ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095206

# VARIOUS PERFORMANCE BOUNDS ON THE ESTIMATION OF LOW-RANK PROBABILITY MASS FUNCTION TENSORS FROM PARTIAL OBSERVATIONS

*Tomer Hershkovitz*$^{\star}$*, Martin Haardt*$^{\dagger}$ *and Arie Yeredor*$^{\star}$

$^{\star}$School of Electrical Engineering
Tel-Aviv University
Tel Aviv, Israel

$^{\dagger}$Communication Research Laboratory
Ilmenau University of Technology
Ilmenau, Germany

## ABSTRACT

Probability mass function (PMF) estimation using a low-rank model for the PMF tensor has gained increased popularity in recent years. However, its performance evaluation relied mostly on empirical testing. In this work, we derive theoretical bounds on the attainable performance under this model assumption. We begin by deriving the constrained Cramér-Rao Bound (CCRB) on the low-rank decomposition parameters, and then extend the CCRB to bounds on the mean square error in the resulting estimates of the PMF tensor's elements, as well as on the mean Kullback-Leibler divergence (KLD) between the estimated and true PMFs. The asymptotic tightness of these bounds is demonstrated by comparing them to the performance of the Maximum Likelihood estimate in a small-scale simulation example.

***Index Terms*—** PMF Estimation, Low-Rank CPD, Constrained Cramér-Rao Bound, KLD Bound.

## 1. INTRODUCTION

ESTIMATION of the probability mass function (PMF) of a discrete random vector (RV) (whose elements take values in finite alphabets) from partial observations thereof is a key problem in many data analysis contexts in signal processing and machine learning (e.g., recommender systems, data completion, features selection, classification). Since the reliable direct estimation of a PMF tensor using naïve histogram methods requires data sizes that grow exponentially with the dimension of the RV, several alternative estimation paradigms have been proposed in recent years [1–12], which are based on a low-rank non-negative canonical polyadic decomposition (CPD) model assumption for the PMF tensor.

However, the performance of the proposed approaches was only demonstrated empirically, without comparing it to a theoretical bound. Our goal in this paper is to derive useful theoretical estimation bounds on three possible performance measures in PMF estimation problems. Such bounds can be useful for determining whether in a given estimation problem, under prescribed conditions, a desired accuracy level is (at least theoretically) attainable or not. We begin by deriving the Constrained Cramér-Rao Bound (CCRB) on the mean square error (MSE) in unbiased estimation of the underlying low-rank CPD parameters (which are constrained to the *probability simplex*, as explained in the sequel). We then use this bound in order to obtain element-wise MSE bounds on the estimated elements of the PMF tensor when its estimate is the "plug-in" estimate obtained by substituting the estimated CPD parameters in the CPD model. We compare this bound to the (nearly trivial) bound on the MSE in direct estimation of each element without the low-rank CPD model assumption, so as to quantify the potential gain (in terms of the element-wise MSE bound) in using the low-rank assumption whenever this assumption is justified.

In addition, we provide a lower-bound on the attainable mean Kullback-Leibler divergence (KLD, [15]) between the estimated and the true PMF tensors (under the low-rank model assumption).

We demonstrate the possible use of these bounds in a small-scale example, showing that the performance of the Maximum Likelihood estimate (MLE) is bounded by the derived bounds, which, as expected, are asymptotically tight.

### 1.1. Notation

We denote scalars with plain letters ($x$), vectors and matrices with boldface letters ($\boldsymbol{x}$, $\boldsymbol{X}$), and tensors with boldface calligraphic letters ($\boldsymbol{\mathcal{X}}$). RVs are denoted using sans-serif fonts ($\mathsf{X}$). We use $\mathbf{1}$ and $\mathbf{0}$ to denote (resp.) an all-ones and an all-zeros vector (or matrix, depending on context). The elements of vectors, matrices and tensors are denoted using indices in parentheses, e.g., $x(4)$, $X(m,n)$, $\mathcal{X}(i_1,i_2,i_3)$. We use colon notation to denote the column vector of a matrix, e.g., $\boldsymbol{A}(:,2)$. The vec$(\cdot)$ operator creates a column vector by concatenating the (mode-1) columns of its argument (matrix or tensor). $\Pr\{\cdot\}$ denotes probability; Trace$\{\cdot\}$ denotes the trace; $I\{\cdot\}$ denotes the Indicator function (taking the value $1$ if the condition in its argument is satisfied, and $0$ otherwise); $\circ$ denotes an outer product; $\|\cdot\|_2$ denotes the $L2$ norm. We

The authors gratefully acknowledge the support of the German Research Foundation (DFG) under the PROMETHEUS project (Project no. HA 2239/16-1).

use the notation $[T]$ to denote the set of integers $\{1,2,...,T\}$.

## 2. PROBLEM FORMULATION

Let $\mathsf{X}=[\mathsf{X}_1,\mathsf{X}_2,...,\mathsf{X}_N]^T\in\mathbb{R}^N$ be a discrete RV with $\mathsf{X}_n$ taking discrete integer values in $[I_n]$. We denote its joint PMF tensor $\boldsymbol{\mathcal{X}}\in\mathbb{R}^{I_1\times I_2\times...\times I_N}$, where $\mathcal{X}(i_1,i_2,...,i_N)\triangleq Pr\{\mathsf{X}_1=i_1,\mathsf{X}_2=i_2,...,\mathsf{X}_N=i_N\}$, and denote $M\triangleq\prod_{n=1}^{N}I_n$ the total number of elements in $\boldsymbol{\mathcal{X}}$. We assume that $\boldsymbol{\mathcal{X}}$ admits a low-rank non-negative CPD [3] with $F$ components, namely there are $N$ factor matrices $\boldsymbol{A}_1,\boldsymbol{A}_2,...,\boldsymbol{A}_N$, $\boldsymbol{A}_n\in\mathbb{R}^{I_n\times F}$, and a "loading vector" $\boldsymbol{\lambda}\in\mathbb{R}^F$ such that

$$\boldsymbol{\mathcal{X}}=\sum_{f=1}^{F}\lambda_f\cdot\boldsymbol{A}_1(:,f)\circ\boldsymbol{A}_2(:,f)\circ...\circ\boldsymbol{A}_N(:,f). \qquad (1)$$

All elements of $\boldsymbol{\lambda}$ are positive, all elements of $\boldsymbol{A}_1,\boldsymbol{A}_2,...,\boldsymbol{A}_N$ are non-negative and $\mathbf{1}^T\boldsymbol{\lambda}=1$, $\mathbf{1}^T\boldsymbol{A}_n=\mathbf{1}^T, n\in[N]$ (these are commonly known as the *probability simplex* constraints).

Let $\mathsf{Y}=[\mathsf{Y}_1,\mathsf{Y}_2,...,\mathsf{Y}_N]^T$ denote an RV obtained as an incomplete observation of $\mathsf{X}$, such that

$$\mathsf{Y}_n=\begin{cases}\mathsf{X}_n & \text{w.p. } 1-p\\ 0 & \text{w.p. } p\end{cases} \qquad n\in[N], \qquad (2)$$

where $\mathsf{Y}_n$ is drawn independently for each $n$ and independently of all other elements of $\mathsf{X}$ (excluding $X_n$). The known parameter $p$ denotes the outage probability, namely the probability that $X_n$ is unobserved in $\mathsf{Y}$. Assume that $\boldsymbol{y}$ is a realization of $\mathsf{Y}$. Let $B\in[0,N]$ denote the number of non-zero elements of $\boldsymbol{y}$, and let $n_1,...,n_B$ denote their indices. Let $\boldsymbol{\theta}$ denote the vector of all unknown parameters, a concatenation of $\boldsymbol{\lambda}$ with the columns of $\boldsymbol{A}_1,\boldsymbol{A}_2,...,\boldsymbol{A}_N$, such that

$$\boldsymbol{\theta}\triangleq[\boldsymbol{\lambda}^T,\text{vec}^T(\boldsymbol{A}_1),...,\text{vec}^T(\boldsymbol{A}_N)]^T\in\mathbb{R}^K, \qquad (3)$$

where $K\triangleq F\cdot(1+\sum_{n=1}^{N}I_n)$ is the number of unknown parameters. The likelihood function of $\boldsymbol{y}$ parameterized by $\boldsymbol{\theta}$ is given by

$$\begin{aligned}p(\boldsymbol{y};\boldsymbol{\theta})&\triangleq\Pr(\mathsf{Y}=\boldsymbol{y};\boldsymbol{\lambda},\boldsymbol{A}_1,\boldsymbol{A}_2,...,\boldsymbol{A}_N)\\&=\underbrace{p^{N-B}(1-p)^B}_{\triangleq q}\Pr(\mathsf{X}_{n_1}=y_{n_1},...,\mathsf{X}_{n_B}=y_{n_B})\\&=q\sum_{f=1}^{F}\lambda_f\prod_{b=1}^{B}A_{n_b}(y_{n_b},f). \qquad (4)\end{aligned}$$

Our goal is to derive bounds on various quality measures of the estimation of $\boldsymbol{\mathcal{X}}$ from $T$ independent, identically distributed (iid) observations of $\mathsf{Y}$.

## 3. THE CONSTRAINED-CRAMÉR-RAO BOUND ON UNBIASED ESTIMATION OF $\boldsymbol{\theta}$

Given a general probability model $p(\boldsymbol{y};\boldsymbol{\theta})$ specifying the probability of measurements $\boldsymbol{y}\in\mathbb{R}^N$ in terms of unknown parameters $\boldsymbol{\theta}\in\mathbb{R}^K$, where $\boldsymbol{\theta}$ is known to satisfy a set of $R$ constraints of the form $\boldsymbol{c}(\boldsymbol{\theta})=\mathbf{0}\in\mathbb{R}^R$, the CCRB [14] specifies the CRB on any unbiased estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ complying with the same constraints, $\boldsymbol{c}(\widehat{\boldsymbol{\theta}})=\mathbf{0}$:

$$\begin{aligned}&E\left[(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})^T\right]\succeq\\&\boldsymbol{U}(\boldsymbol{\theta})\left[\boldsymbol{U}^T(\boldsymbol{\theta})\boldsymbol{J}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})\right]^{-1}\boldsymbol{U}^T(\boldsymbol{\theta})\triangleq\boldsymbol{B}_{\boldsymbol{\theta}},\end{aligned} \qquad (5)$$

where the columns of $\boldsymbol{U}(\boldsymbol{\theta})\in\mathbb{R}^{K\times(K-R)}$ form an orthonormal basis of the null-space of the gradient matrix[1] $\boldsymbol{C}(\boldsymbol{\theta})\triangleq\frac{\partial\boldsymbol{c}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\in\mathbb{R}^{R\times K}$ of the constraints, namely $\boldsymbol{C}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})=\mathbf{0}\in\mathbb{R}^{R\times(K-R)}$, and where

$$\boldsymbol{J}(\boldsymbol{\theta})\triangleq E\left[\frac{\partial^T\log p(\mathsf{Y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial\log p(\mathsf{Y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right] \qquad (6)$$

is the $K\times K$ Fisher Information Matrix (FIM).

Returning to our specific problem, our model $p(\boldsymbol{y};\boldsymbol{\theta})$ is specified in (4), and the probability simplex constraints give rise to the following set of $R\triangleq 1+F\cdot N$ linear equations:

$$\boldsymbol{c}(\boldsymbol{\theta})\triangleq\begin{bmatrix}\sum_{f=1}^{F}\lambda_f-1\\ \sum_{i=1}^{I_1}A_1(i,1)-1\\ \sum_{i=1}^{I_1}A_1(i,2)-1\\ \vdots\\ \sum_{i=1}^{I_N}A_N(i,F)-1\end{bmatrix}=\mathbf{0}\in\mathbb{R}^R. \qquad (7)$$

$\boldsymbol{C}(\boldsymbol{\theta})$ is then a constant block-diagonal matrix $\boldsymbol{C}$ with $R$ single-row blocks of ones, and $\boldsymbol{U}(\boldsymbol{\theta})$ is therefore also a constant block-diagonal matrix $\boldsymbol{U}$ with $R$ blocks, such that each block is an orthonormal basis of the null-space of an all-ones row-vector (block) of the corresponding length in $\boldsymbol{C}$.

Explicit expressions for the elements of the score vector $\partial\log p(\boldsymbol{y};\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ and of the FIM $\boldsymbol{J}(\boldsymbol{\theta})$ are obtained as follows. For $f\in[F]$,

$$\frac{\partial\log p(\boldsymbol{y};\boldsymbol{\theta})}{\partial\lambda_f}=\frac{q\cdot\alpha_f(\boldsymbol{y};\boldsymbol{\theta})}{p(\boldsymbol{y};\boldsymbol{\theta})} \qquad (8)$$

$$\frac{\partial\log p(\boldsymbol{y};\boldsymbol{\theta})}{\partial A_n(i,f)}=\frac{q\cdot\beta_f(\boldsymbol{y};\boldsymbol{\theta},n,i)}{p(\boldsymbol{y};\boldsymbol{\theta})} \quad \begin{matrix}n\in[N]\\ i\in[I_n]\end{matrix} \qquad (9)$$

where

$$\alpha_f(\boldsymbol{y};\boldsymbol{\theta})\triangleq\prod_{b=1}^{B}A_{n_b}(y_{n_b},f) \qquad (10)$$

and

$$\begin{aligned}\beta_f(\boldsymbol{y};\boldsymbol{\theta},n,i)\triangleq\lambda_f\prod_{\substack{b=1\\ n_b\neq n}}^{B}A_{n_b}(y_{n_b},f)\\ \cdot I\{y_n=i\}\cdot I\left\{n\in\{n_1\dots n_B\}\right\}. \qquad (11)\end{aligned}$$

[1] Assumed to have full row rank.

In order to calculate the elements of the FIM $\boldsymbol{J}(\boldsymbol{\theta})$ we need to take the mean in (6) by summing the product over all $L \triangleq \prod_{n=1}^{N}(1+I_n)$ possible values of $\mathsf{Y}$, denoted $\boldsymbol{y}_1, ..., \boldsymbol{y}_L$, each multiplied by $p(\boldsymbol{y}_\ell;\boldsymbol{\theta})$. For example, for $f_1, f_2 \in [F]$ we have

$$J(f_1,f_2) = \sum_{\ell=1}^{L} \frac{q_\ell^2 \alpha_{f_1}(\boldsymbol{y}_\ell;\boldsymbol{\theta})\alpha_{f_2}(\boldsymbol{y}_\ell;\boldsymbol{\theta})}{p(\boldsymbol{y}_\ell;\boldsymbol{\theta})}, \tag{12}$$

where $q_\ell$ is obtained as in (4), substituting $B$ with $B_\ell$, the number of non-zero elements in $\boldsymbol{y}_\ell$.

Likewise, for $F < m_1, m_2 \leq K$ with one-to-one mappings $m_1 \leftrightarrow (\breve{n}_1, i_1, f_1)$ (such that $\theta(m_1) = A_{\breve{n}_1}(i_1, f_1)$) and $m_2 \leftrightarrow (\breve{n}_2, i_2, f_2)$ we have

$$J(m_1,m_2) = \sum_{\ell=1}^{L} \frac{q_\ell^2 \beta_{f_1}(\boldsymbol{y}_\ell;\boldsymbol{\theta},\breve{n}_1,i_1)\beta_{f_2}(\boldsymbol{y}_\ell;\boldsymbol{\theta},\breve{n}_2,i_2)}{p(\boldsymbol{y}_\ell;\boldsymbol{\theta})}, \tag{13}$$

and for $f_1 \in [F]$ and $m_2 \leftrightarrow (\breve{n}_2, i_2, f_2)$ we have

$$J(f_1,m_2) = \sum_{\ell=1}^{L} \frac{q_\ell^2 \alpha_{f_1}(\boldsymbol{y}_\ell;\boldsymbol{\theta})\beta_{f_2}(\boldsymbol{y}_\ell;\boldsymbol{\theta},\breve{n}_2,i_2)}{p(\boldsymbol{y}_\ell;\boldsymbol{\theta})}, \tag{14}$$

with $J(m_2, f_1) = J(f_1, m_2)$. The FIM of $T$ iid observations of $\mathsf{Y}$ is given by $T$ times the single-observation FIM.

## 4. ELEMENT-WISE MSE BOUNDS ON PLUG-IN ESTIMATES OF THE PMF

Let $\mathcal{X}_m$, $m \in [M]$ denote an element in $\boldsymbol{\mathcal{X}}$, with a one-to-one mapping $m \leftrightarrow (i_1, i_2, ..., i_N)$, such that $\mathcal{X}_m = \mathcal{X}(i_1, i_2, ..., i_N)$. Given $T$ iid observations of $\mathsf{X}$, the model-free CRB on any unbiased estimate of $\mathcal{X}_m$ is well-known to take the Bernoulli form

$$\text{MSE}[\widehat{\mathcal{X}}_m] \geq \frac{\mathcal{X}_m(1-\mathcal{X}_m)}{T} \overset{\mathcal{X}_m \ll 1}{\approx} \frac{\mathcal{X}_m}{T}. \tag{15}$$

Having found the CCRB on unbiased, constrained estimation of $\boldsymbol{\theta}$ in Section 3, we now proceed to derive a lower bound on the element-wise MSE in estimating $\boldsymbol{\mathcal{X}}$ via a plug-in estimate of $\boldsymbol{\theta}$ in our low-rank model (4), and compare it to the Bernoulli CRB (15). Define

$$\boldsymbol{\phi}(\boldsymbol{\theta}) \triangleq \text{vec}(\boldsymbol{\mathcal{X}}) \in \mathbb{R}^M. \tag{16}$$

Let $\widehat{\boldsymbol{\mathcal{X}}}(\widehat{\boldsymbol{\theta}})$ denote the plug-in estimate of $\boldsymbol{\mathcal{X}}$ using an arbitrary unbiased, constrained estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, and let $\boldsymbol{\phi}(\widehat{\boldsymbol{\theta}}) = \text{vec}(\widehat{\boldsymbol{\mathcal{X}}}(\widehat{\boldsymbol{\theta}}))$ denote its vectorized form. Using a Taylor series expansion of $\boldsymbol{\phi}(\widehat{\boldsymbol{\theta}})$ around the true $\boldsymbol{\theta}$, and assuming small errors, we have

$$\boldsymbol{\phi}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\phi}(\boldsymbol{\theta}) \approx \left.\frac{\partial \boldsymbol{\phi}(\widehat{\boldsymbol{\theta}})}{\partial \widehat{\boldsymbol{\theta}}}\right|_{\widehat{\boldsymbol{\theta}}=\boldsymbol{\theta}} \cdot (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}). \tag{17}$$

Taking the (matrix) second moment we get

$$E\left[\left(\boldsymbol{\phi}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\phi}(\boldsymbol{\theta})\right) \cdot \left(\boldsymbol{\phi}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\phi}(\boldsymbol{\theta})\right)^T\right] \approx \\ \left.\frac{\partial \boldsymbol{\phi}(\widehat{\boldsymbol{\theta}})}{\partial \widehat{\boldsymbol{\theta}}}\right|_{\widehat{\boldsymbol{\theta}}=\boldsymbol{\theta}} \cdot E\left[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \cdot (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\right] \cdot \left.\frac{\partial^T \boldsymbol{\phi}(\widehat{\boldsymbol{\theta}})}{\partial \widehat{\boldsymbol{\theta}}}\right|_{\widehat{\boldsymbol{\theta}}=\boldsymbol{\theta}} \tag{18}$$

which implies the (small errors) MSE bound

$$E\left[\left(\boldsymbol{\phi}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\phi}(\boldsymbol{\theta})\right) \cdot \left(\boldsymbol{\phi}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\phi}(\boldsymbol{\theta})\right)^T\right] \succeq \\ \left.\frac{\partial \boldsymbol{\phi}(\widehat{\boldsymbol{\theta}})}{\partial \widehat{\boldsymbol{\theta}}}\right|_{\widehat{\boldsymbol{\theta}}=\boldsymbol{\theta}} \cdot \boldsymbol{B}_{\boldsymbol{\theta}} \cdot \left.\frac{\partial^T \boldsymbol{\phi}(\widehat{\boldsymbol{\theta}})}{\partial \widehat{\boldsymbol{\theta}}}\right|_{\widehat{\boldsymbol{\theta}}=\boldsymbol{\theta}} \triangleq \boldsymbol{B}_{\mathcal{X}} \in \mathbb{R}^{M \times M}, \tag{19}$$

where for the $m$-th element of $\boldsymbol{\phi}(\boldsymbol{\theta})$ (with the same mapping $m \leftrightarrow (i_1, i_2, ..., i_N)$ used above), we have

$$\left.\frac{\partial \phi_m(\widehat{\boldsymbol{\theta}})}{\partial \hat{\lambda}_f}\right|_{\widehat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = \prod_{n=1}^{N} A_n(i_n, f) \tag{20}$$

$$\left.\frac{\partial \phi_m(\widehat{\boldsymbol{\theta}})}{\partial \hat{A}_{\breve{n}}(i,f)}\right|_{\widehat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = \lambda_f I\{i = i_{\breve{n}}\} \prod_{\substack{n=1 \\ n \neq \breve{n}}}^{N} A_n(i_n, f) \tag{21}$$

## 5. A LOWER BOUND ON THE MEAN KLD

The KLD is a common measure of (dis-)similaritiy between two distributions. We wish to find a lower bound on the expected value of the KLD between the estimated and true PMFs. Under a small-errors assumption, this bound can be expressed using the element-wise bounds in (19) as follows. By definition of the KLD, we have

$$\begin{aligned} D(\boldsymbol{\mathcal{X}}||\widehat{\boldsymbol{\mathcal{X}}}) &= \sum_{m=1}^{M} \mathcal{X}_m \cdot \log\left(\frac{\mathcal{X}_m}{\widehat{\mathcal{X}}_m}\right) \\ &= \sum_{m=1}^{M} \mathcal{X}_m \cdot \left(\log(\mathcal{X}_m) - \log(\widehat{\mathcal{X}}_m)\right). \end{aligned} \tag{22}$$

Defining $\boldsymbol{\mathcal{E}} \triangleq \hat{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}$ as the estimation errors tensor, we have

$$\begin{aligned} \log(\widehat{\mathcal{X}}_m) &= \log(\mathcal{X}_m + \mathcal{E}_m) \\ &= \log(\mathcal{X}_m) + \log\left(1 + \frac{\mathcal{E}_m}{\mathcal{X}_m}\right). \end{aligned} \tag{23}$$

Substituting (23) in (22), we get

$$D(\boldsymbol{\mathcal{X}}||\widehat{\boldsymbol{\mathcal{X}}}) = -\sum_{m=1}^{M} \mathcal{X}_m \cdot \log\left(1 + \frac{\mathcal{E}_m}{\mathcal{X}_m}\right). \tag{24}$$

Using a second-order Taylor series expansion of $\log(1+z) \approx z - \frac{1}{2}z^2$ for $|z| \ll 1$ we get, under a small-errors assumption

$$D(\boldsymbol{\mathcal{X}}||\widehat{\boldsymbol{\mathcal{X}}}) \approx -\sum_{m=1}^{M} \mathcal{X}_m \cdot \left(\frac{\mathcal{E}_m}{\mathcal{X}_m} - \frac{1}{2} \cdot \frac{\mathcal{E}_m^2}{\mathcal{X}_m^2}\right) = \frac{1}{2}\sum_{m=1}^{M} \frac{\mathcal{E}_m^2}{\mathcal{X}_m} \tag{25}$$

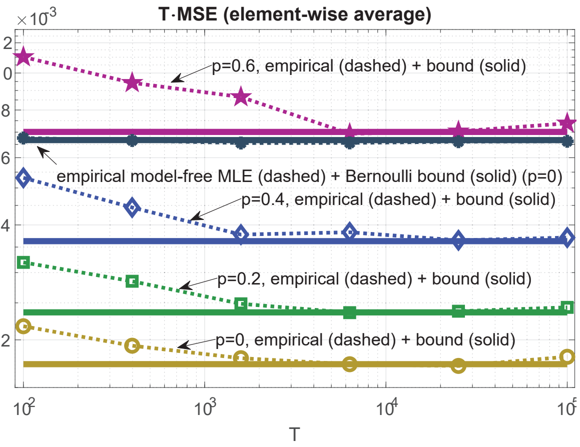

**Fig. 1**. Element-wise $T \cdot$ MSE vs $T$: Bernoulli CRB and our low-rank bound (solid); Empirical MLE-based MSEs (dashed) from 200 independent trials. All based on averaging over all elements of $\boldsymbol{\mathcal{X}}$.

where the last transition is due to $\boldsymbol{\mathcal{X}}$ and $\widehat{\boldsymbol{\mathcal{X}}}$ being PMF tensors whose elements must add up to 1, hence $\sum_{m=1}^{M} \mathcal{E}_m = 0$. Taking the expected value of the KLD we obtain

$$E\left[D(\boldsymbol{\mathcal{X}}||\widehat{\boldsymbol{\mathcal{X}}})\right] \approx \frac{1}{2}\sum_{m=1}^{M}\frac{E\left[\mathcal{E}_m^2\right]}{\mathcal{X}_m} \geq \frac{1}{2}\sum_{m=1}^{M}\frac{B_{\mathcal{X}}(m,m)}{\mathcal{X}_m}, \tag{26}$$

where $B_{\mathcal{X}}(m,m)$ are the diagonal elements of $\boldsymbol{B}_{\mathcal{X}}$ (19), the lower bounds on the MSE of a plug-in estimate of the PMF elements based on any unbiased, constrained estimate of $\boldsymbol{\theta}$.

## 6. SIMULATION RESULTS

To demonstrate the performance bounds on the element-wise MSE and on the mean KLD, we created a small-scale simulation experiment with a rank-2 ($F = 2$) 5-way ($N = 5$) PMF tensor $\boldsymbol{\mathcal{X}}$ of dimensions $[2, 3, 4, 3, 2]$ with arbitrarily chosen $\boldsymbol{A}_1, ..., \boldsymbol{A}_5$ and $\boldsymbol{\lambda}$ as in [4].

Let $\boldsymbol{x}_t$, $t \in [T]$ be the $t$-th observation of $\mathsf{X}$ drawn according to $\boldsymbol{\mathcal{X}}$. The observation $\boldsymbol{y}_t$ of $\mathsf{Y}$ is then obtained by randomly and independently zeroing-out elements of $\boldsymbol{x}_t$ with varying outage probability $p$ (cf. (2)).

Figure 1 shows the Bernoulli CRB (relevant for outage probability $p = 0$ only) and our plug-in low-rank CPD bound (19) (for outage probabilities $p = 0, 0.2, 0.4, 0.6$), averaged over all $M = 144$ elements of $\boldsymbol{\mathcal{X}}$. In addition, we show the MSEs attained by the respective MLE-based estimates, which are asymptotically efficient (we present $T \cdot$ MSE, rather than the MSE itself, in order to keep the differences between the bounds and the empirical MSEs clearly visible as $T$ grows). Note that the model-free (histogram-based) MLE is efficient for all $T$, and therefore coincides with the Bernoulli CRB for all $T$. The advantage of using the low-rank model (when applicable) is evident even in this small-scale example: The zero-outage Bernoulli CRB is more than 150 times larger than the respective zero-outage element-wise bound, and is in fact comparable to the element-wise bound attained at the relatively large outage probability of $p = 0.6$.

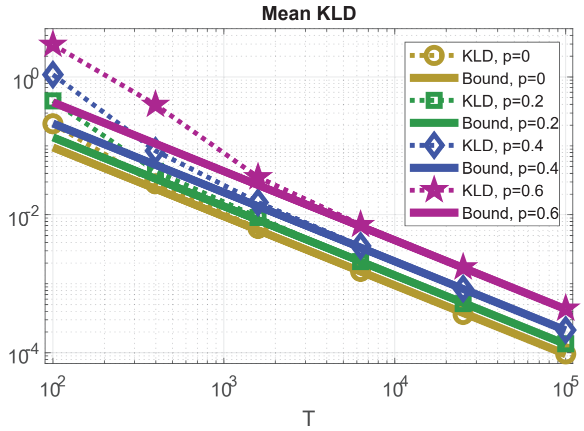

**Fig. 2**. Mean KLD vs $T$: Our bound (solid); Empirical MLE-based KLDs (dashed) averaged over 200 independent trials.

Figure 2 shows our bound (26) on the mean KLD (for $p = 0, 0.2, 0.4, 0.6$), together with the mean empirical KLD between the MLE-based plug-in estimate and the true PMF.

We note that in both figures the plug-in ML estimate asymptotically coincides with the derived performance bounds, which are only accurate for small errors. We can also use these bounds to observe the relation between the outage probability and the sample size needed to attain a certain level of estimation accuracy. Note that this relation is not trivial, as it depends on the dimension $N$ of the RV, as well as on the alphabet size $I_n$ in each dimension.

## 7. CONCLUSION

We derived performance bounds on the estimation of a PMF tensor from partial observations, under the assumption that the tensor admits a low-rank non-negative CPD. Our bounds are based on the CCRB for the CPD parameters (constrained to the *probability simplex*), from which we obtained element-wise MSE bounds for the plug-in estimation error of each element of the tensor, as well as a lower bound on the mean KLD between the estimated and true PMFs. Since the MLE of the model parameters is asymptotically efficient (asymptotics in the sense of $T \to \infty$), all the derived bounds are asymptotically tight (namely, they are asymptotically attainable).

The bounds enable to determine a minimal observation length for attaining prescribed accuracy measures (either in terms of element-wise MSEs or in terms of the mean KLD) in a given estimation problem. They also allow to further explore the non-trivial relation between the outage probability and the attainable accuracy.

Future work would include expansion of the bound to additional accuracy measures, such as the Factor Match Score (FMS, [13]) between the estimated and true CPD parameters, as well as possible extensions of the PMF model or of the outage model.

## 8. REFERENCES

[1] B.-J. Yoon and P. P. Vaidyanathan, ”A multirate DSP model for estimation of discrete probability density functions,” in IEEE Transactions on Signal Processing, vol. 53, no. 1, pp. 252-264, Jan. 2005.

[2] N. Kargas and N. D. Sidiropoulos, ”Completing a joint PMF from projections: A low-rank coupled tensor factorization approach,” in Proc. Information Theory and Applications Workshop (ITA), pp. 1-6, 2017.

[3] N. Kargas, N. D. Sidiropoulos and X. Fu, ”Tensors, Learning, and “Kolmogorov Extension” for Finite-Alphabet Random Vectors,” IEEE Transactions on Signal Processing, vol. 66, no. 18, pp. 4854-4868, 15 Sept.15, 2018.

[4] A. Yeredor and M. Haardt, ”Estimation of a Low-Rank Probability-Tensor from Sample Sub-Tensors via Joint Factorization Minimizing the Kullback-Leibler Divergence,” in Proc. 27th European Signal Processing Conference (EUSIPCO), 2019.

[5] A. Yeredor and M. Haardt, ”Maximum Likelihood Estimation of a Low-Rank Probability Mass Tensor From Partial Observations,” IEEE Signal Processing Letters, vol. 26, no. 10, pp. 1551-1555, Oct. 2019.

[6] M. Amiridi, N. Kargas and N. D. Sidiropoulos, ”Statistical Learning Using Hierarchical Modeling of Probability Tensors,” in Proc. IEEE Data Science Workshop (DSW), pp. 290-294, 2019.

[7] M. Amiridi, N. Kargas and N. D. Sidiropoulos, ”Low-Rank Characteristic Tensor Density Estimation Part I: Foundations,” in IEEE Transactions on Signal Processing, vol. 70, pp. 2654-2668, 2022.

[8] M. Amiridi, N. Kargas and N. D. Sidiropoulos, ”Low-Rank Characteristic Tensor Density Estimation Part II: Compression and Latent Density Estimation,” in IEEE Transactions on Signal Processing, vol. 70, pp. 2669-2680, 2022.

[9] J. Vora, K. S. Gurumoorthy and A. Rajwade, ”Recovery of Joint Probability Distribution from One-Way Marginals: Low Rank Tensors and Random Projections,” in Proc. IEEE Statistical Signal Processing Workshop (SSP), pp. 481-485, 2021.

[10] S. Ibrahim and X. Fu, ”Recovering Joint Probability of Discrete Random Variables From Pairwise Marginals,” IEEE Transactions on Signal Processing, vol. 69, pp. 4116-4131, 2021.

[11] S. ul Haque, A. Rajwade and K. S. Gurumoorthy, ”Joint Probability Estimation Using Tensor Decomposition and Dictionaries,” EUSIPCO 2022, Belgrade, Serbia, 2022, pp. 2226-2230.

[12] D. E. W. Peerlings, J. A. van den Brakel, N. Baştürk and M. J. H. Puts, ”Multivariate Density Estimation by Neural Networks,” in IEEE Transactions on Neural Networks and Learning Systems, 2023 (to appear).

[13] E. C. Chi and T. G. Kolda, “On Tensors, Sparsity, and Nonnegative Factorizations,” SIAM J. Matrix Anal. & Appl., vol. 33, no. 4, pp. 1272–1299, Jan. 2012

[14] P. Stoica and Boon Chong Ng, ”On the Cramer-Rao bound under parametric constraints,” IEEE Signal Processing Letters, vol. 5, no. 7, pp. 177-179, July 1998

[15] S. Kullback and R. A. Leibler, “On Information and Sufficiency”, The Annals of Mathematical Statistics, vol. 22, no. 1, pp. 79–86, 1951.