, 2013.

[25] Y. G. Yatracos, "On the rare species of a population," *Journal of Statistical Planning and Inference*, vol. 48, no. 3, pp. 321 – 329, 1995.

[26] J. Gorman and A. Hero, "Lower bounds for parametric estimation with constraints," *IEEE Trans. Information Theory*, vol. 36, no. 6, pp. 1285–1301, Nov. 1990.

[27] P. Stoica and B. C. Ng, "On the Cramér-Rao bound under parametric constraints," *IEEE Signal Processing Letters*, vol. 5, no. 7, pp. 177–179, July 1998.

[28] E. Nitzan, T. Routtenberg, and J. Tabrikian, "Cramér-Rao bound for constrained parameter estimation using Lehmann-unbiasedness," *IEEE Trans. Signal Processing*, vol. 67, no. 3, pp. 753–768, Feb 2019.

[29] E. Nitzan, T. Routtenberg, and J. Tabrikian, "Cramér-Rao bound under norm constraint," *IEEE Signal Processing Letters*, vol. 26, no. 9, pp. 1393–1397, 2019.

[30] Z. Ben-Haim and Y. C. Eldar, "The Cramér-Rao bound for estimating a sparse parameter vector," *IEEE Trans. Signal Processing*, vol. 58, no. 6, pp. 3384–3389, June 2010.

[31] T. Routtenberg and L. Tong, "Estimation after parameter selection: Performance analysis and estimation methods," *IEEE Trans. Signal Processing*, vol. 64, no. 20, pp. 5268–5281, Oct. 2016.

[32] N. Harel and T. Routtenberg, "Low-complexity methods for estimation after parameter selection," *IEEE Trans. Signal Processing*, vol. 68, pp. 1152–1167, 2020.

[33] E. Meir and T. Routtenberg, "Cramér-Rao bound for estimation after model selection and its application to sparse vector estimation," *ArXiv e-prints, https://arxiv.org/abs/1904.06837*, 2019.

[34] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, 3rd ed. New York: Springer Texts in Statistics, 2005.

[35] H. Kesten and N. Morse, "A property of the multinomial distribution," *The Annals of Mathematical Statistics*, vol. 30, no. 1, pp. 120–127, 1959.

[36] A. Cohen and H. B. Sackrowitz, "Admissibility of estimators of the probability of unobserved outcomes," *Annals of the Institute of Statistical Mathematics*, vol. 42, no. 4, pp. 623–636, 1990.

[37] J. Acharya, Y. Bao, Y. Kang, and Z. Sun, "Improved bounds for minimax risk of estimating missing mass," in *2018 IEEE International Symposium on Information Theory (ISIT)*, 2018, pp. 326–330.

[38] F. Gini, "Estimation strategies in the presence of nuisance parameters," *Signal processing*, vol. 55, no. 2, pp. 241–245, 1996.

[39] S. Bar and J. Tabrikian, "Bayesian estimation in the presence of deterministic nuisance parameters; Part I: Performance bounds," *IEEE Trans. Signal Processing*, vol. 63, no. 24, pp. 6632–6646, Dec. 2015.

[40] Y. Hao, A. Orlitsky, and V. Pichapati, "On learning Markov chains," in *Advances in Neural Information Processing Systems*, 2018, pp. 648–657.

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

[42] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice Hall, 1993.

[43] T. Routtenberg and J. Tabrikian, "Non-Bayesian periodic Cramér-Rao bound," *IEEE Trans. Signal Processing*, vol. 61, no. 4, pp. 1019–1032, Feb. 2013.

[44] B.-H. Juang and S. Lo, "On the bias of the Turing-Good estimate of probabilities," *IEEE Trans. signal processing*, vol. 42, no. 2, pp. 496–498, 1994.

[45] V. Berisha and A. O. Hero, "Empirical non-parametric estimation of the Fisher information," *IEEE Signal Processing Letters*, vol. 22, no. 7, pp. 988–992, 2014.

[46] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.