Under-Counted Tensor Completion with Neural Incorporation of Attributes

Shahana Ibrahim [1]   **Xiao Fu** [1]   **Rebecca Hutchinson** [1]   **Eugene Seo** [2]

Abstract

Systematic under-counting effects are observed in data collected across many disciplines, e.g., epidemiology and ecology. *Under-counted tensor completion* (UC-TC) is well-motivated for many data analytics tasks, e.g., inferring the case numbers of infectious diseases at unobserved locations from under-counted case numbers in neighboring regions. However, existing methods for similar problems often lack supports in theory, making it hard to understand the underlying principles and conditions beyond empirical successes. In this work, a low-rank Poisson tensor model with an expressive unknown *nonlinear* side information extractor is proposed for under-counted multi-aspect data. A joint low-rank tensor completion and neural network learning algorithm is designed to recover the model. Moreover, the UC-TC formulation is supported by theoretical analysis showing that the fully counted entries of the tensor and each entry's under-counting probability can be provably recovered from partial observations—under reasonable conditions. To our best knowledge, the result is the first to offer theoretical supports for under-counted multi-aspect data completion. Simulations and real-data experiments corroborate the theoretical claims.

1. Introduction

Tensor completion (TC) has been a workhorse for a large variety of data analytics tasks, e.g., image/video inpainting (Liu et al., 2013), hyperspectral denoising (Wang et al., 2018), sampling and recovery in magnetic resonance imaging (MRI) (Kanatsoulis et al., 2020), and multimodal data mining (Papalexakis et al., 2017)—just to name a few. In the past decade, theory and methods of TC progressed

significantly—aspects such as recoverability, sample complexity, optimal algorithm design were extensively studied under various tensor models, e.g., *Tucker decomposition* (Mu et al., 2014), *canonical polyadic decomposition* (CPD) (Kanatsoulis et al., 2020; Sørensen & De Lathauwer, 2019), and *block-term decomposition* (BTD) (Zhang et al., 2020a; Ding et al., 2020). The vast majority of classic TC methods were proposed for handling real-valued data, but TC algorithms designed for integer data, e.g., binary data (Ghadermarzy et al., 2018) and count data (Chi & Kolda, 2012), also exist. This is because such data often arise in real-world problems, e.g., movie recommendation (Karatzoglou et al., 2010), knowledge graph completion (Balazevic et al., 2019), and social network analysis (Kashima et al., 2009).

However, less attention has been paid to integer data TC problems where the observed entries suffer from systematical *under-counting*—but (strong) under-counting effects are encountered in many disciplines, e.g., epidemiology and ecology. In epidemiology, the cases of an infectious disease (e.g., COVID-19) may be under-counted due to the existence of symptom-free patients and the lack of testing (Schneider, 2020). In ecology, when an observer sees a species at a certain location, the observer likely just observes a small portion of this species' population (MacKenzie et al., 2006; Tylianakis et al., 2010). Hence, taking under-counting effects into consideration for TC is meaningful in these domains. For example, recovering the "true" (fully counted) number of COVID-19 cases at a certain location and time helps estimate/understand the infection situation. The task of *under-counted tensor completion* (UC-TC) is naturally more challenging than the conventional (integer data) TC, as it implicitly involves an extra task of compensating the *unknown* under-counting effect.

1.1. Prior Work and Challenges

In the literature, there exist some limited efforts towards dealing with under-counted data completion/prediction. For example, the line of work for user exposure matrix completion considered link prediction under undetected links and unobserved links simultaneously (Liang et al., 2016). The work on "zero-inflated" matrix/tensor factorization and related models (Billio et al., 2022; Simchowitz, 2013; Durif et al., 2019) considered a similar scenario, where the datasets' observed entries are zeros with a certain non-zero

---

probability. The recent work by (Fu et al., 2021) proposed an approach to tackle under-counted matrix completion (UC-MC) using a Poisson-Binomial matrix factorization model —which was inspired by the species' abundance modeling in ecology (Joseph et al., 2009; Royle, 2004; Hutchinson et al., 2011).

Challenges. The existing works by (Liang et al., 2016; Billio et al., 2022; Simchowitz, 2013; Durif et al., 2019; Fu et al., 2021) were shown effective on many UC-MC tasks. With proper modifications, these computational frameworks may be generalized to handle tensor data. However, some notable challenges remain. First, how to effectively model the under-counting effect has been an open question. For example, the works by (Liang et al., 2016; Billio et al., 2022; Simchowitz, 2013; Durif et al., 2019) all considered zero-inflated MC models, which dealt with a special case of under-counting, i.e., the observed matrix entries are either the true counts or zeros. For more general under-counted data (i.e., the observed counts are smaller than or equal to the true counts), the recent work by (Fu et al., 2021) proposed to model the probability of missing a count (e.g., an interaction of a pollinator and a plant) as a *linear function* of side attributes, e.g., the location, temperature, and humidity—which is reminiscent of the classic species distribution models (SDMs) (Hutchinson et al., 2011). However, using simple linear functions is a compromise between model expressiveness and computational convenience. Second, the existing works demonstrated their effectiveness only *empirically*. It has been unclear under what conditions the underlying true counts and some key model parameters (e.g., the detection probabilities of the counts) could be *provably* recovered.

1.2. Contributions

In this work, we propose a UC-TC framework based on the Poisson-Binomial model (Joseph et al., 2009; Royle, 2004; Hutchinson et al., 2011) as in the MC work by (Fu et al., 2021). Unlike (Fu et al., 2021) that used a simple side information model and was largely empirical, we address a number of key challenges in modeling, computation, and theory. Our contribution is twofold:

UC-TC with Nonlinear Incorporation of Side Information. To model and learn the under-counting effect in tensor data, we propose a UC-TC framework with neural network-based incorporation of attributes. Specifically, we extend the UC-MC idea in (Fu et al., 2021) to the higher-order tensor regime and model the underlying fully counted data as a Poisson tensor that admits a low-rank *canonical polyadic decomposition* (CPD) (Harshman, 1970); the Poisson tensor is then under-counted via a Binomial detection procedure. Different from the existing work that uses a linear function to model the relationship between the attributes and the

detection probability of a count, we represent the relationship as an *unknown nonlinear function*, which substantially generalizes the modeling capacity. We recast the UC-TC problem as a *maximum likelihood estimation* (MLE) criterion, where the unknown nonlinear function is represented by a neural network (NN) for side information incorporation. We also design a *block coordinate descent* (BCD) algorithm (Razaviyayn et al., 2013) to tackle the formulated MLE effectively.

Recoverability Analysis of UC-TC. We present theoretical characterizations of the proposed UC-TC criterion under the Poisson-Binomial framework. We show that, under reasonable conditions, the fully counted tensor is recoverable, and the associated detection probabilities of all the entries are identifiable—both are up to a global scaling/counter-scaling ambiguity. Our analysis reveals interesting performance-deciding factors such as the similarity of the under-counting effects across some entries. The result also shows that the nonlinear incorporation of side information imposes an intuitive trade-off between model complexity and recovery accuracy, given a fixed amount of samples—which also justifies our proposal of using neural nonlinear models. To our best knowledge, this is the first provable learning criterion for under-counted factor analysis models.

We conduct extensive evaluation of the proposed approach on synthetic data to validate our theoretical findings. We also test our approach on real-world prediction tasks in epidemiology (Zhang et al., 2020b) and ecology (Seo et al., 2022; Daly et al., 2019).

2. Background

2.1. UC-TC: Problem Statement

A Kth-order tensor $\underline{\mathbf{Y}} \in \mathbb{R}^{I_1 \times \cdots \times I_K}$ is a multi-way array whose entries are indexed by K coordinates. An entry of such a tensor can be expressed as follows:

$$y_{\boldsymbol{i}} = [\underline{\mathbf{Y}}]_{\boldsymbol{i}}, \ \boldsymbol{i} = (i_1, \ldots, i_K) \in [I_1] \times \ldots \times [I_K]. \quad (1)$$

Tensors are widely used to represent multi-aspect data. For example, in a four-aspect epidemic data, $y_{\boldsymbol{i}}$ can represent the number of recorded infected cases in city i_1, during week i_2, among age group i_3 and ethnicity group i_4.

TC and MC tasks naturally arise in many applications, where multi-aspect data are only observed partially. A classic use case is recommender systems (Hu et al., 2008; Karatzoglou et al., 2010), where the ratings that users provide to products are used to predict the unseen ratings—which can be done via completing the rating matrix/tensor. TC problems were primarily considered for continuous data (Gandy et al., 2011; Montanari & Sun, 2016; Yuan & Zhang, 2016; Zhang & Aeron, 2017; Sørensen & De Lathauwer, 2019; Zhang et al., 2020a), but TC for other data formats (e.g., binary

and integer data) were also studied in the literature (Ghadermarzy et al., 2018; Chi & Kolda, 2012; Lee & Wang, 2020)—as the latter found many applications in real data analytics problems, e.g., 5-star rating-based recommender systems (Karatzoglou et al., 2010), adjacency network analysis (Acar et al., 2009) and traffic flow prediction (Tan et al., 2016).

In this work, our interest lies in integer TC problems where the partially observed tensor data suffer from systematical under-counting. To be specific, we consider integer tensor data as in (Chi & Kolda, 2012; Lee & Wang, 2020). Like in conventional integer TC problems, the observed tensor $\underline{\boldsymbol{Y}}$ is highly incomplete; i.e., $\boldsymbol{\Omega} \subseteq [I_1] \times \ldots \times [I_K]$ are the indices of the observed entries, where $|\boldsymbol{\Omega}| \ll \prod_{k=1}^{K} I_k$. The key difference from conventional integer TC is that in our setting, the observed entries are smaller than or equal to their actual values, i.e., $y_{\boldsymbol{i}} \leq n_{\boldsymbol{i}}$, $\boldsymbol{i} \in \boldsymbol{\Omega}$, where $n_{\boldsymbol{i}} = [\underline{\boldsymbol{N}}]_{\boldsymbol{i}}$, $\underline{\boldsymbol{N}} \in \mathbb{Z}_{+}^{I_1 \times \ldots \times I_K}$ represents the (unobserved) true-count tensor, and $y_{\boldsymbol{i}} \in \mathbb{Z}_{+}$ is the observed and under-counted version of $n_{\boldsymbol{i}}$.

We assume that for each entry of $\underline{\boldsymbol{Y}}$, there is an associated attribute/feature vector $\boldsymbol{z_i} \in \mathbb{R}^D$, capturing side information regarding the observation. The attributes reflect the *detection probability* of each count in $n_{\boldsymbol{i}}$. We note that such entry attributes are often available in real-world applications. Two examples are as follows:

• **Epidemiology Data**: Consider tensor data in epidemiology that records infected case counts. The count tensor could be indexed by 'city'×'age group'×'profession'. Then, $\boldsymbol{z_i}$ may contain observation-affecting features such as the testing capacity of city i_1, population of age group i_2, and the level of exposure to the virus for profession i_3.

• **Ecology Data**: In ecology, the counts of interactions between pollinators and plants at different times could be represented by a tensor (i.e., a 'pollinator'×'plant'×'time' tensor). There, $\boldsymbol{z_i}$ could contain the properties of the pollinator i_1 and plant i_2, the temperature, humidity and other observation-affecting factors of site i_3.

In this work, we assume that the entry features $\boldsymbol{z_i}$ are available for a subset of entries indexed by $\boldsymbol{\Xi} \subseteq [I_1] \times \ldots \times [I_K]$. Note that $\boldsymbol{\Xi}$ and $\boldsymbol{\Omega}$ need not be identical.

2.2. Existing Work: The Poisson-Binomial MC Model

The work in (Fu et al., 2021) proposed an MC framework with the following generative model for the under-counted entries of a pollinator-plant interaction count matrix:

$$n_{i,j} \sim \mathsf{Poisson}(\lambda_{i,j}), \ \lambda_{i,j} = \boldsymbol{u}_i^\top \boldsymbol{v}_j, \ \boldsymbol{u}_i, \boldsymbol{v}_j \geq \boldsymbol{0}, \quad (2\mathrm{a})$$

$$y_{i,j} \sim \mathsf{Binomial}(n_{i,j}, p_{i,j}), \ p_{i,j} = \boldsymbol{z}_{i,j}^\top \boldsymbol{\theta}, \quad (2\mathrm{b})$$

where $\boldsymbol{u}_i \in \mathbb{R}^F, \boldsymbol{v}_j \in \mathbb{R}^F$ and $\boldsymbol{\theta} \in \mathbb{R}^D$ are the model parameters to estimate. To be more specific, $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ are the F-dimensional latent representations of pollinator i and plant j. The parameter $\lambda_{i,j}$ stands for the average number of interactions between pollinator i and plant j, and the ground-truth number of interactions $n_{i,j}$ is the realization of a Poisson sampling process with parameter $\lambda_{i,j}$. The observed data $y_{i,j}$ is under-counted through a Binomial detection process, with the detection probability $p_{i,j}$. The detection probability $p_{i,j}$ was modeled as a linear function of the entry features $\boldsymbol{z}_{i,j}$. Under this model, $\boldsymbol{\Lambda} = \boldsymbol{U}\boldsymbol{V}^\top$ where $[\boldsymbol{\Lambda}]_{i,j} = \lambda_{i,j}$ is a low-rank matrix model if F is relatively small. Hence, estimating $\boldsymbol{u}_i$'s, $\boldsymbol{v}_j$'s, and $\boldsymbol{\theta}$ can be regarded as a variant of Poisson matrix completion problem.

2.3. Challenges

To extend the Poisson-Binomial model to tensor cases and more general settings, there are a couple of key challenges to be addressed. First, the linear model in (2b), i.e., $p_{i,j} = \boldsymbol{z}_{i,j}^\top \boldsymbol{\theta}$, may not be expressive enough, as the relation between the features and the detection probability is highly likely to be *nonlinear*. The linear model assumption was used for computational convenience, but is a performance-limiting factor in the model of (Fu et al., 2021) (as will be seen in the experiments). Second, perhaps more importantly, there were no theoretical characterizations of the Poisson-Binomial model, despite of its popularity in ecological data analysis (Fu et al., 2021; Hutchinson et al., 2011; Royle, 2004; Dennis et al., 2015). In the context of UC-MC, it is unclear if the underlying Poisson parameters $\lambda_{i,j}$ for all (i, j) could be recovered from a subset of observations $y_{i,j}$, $(i, j) \in \boldsymbol{\Omega}$. The recoverability analysis can not be covered by existing TC theory, as the "extra" task of estimating $p_{i,j}$ makes UC-TC a much harder learning problem compared to the traditional TC problems.

3. Problem Formulation

In this work, we generalize the Poisson-Binomial model to cover tensor data, with *nonlinear* attribute incorporation—and more importantly—with recoverability support.

3.1. Generative Model

We generalize the model (2) by considering under-counted tensors with a size of $I_1 \times \cdots \times I_K$ generated as follows:

$$\lambda_{\boldsymbol{i}} = \sum_{f=1}^{F} \prod_{k=1}^{K} \boldsymbol{U}_k(i_k, f), \ \boldsymbol{U}_k \geq \boldsymbol{0}, \forall k \in [K], \quad (3\mathrm{a})$$

$$n_{\boldsymbol{i}} \sim \mathsf{Poisson}(\lambda_{\boldsymbol{i}}), \quad (3\mathrm{b})$$

$$p_{\boldsymbol{i}} = g\left(\boldsymbol{z_i}\right), \ 0 \leq p_{\boldsymbol{i}} \leq 1, \quad (3\mathrm{c})$$

$$y_{\boldsymbol{i}} \sim \mathsf{Binomial}(n_{\boldsymbol{i}}, p_{\boldsymbol{i}}), \quad (3\mathrm{d})$$

where $\boldsymbol{i} = (i_1, \ldots, i_K)$ is a shorthand notation as defined before, $\boldsymbol{U}_k(i_k, :) \in \mathbb{R}^F$ denotes the F-dimensional latent representation of entity i_k of aspect k, $y_{\boldsymbol{i}}$ denotes the *observed* count indexed by $\boldsymbol{i}$, $n_{\boldsymbol{i}}$ is the corresponding true count, $\lambda_{\boldsymbol{i}}$ stands for the Poisson parameter (which is the expectation of the true count), and $p_{\boldsymbol{i}}$ represents the detection probability as in the model of (2). The vector $\boldsymbol{z}_{\boldsymbol{i}} \in \mathbb{R}^D$ collects the features of entry $\boldsymbol{i}$. The *unknown* nonlinear function $g(\cdot) : \mathbb{R}^D \to [0, 1]$ is used to model the dependence between the detection probability $p_{\boldsymbol{i}}$ and $\boldsymbol{z}_{\boldsymbol{i}}$. In (3), we have imposed nonnegativity constraints on $\boldsymbol{U}_k$'s in order to make sure that the Poisson parameters are all nonnegative.

Given $y_{\boldsymbol{i}}$ for $\boldsymbol{i} \in \boldsymbol{\Omega} \subset [I_1] \times \ldots \times [I_K]$ and $\boldsymbol{z}_{\boldsymbol{i}}$ for $\boldsymbol{i} \in \boldsymbol{\Xi} \subset [I_1] \times \ldots \times [I_K]$, our goal is to estimate $\lambda_{\boldsymbol{i}}$ and $p_{\boldsymbol{i}}$ for all $\boldsymbol{i}$.

3.2. Maximum Likelihood Estimation-Based UC-TC

Assuming that $y_{\boldsymbol{i}}$'s are independently sampled from the generative model in (3), the log-likelihood of the observations can be expressed as follows:

$$
\begin{aligned}
&\log \prod_{\boldsymbol{i} \in \boldsymbol{\Omega}} \mathsf{Pr}(y_{\boldsymbol{i}}; \lambda_{\boldsymbol{i}}, p_{\boldsymbol{i}}) \\
&= \log \left(\prod_{\boldsymbol{i} \in \boldsymbol{\Omega}} \sum_{n=y_{\boldsymbol{i}}}^{\infty} \mathsf{Pr}(N_{\boldsymbol{i}} = n; \lambda_{\boldsymbol{i}}) \mathsf{Pr}(y_{\boldsymbol{i}} | N_{\boldsymbol{i}} = n; p_{\boldsymbol{i}}) \right) \\
&= \sum_{\boldsymbol{i} \in \boldsymbol{\Omega}} \log \left(\sum_{n=y_{\boldsymbol{i}}}^{\infty} \left(\frac{\lambda_{\boldsymbol{i}}^n e^{-\lambda_{\boldsymbol{i}}}}{n!} \frac{n! p_{\boldsymbol{i}}^{y_{\boldsymbol{i}}} (1-p_{\boldsymbol{i}})^{n-y_{\boldsymbol{i}}}}{y_{\boldsymbol{i}}!(n-y_{\boldsymbol{i}})!} \right) \right) \\
&= \sum_{\boldsymbol{i} \in \boldsymbol{\Omega}} [y_{\boldsymbol{i}} \log \lambda_{\boldsymbol{i}} + y_{\boldsymbol{i}} \log p_{\boldsymbol{i}} - \lambda_{\boldsymbol{i}} p_{\boldsymbol{i}} - \log y_{\boldsymbol{i}}!] ; \quad (4)
\end{aligned}
$$

see the detailed derivation in the supplementary material in Sec. B. Under the generative model in (3), we have $p_{\boldsymbol{i}} = g(\boldsymbol{z}_{\boldsymbol{i}})$. Since $g(\cdot)$ is unknown and nonlinear, we represent it using a neural network denoted as $g_{\boldsymbol{\theta}} : \mathbb{R}^D \to [0, 1]$:

$$
g_{\boldsymbol{\theta}}(\boldsymbol{z}) = \sigma(\boldsymbol{w}_L^\top \boldsymbol{\sigma}(\boldsymbol{W}_{L-1} \boldsymbol{\sigma}(\ldots \boldsymbol{\sigma}(\boldsymbol{W}_1 \boldsymbol{z})))), \quad (5)
$$

where $\boldsymbol{W}_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ denotes the weight matrix at ℓth layer ($d_0 = D$), $\boldsymbol{w}_L \in \mathbb{R}^{d_{L-1}}$ denotes the last layer's weights, $\boldsymbol{\theta}$ denotes the parameters of the NN ($\{\boldsymbol{W}_\ell\}_{\ell=1}^{L-1}, \boldsymbol{w}_L$), and $\boldsymbol{\sigma}(\cdot) = [\sigma(\cdot), \ldots, \sigma(\cdot)]^\top$ denotes the activation function with a proper dimension. Hence, combining (4) with (3a) and (3c), we formulate the MLE criterion as follows:

$$
\begin{aligned}
\underset{\{\boldsymbol{U}_k\}, \boldsymbol{\theta}, \{p_{\boldsymbol{i}}\}}{\text{minimize}} \; L(\boldsymbol{U}, \boldsymbol{\theta}, \underline{\boldsymbol{P}}) &:= \sum_{\boldsymbol{i} \in \boldsymbol{\Omega}} \left[\left(\sum_{f=1}^{F} \prod_{k=1}^{K} \boldsymbol{U}_k(i_k, f) \right) p_{\boldsymbol{i}} \right. \\
&\left. - y_{\boldsymbol{i}} \log \left(\sum_{f=1}^{F} \prod_{k=1}^{K} \boldsymbol{U}_k(i_k, f) \right) - y_{\boldsymbol{i}} \log p_{\boldsymbol{i}} \right], \quad (6a)
\end{aligned}
$$

$$
\text{subject to} \; \boldsymbol{U}_k \in \mathcal{U}_k, \; \forall k \in [K], \quad (6b)
$$

$$
p_{\boldsymbol{i}} = g_{\boldsymbol{\theta}}(\boldsymbol{z}_{\boldsymbol{i}}), \; \forall \boldsymbol{i} \in \boldsymbol{\Xi}, \; g_{\boldsymbol{\theta}} \in \mathcal{G}, \quad (6c)
$$

where $\mathcal{U}_k$ is a constraint set that encodes the prior knowledge of $\boldsymbol{U}_k$ (e.g., nonnegativity), and $\mathcal{G}$ represents the function class where $g_{\boldsymbol{\theta}}$ is taken from. Given the formulated MLE, there are a couple of key aspects that are of interest. First, if one solves the formulated learning criterion in (6), can the *optimal* solution recover the latent parameters $\lambda_{\boldsymbol{i}}$ and $p_{\boldsymbol{i}}$ over all $\boldsymbol{i}$? This question will be answered in Sec. 4. Second, the criterion (6) presents a challenging optimization problem that involves tensor decomposition and neural network learning—both are nontrivial problems. A BCD algorithm to effectively tackle this problem will be proposed in Sec. 5.

4. Recoverability Analysis

Our goal is to show that the optimal solution given by (6) provably recovers $\lambda_{\boldsymbol{i}}$'s and $p_{\boldsymbol{i}}$'s under the generative model (3) (up to certain reasonable ambiguities).

4.1. Analysis Setup

To better present the analysis, we use "$\natural$" to denote the ground-truth parameters in the model (3)—e.g., $\lambda_{\boldsymbol{i}}^\natural$ and $p_{\boldsymbol{i}}^\natural$ denote the ground-truth Poisson parameter and ground-truth detection probability of entry $\boldsymbol{i}$, respectively. We have the corresponding tensor notations $[\underline{\boldsymbol{\Lambda}}^\natural]_{\boldsymbol{i}} = \lambda_{\boldsymbol{i}}^\natural$ and $[\underline{\boldsymbol{P}}^\natural]_{\boldsymbol{i}} = p_{\boldsymbol{i}}^\natural$. The "hat" notation is used to denote the terms constructed from the optimal solution given by (6). For instance, $\{\widehat{\boldsymbol{U}}_k\}_{k=1}^K$, $\widehat{g}_{\boldsymbol{\theta}}$ and $\widehat{p}_{\boldsymbol{i}}, \forall \boldsymbol{i}$ denote the estimates given by the solution obtained from solving (6). We consider the following assumptions:

Assumption 4.1 (**Bounded Parameters**). There exist scalars $p_{\min}, p_{\max}, \beta_u, \alpha_u > 0$ such that $p_{\boldsymbol{i}}^\natural = g^\natural(\boldsymbol{z}_{\boldsymbol{i}})$, $0 < p_{\min} \leq p_{\boldsymbol{i}}^\natural \leq p_{\max}$, $\forall \boldsymbol{i}$ and $0 < \beta_u \leq \boldsymbol{U}_k^\natural(i_k, f) \leq \alpha_u$, $\forall i_k \in [I_k], f \in [F], k \in [K]$. In addition, the constraints in (6), namely, $\mathcal{U}_k$ and $\mathcal{G}$ satisfy $\mathcal{U}_k = \{\boldsymbol{U} \in \mathbb{R}^{I_k \times F} \mid \beta_u \leq \boldsymbol{U}(i_k, f) \leq \alpha_u, \forall i_k, f\}$ and $\mathcal{G} = \{g_{\boldsymbol{\theta}} : \mathbb{R}^D \to [p_{\min}, p_{\max}]\}$, respectively.

Assumption 4.2 (**Approximation Error**). Assume that there exists $\widetilde{g}_{\boldsymbol{\theta}} \in \mathcal{G}$ such that $|\widetilde{g}_{\boldsymbol{\theta}}(\boldsymbol{z}) - g^\natural(\boldsymbol{z})| \leq \nu$ for all $\boldsymbol{z} \in \mathbb{R}^D$, where $0 \leq \nu < \infty$. In addition, assume that the class $\mathcal{G}$ has a complexity measure $\mathscr{R}_{\mathcal{G}}$.

Assumption 4.3 (**Similar Attribute Subset**). There exists an index set $\boldsymbol{\Theta}$ whose elements are sampled uniformly at random (without replacement) from $[I_1] \times \ldots \times [I_K]$ and a scalar $\zeta > 0$ such that $\max_{\boldsymbol{i}, \boldsymbol{j} \in \boldsymbol{\Theta}} \|\boldsymbol{z}_{\boldsymbol{i}} - \boldsymbol{z}_{\boldsymbol{j}}\|_2 \leq \zeta$.

Assumption 4.4 (**Lipschitz Continuity**). Assume that the ground-truth function $g^\natural$ and any function $g_{\boldsymbol{\theta}} \in \mathcal{G}$ are Lipschitz continuous, i.e., for certain $L_g, L_{\boldsymbol{\theta}} > 0$ and for any pair of $\boldsymbol{z}_{\boldsymbol{i}}, \boldsymbol{z}_{\boldsymbol{j}}$, $|g^\natural(\boldsymbol{z}_{\boldsymbol{i}}) - g^\natural(\boldsymbol{z}_{\boldsymbol{j}})| \leq L_g \|\boldsymbol{z}_{\boldsymbol{i}} - \boldsymbol{z}_{\boldsymbol{j}}\|_2$, and $|g_{\boldsymbol{\theta}}(\boldsymbol{z}_{\boldsymbol{i}}) - g_{\boldsymbol{\theta}}(\boldsymbol{z}_{\boldsymbol{j}})| \leq L_{\boldsymbol{\theta}} \|\boldsymbol{z}_{\boldsymbol{i}} - \boldsymbol{z}_{\boldsymbol{j}}\|_2$ hold.

Assumption 4.2 requires that the neural network class $\mathcal{G}$ contains a function $\widetilde{g}_{\boldsymbol{\theta}}$ that is closer to the ground-truth func-

tion $g^\natural$. In practice, if deeper/wider neural networks are employed, the value of ν is smaller. For the complexity measure $\mathscr{R}_\mathcal{G}$ used in Assumption 4.2, we use the *sensitive complexity* parameter introduced in (Lin & Zhang, 2019). For deeper and wider neural network function classes, the parameter $\mathscr{R}_\mathcal{G}$ gets larger—also see supplementary material Sec. M for more details. Assumptions 4.3-4.4 together imply that there exists a set of p_i's that are similar. For example, in an epidemic dataset (e.g., the COVID-19 dataset (Zhang et al., 2020b)) that records the number of detected infectious disease cases, when the testing capacity and the population of a number of locations are similar with each other, it is reasonable to assume that the corresponding detection probabilities p_i's are similar. The Lipschitz continuity assumption on g_θ given by Assumption 4.4 requires that the learned g_θ is smooth enough, which can be achieved via bounding (or penalizing) the norm of the neural network parameter θ during the learning process.

4.2. Main Result

Our main result is as follows:

Theorem 4.5. *Suppose that the Assumptions 4.1-4.4 hold true. Assume that the entries of index sets Ω and Ξ are sampled from $[I_1] \times \ldots \times [I_K]$ uniformly at random with replacement and satisfies $\Omega \subseteq \Xi$. In addition, let $z_i, \ldots, z_{i_S}$ be the set of observed features. Assume that each z_i is drawn from a distribution $\mathcal{D}$. Then, for $\delta > 0$, the following hold with probability of at least $1 - 5\delta - 3e^{-\alpha(e^2-3)}$:*

$$\frac{\left\| \Delta^\natural - \widehat{\xi}\widehat{\Delta} \right\|_F^2}{\prod_k I_k} \leq \mathcal{O}\left(\frac{1}{p_{\min}^2} \varrho_1 \right), \tag{7a}$$

$$\mathbb{E}_{z_i \sim \mathcal{D}}\left[(p_i^\natural - {}^1\!/\!\widehat{\xi}\widehat{g}_\theta(z_i))^2 \right] \leq \mathcal{O}\left(\frac{p_{\max}}{\beta p_{\min}^2} \varrho_1 + \varrho_2 \right), \tag{7b}$$

where $\varrho_1 = F\alpha_u^K \left(\eta + \zeta^2(L_g + L_\theta)^2 + \sqrt{\frac{\log(1/\delta)}{|\Theta|}} \right)$,

$$\varrho_2 = \sqrt{\frac{\log(1/\delta)}{S}} + \frac{(\|Z\|_F \mathscr{R}_\mathcal{G})^{1/4}}{\sqrt{S}},$$

$$\eta^2 = \frac{K\sum_k I_k + \|Z\|_F \mathscr{R}_\mathcal{G}}{\sqrt{T}} + c\sqrt{\frac{\log(1/\delta)}{T}} + \alpha\nu,$$

$c = \alpha\max\{|\log\beta|, \log\alpha\}$, $\alpha = F\alpha_u^K p_{\max}$, $\beta = F\beta_u^K p_{\min}$, $Z = [z_{i_1}, \ldots, z_{i_S}] \in \mathbb{R}^{D \times S}$, $T = |\Omega|$, *and* $\widehat{\xi}$ *denote the global scaling ambiguity.*

The proof of Theorem 4.5 is relegated to the supplementary material in Sec. D. The result reveals several interesting insights. First, the results in (7) show that the proposed MLE criterion can correctly recover the average true counts $\lambda_i^\natural$'s and the detection probabilities $p_i^\natural$'s, up to a certain global scaling ambiguity between the entries. Second, the bounds

indicate that the estimation accuracy is better when there are more observations (i.e., larger $|\Omega|$) and more similar features (i.e., smaller ζ and larger $|\Theta|$). Third, the result suggests that an appropriate choice of the neural network class plays a role in ensuring good estimation accuracy. This is because if one uses deeper/wider neural networks, the approximation error ν becomes smaller, but the complexity paramater $\mathscr{R}_\mathcal{G}$ is larger. Hence, there is a trade-off to strike while choosing the network architecture of g_θ.

5. Algorithm Design

The proposed MLE in (6) is a combination of Poisson tensor decomposition and neural network learning—both are hard optimization problems. To tackle this problem, we design a BCD-based (Razaviyayn et al., 2013; Bertsekas, 1999) algorithm. We first re-formulate (6) using a regularized form by considering nonnegativity constraints for U_k's:

$$\underset{\{U_k\},\theta,\{p_i\}}{\text{minimize}} \frac{1}{|\Omega|} L(U, \theta, \underline{P}) + \frac{\mu}{|\Xi|} \sum_{i \in \Xi} \ell(g_\theta(z_i), p_i), \tag{8a}$$

$$\text{subject to } U_k \geq 0, \forall k \in [K], \tag{8b}$$

$$0 \leq p_i \leq 1, \ \forall i \in \Omega \cup \Xi, \ g_\theta \in \mathcal{G}, \tag{8c}$$

where $\mu > 0$ is a regularization parameter and $\ell(\cdot, \cdot)$ denotes a certain distance/divergence measure, e.g., the least squares function $\ell(x, y) = (x - y)^2$.

Note that in the re-formulated problem (8), we did not explicitly use the bounds on $\mathcal{U}_k$ and $\mathcal{G}$ in Assumption 4.1. The reason is twofold: First, the bounds serve for analytical purposes, yet not easy to know exactly in practice. Second, ignoring the bounds is inconsequential in terms of performance (as will be seen in experiments), but substantially simplifies the algorithm design. Nonetheless, as one will see, the update of U_k naturally results in positive and bounded solutions, under reasonable conditions (cf. the comments after Eq. 10). In addition, as g_θ can be constructed to have a sigmoid output layer, the value of $g_\theta(z_i)$ is naturally bounded away from 0 and 1.

Also note that we adopt this regularized optimization design since it gives flexibility to handle different cases such as $\Omega \subseteq \Xi$ and $\Omega \supset \Xi$—as we will see in the following section. Our BCD updates are designed as follows:

The U_k-Subproblem. When θ, p_i, U_j, $j \neq k$ are fixed, the subproblem w.r.t. U_k is given by

$$\underset{U_k \geq 0}{\text{minimize}} \sum_{i \in \Omega} \left[\left(\sum_{f=1}^F U_k(i_k, f) \prod_{j \neq k} U_j(i_j, f) \right) p_i \right.$$

$$\left. - y_i \log\left(\sum_{f=1}^F U_k(i_k, f) \prod_{j \neq k} U_j(i_j, f) \right) \right]. \tag{9}$$

From (9), the update for $\boldsymbol{U}_k$ is as follows:

$$\boldsymbol{U}_k(i_k, f) \leftarrow \frac{\sum_{\boldsymbol{i} \in \boldsymbol{\Omega}} y_{\boldsymbol{i}} \alpha_{\boldsymbol{i}}^{(f)}}{\sum_{\boldsymbol{i} \in \boldsymbol{\Omega}} \prod_{j \neq k} \boldsymbol{U}_j(i_j, f) p_{\boldsymbol{i}}}, \; \forall \boldsymbol{i}, f, \quad (10)$$

where $\alpha_{\boldsymbol{i}}^{(f)} = \frac{\overline{\boldsymbol{U}}_k(i_k, f) \prod_{j \neq k} \boldsymbol{U}_j(i_j, f)}{\sum_{f=1}^{F} \overline{\boldsymbol{U}}_k(i_k, f) \prod_{j \neq k} \boldsymbol{U}_j(i_j, f)}$ and $\overline{\boldsymbol{U}}_k$ denotes $\boldsymbol{U}_k$ from the previous iteration. Note that the above solution is always strictly positive if $\boldsymbol{U}_k > \boldsymbol{0}$ for all k and $p_{\boldsymbol{i}} > 0$ for all $\boldsymbol{i}$. Our update rule design for $\boldsymbol{U}_k$ follows the ideas in (Chi & Kolda, 2012; Fu et al., 2021), which is essentially a *majorization minimization* step w.r.t. $\boldsymbol{U}_k$; see Sec. C in the supplementary material.

The p-Subproblem. To update $p_{\boldsymbol{i}}$, we fix $\boldsymbol{U}_k$'s and $\boldsymbol{\theta}$ and consider different cases.

Case 1. When both $y_{\boldsymbol{i}}$ and $\boldsymbol{z}_{\boldsymbol{i}}$ are available, i.e., $\boldsymbol{i} \in \boldsymbol{\Omega} \cap \boldsymbol{\Xi}$, $p_{\boldsymbol{i}}$ is updated via optimizing the following subproblem:

$$\min_{p_{\boldsymbol{i}} \in [0,1]} \frac{1}{|\boldsymbol{\Omega}|} (\lambda_{\boldsymbol{i}} p_{\boldsymbol{i}} - y_{\boldsymbol{i}} \log(p_{\boldsymbol{i}})) + \frac{\mu}{|\boldsymbol{\Xi}|} \ell(g_{\boldsymbol{\theta}}(\boldsymbol{z}_{\boldsymbol{i}}), p_{\boldsymbol{i}}), \quad (11)$$

where $\lambda_{\boldsymbol{i}} = \sum_{f=1}^{F} \prod_{k=1}^{K} \boldsymbol{U}_k(i_k, f)$ is obtained using the current estimates of $\boldsymbol{U}_k$'s.

Case 2. Suppose that the observation $y_{\boldsymbol{i}}$ is not available, but $\boldsymbol{z}_{\boldsymbol{i}}$ is available, i.e., $\boldsymbol{i} \in \boldsymbol{\Xi} - \boldsymbol{\Omega} \cap \boldsymbol{\Xi}$. Then, we consider the following subproblem to update such $p_{\boldsymbol{i}}$'s:

$$\min_{p_{\boldsymbol{i}} \in [0,1]} \ell(g_{\boldsymbol{\theta}}(\boldsymbol{z}_{\boldsymbol{i}}), p_{\boldsymbol{i}}). \quad (12)$$

If we choose a convex $\ell(\cdot)$ in (11)-(12), the problems can be optimally solved using any standard techniques such as projected gradient decent. Moreover, if we set $\ell(\cdot)$ to be one of the popular choices such Euclidean distance and KL divergence, closed-form solutions for the problems in (11)-(12) exist—see Table 6 in the supplementary material.

Case 3.[1] When $y_{\boldsymbol{i}}$ is observed, but $\boldsymbol{z}_{\boldsymbol{i}}$ is not accessible, i.e., $\boldsymbol{i} \in \boldsymbol{\Omega} - \boldsymbol{\Omega} \cap \boldsymbol{\Xi}$, we update $p_{\boldsymbol{i}}$ via solving the following subproblem:

$$\min_{p_{\boldsymbol{i}} \in (0,1]} \lambda_{\boldsymbol{i}} p_{\boldsymbol{i}} - y_{\boldsymbol{i}} \log(p_{\boldsymbol{i}}), \quad (13)$$

which gives the update rule for $p_{\boldsymbol{i}}$ as $p_{\boldsymbol{i}} \leftarrow [y_{\boldsymbol{i}}/\lambda_{\boldsymbol{i}}]_{[0,1]}$.

The $\boldsymbol{\theta}$-Subproblem. By fixing the $\boldsymbol{U}_k$'s and all the $p_{\boldsymbol{i}}$'s, the subproblem for updating $\boldsymbol{\theta}$ is as follows:

$$\min_{\boldsymbol{\theta}} \sum_{\boldsymbol{i} \in \boldsymbol{\Xi}} \ell(g_{\boldsymbol{\theta}}(\boldsymbol{z}_{\boldsymbol{i}}), p_{\boldsymbol{i}}).$$

---

[1] Our recoverability analysis is built upon the assumption $\boldsymbol{\Omega} \subseteq \boldsymbol{\Xi}$; see Theorem 4.5. However, in practice, there are cases where $y_{\boldsymbol{i}}$ is observed, but $\boldsymbol{z}_{\boldsymbol{i}}$ is not available (i.e., $\boldsymbol{\Omega} - \boldsymbol{\Omega} \cap \boldsymbol{\Xi} \neq \emptyset$). Our algorithm can still operate under such cases.

This is an unconstrained neural network learning problem. Many off-the-shelf neural network learning algorithms that use gradient, e.g., `Adam` (Kingma & Ba, 2015) and `Adagrad` (Duchi et al., 2011), can be used to handle this subproblem.

More design details and the description of the algorithm are provided in supplementary material in Sec. C. The proposed algorithm is referred to as the *Under-Counted Data Prediction Via Learner-Aided Tensor Completion* (`UncleTC`) algorithm.

A remark is that our algorithm uses a batch (instead of stochastic) update rule for the $\boldsymbol{U}_k$'s. This can also be replaced by stochastic tensor decomposition algorithms (see, e.g., (Fu et al., 2020a; Pu et al., 2022) and the references in (Fu et al., 2020b)). Using stochastic algorithms for the $\boldsymbol{U}_k$-subproblems may further improve efficiency.

6. Experiments

In this section, we evaluate our proposed method through a series of synthetic and real data experiments. The source code is available at `https://github.com/shahanaibrahimosu/ undercounted-tensor-completion`.

Baselines. We employ a number of low-rank tensor completion-based baselines, namely `NTF-CPD-KL`(Chi & Kolda, 2012), `HaLRTC` (Liu et al., 2013), `NTF-CPD-LS` (Shashua & Hazan, 2005), `BPTF-CPD` (Schein et al., 2015), and `NTF-Tucker-LS` (Kim & Choi, 2007). Among them, both `NTF-CPD-KL` and `BPTF` consider Poisson modeling, but they do not have a Binomial detection stage in their models to accommodate the under-counting effect. We also consider two recent NN-based tensor completion methods, `CostCo` (Liu et al., 2019) and `POND` (Tillinghast et al., 2020), as baselines. In order to incorporate the side features, the `CostCo` method is slightly modified from its original implementation by concatenating side features with the indices of the observed entries as the input to their networks. In addition, we also include a tensor version of (Fu et al., 2021), where a linear function $g_{\boldsymbol{\theta}}(\boldsymbol{z}_{\boldsymbol{i}}) = \boldsymbol{\theta}_1^{\top} \boldsymbol{z}_{\boldsymbol{i}} + \theta_2$ is used to learn g function. This baseline is referred to as `UncleTC(Linear)`. More details on the implementation are provided in the supplementary material in Sec. N.

6.1. Synthetic-Data Simulations

Data Generation. The data generating process follows (3). We first generate $\boldsymbol{U}_k \in \mathbb{R}^{I_k \times F}, \; \forall k$ by randomly sampling its entries from a uniform distribution between 0 and κ. We fix $K = 3$, $F = 3$ and $I_k = 20$, $\forall k$, unless specified otherwise. We also fix $\kappa = 15$ so that $\lambda_{\boldsymbol{i}}$'s are not unreasonably small. The feature vectors $\boldsymbol{z}_{\boldsymbol{i}} \in \mathbb{R}^D$

Table 1. Average MAE_p and MAE_λ over 20 random trials under various values of γ_Ω; $\gamma_\Xi = 0.3, \gamma_\Theta = 0.2, \text{SNR} = 40\text{dB}$ with $\Theta \cup \Omega \subseteq \Xi$.

Algorithm	Metric	$\gamma_\Omega = 0.05$	$\gamma_\Omega = 0.1$	$\gamma_\Omega = 0.3$
UncleTC	MAE_p	$\mathbf{0.156 \pm 0.055}$	$\mathbf{0.111 \pm 0.038}$	$\mathbf{0.063 \pm 0.047}$
	MAE_λ	$\mathbf{0.104 \pm 0.063}$	$\mathbf{0.063 \pm 0.041}$	$\mathbf{0.037 \pm 0.031}$
UncleTC(Linear)	MAE_p	0.231 ± 0.072	0.189 ± 0.046	0.193 ± 0.180
	MAE_λ	0.135 ± 0.069	0.092 ± 0.039	0.090 ± 0.089
NTF-CPD-KL	MAE_λ	0.306 ± 0.119	0.258 ± 0.128	0.190 ± 0.124
BPTF	MAE_λ	0.547 ± 0.143	0.493 ± 0.161	0.345 ± 0.104
HaLRTC	MAE_λ	0.503 ± 0.382	0.495 ± 0.391	0.436 ± 0.282
NTF-LS	MAE_λ	0.742 ± 0.462	0.684 ± 0.421	0.403 ± 0.099
NTF-Tucker-LS	MAE_λ	0.639 ± 0.494	0.526 ± 0.299	0.341 ± 0.072
CostCo	MAE_λ	0.964 ± 0.235	0.853 ± 0.213	0.809 ± 0.147
POND	MAE_λ	0.841 ± 0.146	0.877 ± 0.098	0.885 ± 0.087

Table 2. Average MAE_p and MAE_λ over 20 random trials under various values of ζ; $\gamma_\Omega = 0.2, \gamma_\Xi = 0.3, \gamma_\Theta = 0.2$ with $\Theta \cup \Omega \subseteq \Xi$.

Algorithm	Metric	$\text{SNR} = 0\text{dB}$	$\text{SNR} = 10\text{dB}$	$\text{SNR} = 40\text{dB}$
UncleTC	MAE_p	$\mathbf{0.154 \pm 0.007}$	$\mathbf{0.047 \pm 0.018}$	$\mathbf{0.035 \pm 0.001}$
	MAE_λ	$\mathbf{0.089 \pm 0.010}$	$\mathbf{0.026 \pm 0.007}$	$\mathbf{0.023 \pm 0.001}$
UncleTC(Linear)	MAE_p	0.304 ± 0.012	0.109 ± 0.041	0.129 ± 0.049
	MAE_λ	0.193 ± 0.054	0.052 ± 0.016	0.065 ± 0.020
NTF-CPD-KL	MAE_λ	0.299 ± 0.055	0.111 ± 0.001	0.107 ± 0.001
BPTF	MAE_λ	0.420 ± 0.051	0.414 ± 0.047	0.427 ± 0.070
HaLRTC	MAE_λ	0.423 ± 0.068	0.441 ± 0.109	0.440 ± 0.126
NTF-LS	MAE_λ	0.479 ± 0.062	0.474 ± 0.057	0.472 ± 0.046
NTF-Tucker-LS	MAE_λ	0.366 ± 0.047	0.375 ± 0.050	0.378 ± 0.045
CostCo	MAE_λ	0.987 ± 0.256	0.876 ± 0.432	0.890 ± 0.357
POND	MAE_λ	0.824 ± 0.096	0.785 ± 0.131	0.837 ± 0.087

Table 3. Average MAE_p and MAE_λ over 20 random trials under various values of γ_Θ; $\gamma_\Omega = 0.2, \gamma_\Xi = 0.7, \text{SNR} = 40\text{dB}$ with $\Theta \cup \Omega \subseteq \Xi$.

Algorithm	Metric	$\gamma_\Theta = 0$	$\gamma_\Theta = 0.3$	$\gamma_\Theta = 0.5$
UncleTC	MAE_p	$\mathbf{0.163 \pm 0.007}$	$\mathbf{0.126 \pm 0.044}$	$\mathbf{0.104 \pm 0.076}$
	MAE_λ	$\mathbf{0.062 \pm 0.003}$	$\mathbf{0.055 \pm 0.028}$	$\mathbf{0.054 \pm 0.148}$
UncleTC(Linear)	MAE_p	0.231 ± 0.005	0.252 ± 0.029	0.201 ± 0.081
	MAE_λ	0.078 ± 0.005	0.104 ± 0.041	0.214 ± 0.174
NTF-CPD-KL	MAE_λ	0.185 ± 0.006	0.107 ± 0.002	0.156 ± 0.078
BPTF	MAE_λ	0.381 ± 0.024	0.445 ± 0.092	0.473 ± 0.174
HaLRTC	MAE_λ	0.436 ± 0.030	0.533 ± 0.212	0.561 ± 0.352
NTF-LS	MAE_λ	0.434 ± 0.030	0.472 ± 0.068	0.644 ± 0.412
NTF-Tucker-LS	MAE_λ	0.345 ± 0.039	0.411 ± 0.080	0.564 ± 0.337
CostCo	MAE_λ	0.876 ± 0.178	0.814 ± 0.301	0.987 ± 0.239
POND	MAE_λ	0.926 ± 0.158	0.851 ± 0.165	0.859 ± 0.112

with $D = 10$ are generated by randomly sampling its entries from the standard normal distribution. We use the ground-truth nonlinear function (unknown to our algorithm) $g(z) = \text{sigmoid}(\nu^\top(0.5z^3 + 0.2z))$, where the vector $\nu \in \mathbb{R}^D$ is generated by randomly sampling its entries from the uniform distribution between 0 and 1. We observe only a subset of y_i's such that each i belongs to Ω with probability $\gamma_\Omega \in (0, 1]$. Similarly, only a portion of z_i's are observed such that each i belongs to Ξ with probability $\gamma_\Xi \in (0, 1]$. We also make a subset of z_i's similar such that $z_i = \varphi + n_i$, $i \in \Theta$, $\Theta \subseteq \Xi$, with each i belongs to Θ with probability $\gamma_\Theta \in (0, 1]$. Here, $\varphi \in \mathbb{R}^D$ is a random vector whose entries are drawn from the standard normal distribution and n_i denotes additive noise whose entries are zero mean Gaussian with variance σ^2. Under this setting, we define the *signal-to-noise* (SNR) ratio $\text{SNR} = 10 \log_{10}\left(\|\varphi\|_2^2/D\sigma^2\right)\text{dB}$. As the value of SNR increases, the feature vectors $z_i \in \Theta$ are more similar.

Metric. To characterize the performance of recovering p_i's and λ_i's, we employ *mean absolute error* (MAE) which is defined as $\text{MAE}_\lambda = {}^1/_{\prod_k I_k} \sum_i \left|\widehat{\lambda}_i/\widehat{\lambda}^{(\text{avg})} - \lambda_i^\natural/\lambda^{\natural(\text{avg})}\right|$, where $\widehat{\lambda}_i$ is the estimated value corresponding to $\lambda_i^\natural$, and the subscript (avg) denotes the average across all values, e.g., $\lambda^{(\text{avg})} = {}^1/_{\prod_k I_k} \sum_i \lambda_i$. Note that the metric MAE_λ involves entry-wise normalization of the estimates and the ground truth by the average in order to remove the scaling ambiguity. The metric MAE_p is defined in the same way,

with all the λ-terms replaced by their p counterparts..

Implementation Settings. To learn the nonlinear function, we use a neural network $g_\theta(\cdot)$ with 3 hidden layers and 20 ReLU activation functions in each hidden layer. The regularization parameter μ is fixed to be 2000. More implementation details and experiments using different μ's are provided in the supplementary material in Sec. N.

Results. Table. 1 shows the performance under various γ_Ω's of all the methods under test. Note that for the baselines that do not take detection probability into their models, only the MAE_λ results are presented. The results show that all the non-Poisson baselines including the NN-based methods struggle to estimate the true counts λ_i. This shows the importance of tailored modeling for integer data. Among the algorithms using Poisson models, the proposed UncleTC mothod estimates λ_i's with much higher accuracy relative to NTF-CPD-KL, as our method explicitly models the undercounting effect. One can also note that UncleTC significantly outperforms UncleTC(Linear), in terms of both MAE_p and MAE_λ. This is due to the ability of UncleTC to better handle the nonlinear relationships between the detection probabilities p_i and the side information z_i, by using its NN-based learning. In addition, we note that as γ_Ω is higher, i.e., more observations are available, the performance becomes better, as expected and indicated in our theoretical claim.

In Table 2 and Table 3, we show the performance of the algorithms under vairous SNR and γ_Θ, respectively. In Table 2, as SNR increases from 0 to 40dB, i.e., ζ in Assumption 4.3 becomes smaller and the feature vectors z_i's become more similar, the MAEs decrease noticeably. The impact on MAE_p is obviously spelled out, as a reduction of around 7 times is observed. Table 3 shows that as the size of Θ gets larger, the estimation accuracy of the proposed method is positively impacted. Both tables echo the results presented by Theorem 4.5.

Table 4 validates a key result from Theorem 4.5. That is, there exists a global scaling factor and a counter-scaling

Table 4. Statistics of $\widehat{\lambda}_i/\lambda_i^{\natural}$ and $\widehat{p}_i/p_i^{\natural}$ for different trials of the proposed `UncleTC` with settings $\gamma_\Omega = 0.3, \gamma_\Xi = 0.3, \gamma_\Theta = 0.2, \zeta = 40$dB, and $g(\boldsymbol{z}) = \mathrm{sigmoid}(\boldsymbol{\nu}^\top (0.1 \log(\boldsymbol{z}^2) + 0.1\boldsymbol{z}^2))$

#Trial	mean±std of $\widehat{\lambda}_i/\lambda_i^{\natural}$	mean±std of $\widehat{p}_i/p_i^{\natural}$	mean±std of $\widehat{\lambda}_i\widehat{p}_i/\lambda_i^{\natural}p_i^{\natural}$
1	1.16 ± 0.04	0.86 ± 0.08	0.99 ± 0.08
2	1.29 ± 0.04	0.78 ± 0.08	1.00 ± 0.07
3	0.98 ± 0.03	1.02 ± 0.06	0.99 ± 0.07
4	1.33 ± 0.05	0.76 ± 0.10	1.00 ± 0.09
5	1.09 ± 0.03	0.93 ± 0.06	1.00 ± 0.06
6	1.16 ± 0.05	0.86 ± 0.11	0.99 ± 0.12
7	1.07 ± 0.03	0.94 ± 0.07	1.00 ± 0.06
8	1.24 ± 0.03	0.81 ± 0.05	0.99 ± 0.05
9	0.83 ± 0.02	1.21 ± 0.05	1.00 ± 0.04
10	0.80 ± 0.02	1.23 ± 0.06	0.99 ± 0.07

factor for λ_i and p_i, respectively. In other words, assuming perfect estimation for $\widehat{\lambda}_i$ and $\widehat{p}_i$ by the proposed `UncleTC`, we have $\widehat{\lambda}_i = \xi\lambda_i^{\natural}$, $\widehat{p}_i = 1/\xi p_i^{\natural}, \forall i$ for a certain scalar $\xi > 0$. To see if the `UncleTC` algorithm attains good estimation of λ_i and p_i up to a global scaling/counter-scaling ambiguity, we report the mean and standard deviation of $\widehat{\lambda}_i/\lambda_i^{\natural}$ and $\widehat{p}_i/p_i^{\natural}$ taken over all the i's in each random trial. One can see that only small standard deviations exist in all the trials, suggesting that the scaling/counter-scaling factors are approximately identical over all the indices. Similar observations are made on $\widehat{p}_i/p_i^{\natural}$. One can also note that the scaling factors of $\widehat{\lambda}_i$ and $\widehat{p}_i$ are approximately reciprocal to each other (cf. the last column of Table 4). This observations corroborate the findings in Theorem 4.5.

6.2. Real-Data Experiments

COVID-19 Data. The dataset is obtained from the COVID-19 impact analysis platform hosted by the University of Maryland, College Park, United States (Zhang et al., 2020b). The dataset includes the number of reported COVID-19 cases of about 2270 US counties during 2020-2021, with 25 different attributes such as social distancing index, population density, testing capacity and so on. In order to extract a count-type tensor from the COVID-19 dataset, we represent each county using two coordinates, i.e., the rounded value of latitude and longitude of its geographical center. This way, we obtain an integer tensor of size $43 \times 80 \times 60$ (with 18.42% of entries observed), recording the COVID-19 case numbers of 642 counties over 60 days from January 1, 2021 to March 1, 2021.

Pollinator-Plant Interaction (PPI) Data. The Plant-Pollinator Interaction (PPI) dataset is a publicly available dataset, collected by the researchers at the H.J. Andrews (HJA) Long-term Ecological Research site in Oregon, USA (Seo et al., 2022; Daly et al., 2019). We consider a subset of the dataset by selecting 50 plants and 50 pollinators that have the highest numbers of interactions. This leads to a $50 \times 50 \times 37$ integer tensor with 22.06% observed entries.

Table 5. The rRMSE of various methods on the COVID-19 and PPI datasets.

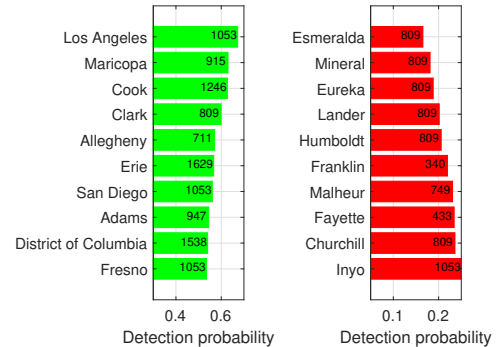| Method	COVID-19	PPI
`UncleTC`	**3.52 ± 0.38**	**9.89 ± 1.67**
`UncleTC(Linear)`	3.90 ± 0.33	10.21 ± 1.45
`HaLRTC`	6.14 ± 0.45	10.77 ± 1.25
`BPTF`	6.72 ± 0.42	11.02 ± 1.39
`NTF-CPD-KL`	**3.61 ± 0.52**	**10.17 ± 1.12**
`NTF-CPD-LS`	7.00 ± 0.38	11.25 ± 1.27
`NTF-Tucker-LS`	6.82 ± 0.48	11.22 ± 1.27
`CostCo`	7.12 ± 0.34	11.01 ± 1.14
`POND`	6.92 ± 0.51	10.95 ± 1.31

Figure 1. The top-10 counties that have the highest `UncleTC`-estimated average detection probabilities (left) and lowest average detection probabilities (right) for COVID-19. Inside the bar, we show the average numbers of COVID-19 tests done per 1000 people during the selected period.

The dataset also includes about 37 entry attributes corresponding to each observation, such as flower color of the plant, pollinator tongue length, weather conditions, and so on.

For both datasets, since the features are in different numerical units and range, we normalize each feature value using min-max normalization (Jain et al., 2005), across all observations. More details on the datasets and tensor construction are given in the supplementary material in Sec. N.

Quantitative Evaluation. Since the ground-truth $\lambda_i^{\natural}$'s and $p_i^{\natural}$'s are unknown for real data, we employ a count data prediction metric, i.e., *relative root mean square* (rRMSE) (Fu et al., 2021) for evaluation. Let $\{y_i\}_{i\in\Delta}$ be the set of actual holdout observations and $\{\widehat{y}_i\}_{i\in\Delta}$ be the corresponding predicted values. Then, the rRMSE is defined as $\mathrm{rRMSE} = 1/\overline{y}\sqrt{\sum_{i\in\Delta}|y_i - \widehat{y}_i|^2/|\Delta|}$, where $\overline{y}$ is the mean value of the actual observations $\{y_i\}_{i\in\Delta}$. Note that, for the proposed method, the observations are predicted via $\widehat{y}_i = \widehat{\lambda}_i\widehat{p}_i$.

Table 5 presents the performance of various approaches on the COVID-19 and PPI datasets. One can see that the pro-

posed `UncleTC` exhibits promising performance on both datasets. One can also see that the performance of the Poisson model-based method `NTF-CPD-KL` has a comparable `rRMSE` performance relative to our method. This is expected, as the Poisson-Binomial model with parameters λ and p can be considered as a Poisson model with the parameter λp (see Lemma B.1). Nontheless, our method still outperforms `NTF-CPD-KL`, as the incorporation of the entry attributes in our framework may be beneficial. Compared to the non-Poisson methods, our approach stands out with obviously lower `rRMSEs`, which is similar to what was observed in the simulations. It can also be noted that `UncleTC` outperforms its linear function-based counterpart `UncleTC(Linear)` in both cases, which is also consistent with our simulation results.

Qualitative Evaluation. In Fig. 1, we analyze the detection probabilities output by the proposed `UncleTC` for COVID-19 dataset. Here, we present the top 10 counties having the highest and lowest detection probabilities for COVID-19 cases found by the proposed `UncleTC` (averaged over all days). Note that the estimated detection probabilities have a global scaling ambiguity—see Theorem 4.5—but their relative ranking order is not affected by this ambiguity. We also report the average number of COVID-19 tests done per 1000 people during the selected period (we removed this feature while training the model to avoid correlated results). One can see that the counties with more tests done are mostly aligned with higher detection probabilities estimated by our algorithm. Also, most counties reported with high detection probabilities are populous, urban counties, which enjoy better infrastructure. These observations suggest that the proposed algorithm outputs plausible results for the detection probabilities—also see the detection probability results for the PPI dataset in the supplementary material in Sec. N.

7. Conclusion

We proposed an under-counted tensor completion framework, which is motivated by the commonly seen under-counting effects in data acquisition across many domains, e.g., ecology and epidemiology. Our model uses a Poisson-Binomial data-generating perspective as often advocated in computational ecology, where the true integer data are under-counted by a Binomial detection process. Unlike existing approaches that treat the under-counting effects using relatively simple models, we used a neural network-based design to account for a wide range of unknown nonlinear relations between the side information (attributes) and the under-counting effects. More importantly, we showed that the proposed under-counted tensor completion criterion can *provably* recover the underlying Poisson tensor and its entries' under-counting characteristics, under reasonable

conditions—which is the first theoretical result of its kind. We validated our theorem and algorithm using simulated and real data and observed intuitively plausible results from our real data results.

Limitations. A key limitation lies in the model assumption that the under-counted probabilities are functions of observed side attributes. However, such attributes may not always be available. Provable UC-TC criteria that do not rely on such assumption are highly desirable, yet the current formulation and analysis cannot cover such settings without substantial changes and re-design.

References

Acar, E., Dunlavy, D. M., and Kolda, T. G. Link prediction on evolving data using matrix and tensor factorizations. In *IEEE International Conference on Data Mining Workshop*, pp. 262–269, 2009.

Balazevic, I., Allen, C., and Hospedales, T. TuckER: Tensor factorization for knowledge graph completion. In *Proceedings of Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, pp. 5185–5194, 2019.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Bertsekas, D. P. *Nonlinear programming*. Athena Scientific, 1999.

Billio, M., Casarin, R., and Iacopini, M. Bayesian markov-switching tensor regression for time-varying networks. *Journal of the American Statistical Association*, pp. 1–13, 2022.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004.

Cao, Y. and Xie, Y. Poisson matrix recovery and completion. *IEEE Transactions on Signal Processing*, 64(6):1609–1620, 2015.

Chacoff, N. P., Vázquez, D. P., Lomáscolo, S. B., Stevani, E. L., Dorado, J., and Padrón, B. Evaluating sampling

completeness in a desert plant–pollinator network. *Journal of Animal Ecology*, 81(1):190–200, 2012.

Chi, E. C. and Kolda, T. G. On tensors, sparsity, and nonnegative factorizations. *SIAM Journal on Matrix Analysis and Applications*, 33(4):1272–1299, 2012.

Daly, C., Schulze, M. D., and McKee, W. A. Meteorological data from benchmark stations at the HJ Andrews experimental forest, 1957 to present, 2019. URL `http://andlter.forestry.oregonstate.edu/data/abstract.aspx?dbcode=MS001.https://doi.org/10.6073/pasta/c021a2ebf1f91adf0ba3b5e53189c84f`.

Dennis, E. B., Morgan, B. J., and Ridout, M. S. Computational aspects of N-mixture models. *Biometrics*, 71(1): 237–246, 2015.

Ding, M., Fu, X., Huang, T.-Z., Wang, J., and Zhao, X.-L. Hyperspectral super-resolution via interpretable block-term tensor modeling. *IEEE Journal of Selected Topics in Signal Processing*, 15(3):641–656, 2020.

Dorado, J., Vázquez, D. P., Stevani, E. L., and Chacoff, N. P. Rareness and specialization in plant–pollinator networks. *Ecology*, 92(1):19–25, 2011.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

Durif, G., Modolo, L., Mold, J. E., Lambert-Lacroix, S., and Picard, F. Probabilistic count matrix factorization for single cell expression data analysis. *Bioinformatics*, 35 (20):4011–4019, 2019.

Fan, J., Ding, L., Yang, C., and Udell, M. Low-rank tensor recovery with euclidean-norm-induced schatten-p quasi-norm regularization. *arXiv preprint arXiv:2012.03436*, 2020.

Fu, X., Ibrahim, S., Wai, H.-T., Gao, C., and Huang, K. Block-randomized stochastic proximal gradient for low-rank tensor factorization. *IEEE Transactions on Signal Processing*, 68:2170–2185, 2020a.

Fu, X., Vervliet, N., De Lathauwer, L., Huang, K., and Gillis, N. Computing large-scale matrix and tensor decomposition with structured factors: A unified nonconvex optimization perspective. *IEEE Signal Processing Magazine*, 37(5):78–94, 2020b.

Fu, X., Seo, E., Clarke, J., and Hutchinson, R. A. Link prediction under imperfect detection: Collaborative filtering for ecological networks. *IEEE Transactions on Knowledge and Data Engineering*, 33(8):3117–3128, 2021.

Gandy, S., Recht, B., and Yamada, I. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27:025010, 2011.

Ghadermarzy, N., Plan, Y., and Yilmaz, O. Learning tensors from partial binary measurements. *IEEE Transactions on Signal Processing*, 67(1):29–40, 2018.

Harshman, R. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics*, 16, 1970.

Hu, Y., Koren, Y., and Volinsky, C. Collaborative Filtering for Implicit Feedback Datasets. *IEEE International Conference on Data Mining*, 2008.

Hutchinson, R., Liu, L.-P., and Dietterich, T. Incorporating boosted regression trees into ecological latent variable models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pp. 1343–1348, 2011.

Jain, A., Nandakumar, K., and Ross, A. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38(12):2270–2285, 2005.

Joseph, L. N., Elkin, C., Martin, T. G., and Possingham, H. P. Modeling abundance using n-mixture models: the importance of considering ecological mechanisms. *Ecological Applications*, 19(3):631–642, 2009.

Kanatsoulis, C. I., Fu, X., Sidiropoulos, N. D., and Akcakaya, M. Tensor completion from regular sub-nyquist samples. *IEEE Transactions on Signal Processing*, 68: 1–16, 2020.

Karatzoglou, A., Amatriain, X., Baltrunas, L., and Oliver, N. Multiverse recommendation: N-dimensional tensor factorization for context-aware collaborative filtering. In *Proceedings of the ACM Conference on Recommender Systems*, pp. 79–86, 2010.

Kashima, H., Kato, T., Yamanishi, Y., Sugiyama, M., and Tsuda, K. Link propagation: A fast semi-supervised learning algorithm for link prediction. In *Proceedings of the SIAM international conference on data mining*, pp. 1100–1111, 2009.

Kim, Y.-D. and Choi, S. Nonnegative tucker decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

Lee, C. and Wang, M. Tensor denoising and completion based on ordinal observations. In *International Conference on Machine Learning*, pp. 5778–5788. PMLR, 2020.

Liang, D., Charlin, L., McInerney, J., and Blei, D. M. Modeling user exposure in recommendation. In *Proceedings of the International Conference on World Wide Web*, pp. 951–961, 2016.

Lin, S. and Zhang, J. Generalization bounds for convolutional neural networks. *arXiv preprint arXiv:1910.01487*, 2019.

Liu, H., Li, Y., Tsang, M., and Liu, Y. Costco: A neural tensor completion model for sparse tensors. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 324–334, 2019.

Liu, J., Musialski, P., Wonka, P., and Ye, J. Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):208–220, 2013.

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. L., and Hines, J. E. *Occupancy estimation and modeling: Inferring patterns and dynamics of species occurrence*. Elsevier, San Diego, USA, 2006.

Montanari, A. and Sun, N. Spectral algorithms for tensor completion. *Communications on Pure and Applied Mathematics*, 71, 2016.

Mu, C., Huang, B., Wright, J., and Goldfarb, D. Square deal: Lower bounds and improved relaxations for tensor recovery. In *Proceedings of the International Conference on Machine Learning*, volume 32, pp. 73–81, 2014.

Papalexakis, E. E., Faloutsos, C., and Sidiropoulos, N. D. Tensors for data mining and data fusion: Models, applications, and scalable algorithms. *ACM Transactions on Intelligent Systems and Technology*, 8(2):16, 2017.

Pu, W., Ibrahim, S., Fu, X., and Hong, M. Stochastic mirror descent for low-rank tensor decomposition under non-euclidean losses. *IEEE Transactions on Signal Processing*, 70:1803–1818, 2022.

Razaviyayn, M., Hong, M., and Luo, Z.-Q. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013.

Royle, J. A. N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, 60(1):108–115, 2004.

Schein, A., Paisley, J., Blei, D. M., and Wallach, H. Bayesian poisson tensor factorization for inferring multilateral relations from sparse dyadic event counts. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1045–1054, 2015.

Schneider, E. C. Failing the test — the tragic data gap undermining the U.S. pandemic response. *New England Journal of Medicine*, 383(4):299–302, 2020.

Seo, E. and Hutchinson, R. Predicting links in plant-pollinator interaction networks using latent factor models with implicit feedback. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

Seo, E., Jones, J. A., Hutchinson, R. A., and Pfeiffer, V. W. Plant pollinator data at HJ Andrews experimental forest, 2011 to 2021, 2022. URL http://andlter.forestry.oregonstate.edu/data/abstract.aspx?dbcode=SA026.https://doi.org/10.6073/pasta/eec55cbc3dbfc56428629773737ab3e5.

Serfling, R. J. Probability Inequalities for the Sum in Sampling without Replacement. *The Annals of Statistics*, 2(1):39 – 48, 1974.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shashua, A. and Hazan, T. Non-negative tensor factorization with applications to statistics and computer vision. In *Proceedings of the International Conference on Machine Learning*, pp. 792–799, 2005.

Simchowitz, M. Zero-inflated poisson factorization for recommendation systems. *Junior Independent Work (advised by D. Blei), Princeton University, Department of Mathematics*, 2013.

Sørensen, M. and De Lathauwer, L. Fiber sampling approach to canonical polyadic decomposition and application to tensor completion. *SIAM Journal on Matrix Analysis and Applications*, 40(3):888–917, 2019.

Tan, H., Wu, Y., Shen, B., Jin, P. J., and Ran, B. Short-term traffic prediction based on dynamic tensor completion. *IEEE Transactions on Intelligent Transportation Systems*, 17(8):2123–2133, 2016.

Tillinghast, C., Fang, S., Zhang, K., and Zhe, S. Probabilistic neural-kernel tensor decomposition. In *IEEE International Conference on Data Mining*, pp. 531–540, 2020.

Tylianakis, J. M., Laliberté, E., Nielsen, A., and Bascompte, J. Conservation of species interaction networks. *Biological Conservation*, 143(10):2270–2279, 2010.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications*, pp. 210–268. Cambridge University Press, 2012.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Wang, Y., Peng, J., Zhao, Q., Leung, Y., Zhao, X.-L., and Meng, D. Hyperspectral image restoration via total variation regularized low-rank tensor decomposition. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(4):1227–1243, 2018.

Yuan, M. and Zhang, C.-H. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, 16(4):1031–1068, 2016.

Zhang, G., Fu, X., Wang, J., Zhao, X.-L., and Hong, M. Spectrum cartography via coupled block-term tensor decomposition. *IEEE Transactions on Signal Processing*, 68:3660–3675, 2020a.

Zhang, L., Ghader, S., Pack, M. L., Xiong, C., Darzi, A., Yang, M., Sun, Q., Kabiri, A., and Hu, S. An interactive COVID-19 mobility impact and social distancing analysis platform. *medRxiv*, 2020b.

Zhang, Z. and Aeron, S. Exact tensor completion using t-SVD. *IEEE Transactions on Signal Processing*, 65(6):1511–1526, 2017.

Supplementary Material of " Under-Counted Tensor Completion with Neural Incorporation of Attributes"

A. Notation

The notations x, $\boldsymbol{x}$, $\boldsymbol{X}$ and $\underline{\boldsymbol{X}}$ represent a scalar, a vector, a matrix, and a tensor, respectively. x_i denotes the ith element of the vector $\boldsymbol{x}$. $[\boldsymbol{X}]_{i,j}$ and $\boldsymbol{X}(i,j)$ both mean the (i,j)th entry of $\boldsymbol{X}$. Similarly, $[\underline{\boldsymbol{X}}]_{\boldsymbol{i}}$ belongs to the $\boldsymbol{i}$th entry of the tensor $\underline{\boldsymbol{X}} \in \mathbb{R}^{I_1 \times \cdots \times I_K}$ where $\boldsymbol{i} = (i_1, \ldots, i_K)$. $\mathbb{Z}_+$ denote the set of all nonnegative integers. $\boldsymbol{x} \geq \boldsymbol{0}$, $\boldsymbol{X} \geq \boldsymbol{0}$, and $\underline{\boldsymbol{X}} \geq \boldsymbol{0}$ imply that all the associated entries are greater than 0. $\|\boldsymbol{x}\|_2$, $\|\boldsymbol{X}\|_F$ and $\|\underline{\boldsymbol{X}}\|_F$ all mean the Euclidean (Frobenius) norm of the augment. $\|[\underline{\boldsymbol{X}}]_{\boldsymbol{\Omega}}\|_F$ indicates the Euclidean norm of the vector formed by concatenating the entries indexed by all unique $\boldsymbol{i}$'s belonging to $\boldsymbol{\Omega}$, i.e., $\|[\underline{\boldsymbol{X}}]_{\boldsymbol{\Omega}}\|_F^2 = \sum_{\boldsymbol{i} \in \mathrm{Supp}(\boldsymbol{\Omega})} x_{\boldsymbol{i}}^2$, where $\mathrm{Supp}(\boldsymbol{\Omega})$ denotes the set of all unique entries in $\boldsymbol{\Omega}$. $\|\underline{\boldsymbol{X}}\|_\infty$ denotes the max norm of $\underline{\boldsymbol{X}}$ such that $\|\underline{\boldsymbol{X}}\|_\infty = \max_{\boldsymbol{i}} [\underline{\boldsymbol{X}}]_{\boldsymbol{i}}$. $[I]$ means an integer set $\{1, 2, \ldots, I\}$. $^\top$ and † denote transpose and pseudo-inverse, respectively. $\boldsymbol{x} \circ \boldsymbol{y}$ denotes the outer product of two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, i.e., $\boldsymbol{x} \circ \boldsymbol{y} = \boldsymbol{x}\boldsymbol{y}^\top$. $\circledast$ denotes the element-wise product (also known as Hadamard product). The function $\mathrm{sigmoid}(x)$ represents $\frac{1}{1+\exp(-x)}$. The constant e denotes the Euler's number. $[x]_{[0,1]}$ denotes $\min\{\max\{x, 0\}, 1\}$.

B. MLE Formulation

Assuming that $y_{\boldsymbol{i}}$'s are independently sampled from the generative model in (3), the log-likelihood of the observations can be expressed as follows:

$$\log \prod_{\boldsymbol{i} \in \boldsymbol{\Omega}} \mathrm{Pr}(Y_{\boldsymbol{i}} = y_{\boldsymbol{i}}; \lambda_{\boldsymbol{i}}, p_{\boldsymbol{i}}) = \log \left(\prod_{\boldsymbol{i} \in \boldsymbol{\Omega}} \sum_{n=y_{\boldsymbol{i}}}^{\infty} \mathrm{Pr}(N_{\boldsymbol{i}} = n; \lambda_{\boldsymbol{i}}) \mathrm{Pr}(y_{\boldsymbol{i}} | N_{\boldsymbol{i}} = n; p_{\boldsymbol{i}}) \right)$$
$$= \sum_{\boldsymbol{i} \in \boldsymbol{\Omega}} \log \left(\sum_{n=y_{\boldsymbol{i}}}^{\infty} \left(\frac{\lambda_{\boldsymbol{i}}^n e^{-\lambda_{\boldsymbol{i}}}}{n!} \frac{n! p_{\boldsymbol{i}}^{y_{\boldsymbol{i}}} (1-p_{\boldsymbol{i}})^{n-y_{\boldsymbol{i}}}}{y_{\boldsymbol{i}}!(n-y_{\boldsymbol{i}})!} \right) \right), \tag{14}$$

where $Y_{\boldsymbol{i}}$ denotes the random variable associated with the observation $y_{\boldsymbol{i}}$ and $N_{\boldsymbol{i}}$ denotes the random variable associated with the true count $n_{\boldsymbol{i}}$. To circumvent the infinite sum in the log-likelihood, we invoke the folowing lemma:

Lemma B.1. *(Dennis et al., 2015) Assume that the observation y follows the below generative model:*

$$n \sim \mathrm{Poisson}(\lambda), \tag{15a}$$
$$y \sim \mathrm{Binomial}(n, p). \tag{15b}$$

Then, the model in (15) *is equivalent to* $y \sim \mathrm{Poisson}(\lambda p)$.

Lemma B.1 states that the Poisson-Binomial model admits an equivalent Poisson model whose Poisson parameter is the multiplication of the original Poisson parameter and the detection probability. The derivation is based on an interesting equality:

$$\sum_{n=y}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \frac{n! p^y (1-p)^{n-y}}{y!(n-y)!} = \frac{(p\lambda)^y e^{-\lambda p}}{y!};$$

see the derivation in (Fu et al., 2021).

Applying the above equality in (14), we obtain:

$$\log \prod_{\boldsymbol{i} \in \boldsymbol{\Omega}} \mathrm{Pr}(Y_{\boldsymbol{i}} = y_{\boldsymbol{i}}; \lambda_{\boldsymbol{i}}, p_{\boldsymbol{i}}) = \sum_{\boldsymbol{i} \in \boldsymbol{\Omega}} \log \left(\frac{(p_{\boldsymbol{i}} \lambda_{\boldsymbol{i}})^{y_{\boldsymbol{i}}} e^{-\lambda_{\boldsymbol{i}} p_{\boldsymbol{i}}}}{y_{\boldsymbol{i}}!} \right)$$
$$= \sum_{\boldsymbol{i} \in \boldsymbol{\Omega}} \left[y_{\boldsymbol{i}} \log \lambda_{\boldsymbol{i}} + y_{\boldsymbol{i}} \log p_{\boldsymbol{i}} - \lambda_{\boldsymbol{i}} p_{\boldsymbol{i}} - \log y_{\boldsymbol{i}}! \right]. \tag{16}$$

Here, the MLE is formulated and derived under the same spirit of the MC case in (Fu et al., 2021).

C. More Details on Algorithm Design

In this section, we provide more details on the implementation of the proposed algorithm `UncleTC`. Our formulation is given below:

$$\operatorname*{minimize}_{\{U_k\}_{k=1}^K, \theta, \{p_i\}} \frac{1}{|\Omega|} \sum_{i \in \Omega} \left[\left(\sum_{f=1}^F \prod_{k=1}^K U_k(i_k, f) \right) p_i - y_i \log \left(\sum_{f=1}^F \prod_{k=1}^K U_k(i_k, f) \right) - y_i \log p_i \right] + \frac{\mu}{|\Xi|} \sum_{i \in \Xi} \ell(g_\theta(z_i), p_i),$$

(17a)

$$\text{subject to } U_k \geq 0, \forall k \in [K],$$

(17b)

$$0 \leq p_i \leq 1, \ \forall i \in \Omega \cup \Xi,$$

(17c)

where $\mu > 0$ is a regularization parameter and $\ell(\cdot, \cdot)$ denotes a certain distance/divergence measure, e.g., the least squares function $\ell(x, y) = (x - y)^2$.

C.1. The U_k-Subproblem

The subproblem for each U_k is given below:

$$\operatorname*{minimize}_{U_k \geq 0} w(U_k) := \sum_{i \in \Omega} \left[\left(\sum_{f=1}^F U_k(i_k, f) \prod_{j \neq k} U_j(i_j, f) \right) p_i - y_i \log \left(\sum_{f=1}^F U_k(i_k, f) \prod_{j \neq k} U_j(i_j, f) \right) \right].$$

(18)

To obtain the update rule for U_k from (18), we have the following result:

Lemma C.1. *Let $\overline{U}_k$ denote the current estimate of U_k. Then, the objective function in* (18) *can be upper-bounded by the following function:*

$$s(U_k; \overline{U}_k) = \sum_{i \in \Omega} \left[\left(\sum_{f=1}^F U_k(i_k, f) \prod_{j \neq k} U_j(i_j, f) \right) p_i - y_i \sum_{f=1}^F \alpha_i^{(f)} \log \left(\frac{U_k(i_k, f) \prod_{j \neq k} U_j(i_j, f)}{\alpha_i^{(f)}} \right) \right],$$

(19)

where $\alpha_i^{(f)} = \dfrac{\overline{U}_k(i_k, f) \prod_{j \neq k} U_j(i_j, f)}{\sum_{f=1}^F \overline{U}_k(i_k, f) \prod_{j \neq k} U_j(i_j, f)}.$

The construction of the surrogate function $s(U_k; \overline{U}_k)$ follows the idea in (Chi & Kolda, 2012; Fu et al., 2021); The differences include that the model in (Chi & Kolda, 2012) did not have p_i terms and that the work in (Fu et al., 2021) did not consider the tensor case. Here, we use the Jensen's inequality to construct a "tight" upper bound, i.e., $w(U_k) \leq s(U_k; \overline{U}_k)$ and the equality can be attained at $U_k = \overline{U}_k$; see Sec. L for the proof of Lemma C.1. To be more precise, the update of U_k is through solving the following subproblem:

$$U_k \leftarrow \arg \min_{U_k \geq 0} s(U_k; \overline{U}_k).$$

(20)

By taking the derivative of $s(\cdot; \overline{U}_k)$ w.r.t. U_k and equating the derivative to zero, we obtain:

$$U_k(i_k, f) \leftarrow \frac{\sum_{i \in \Omega} y_i \alpha_i^{(f)}}{\sum_{i \in \Omega} \prod_{j \neq k} U_j(i_j, f) p_i}.$$

(21)

For efficient implementation of the above updates in the presence of missing data, we adopt a strategy proposed in (Chi & Kolda, 2012) for handling sparse tensors.

C.2. The p-Subproblem

Table 6 summarizes the update rules for the p_i-subproblem in Sec. 5 under different choices of the regularization function $\ell(\cdot)$. The complete description of the proposed `UncleTC` algorithm is provided in Algorithm 1.

Table 6. Updates for p_i subproblems for different loss functions

	$\ell(x, y) = (x - y)^2$	$\ell(x, y) = x \log \frac{x}{y} - x + y$
Case 1: $\forall i \in \Omega \cap \Xi$	$p_i \leftarrow \frac{(2\bar{\mu}g_\theta(z_i) - \bar{\lambda}_i) + \sqrt{(2\bar{\mu}g_\theta(z_i) - \bar{\lambda}_i)^2 + 8\bar{\mu}\bar{y}_i}}{4\bar{\mu}}$	$p_i \leftarrow \frac{\bar{y}_i + \bar{\mu}g_\theta(z_i)}{\bar{\lambda}_i + \bar{\mu}}$
Case 2: $\forall i \in \Xi - \Omega \cap \Xi$	$p_i \leftarrow g_\theta(z_i)$	$p_i \leftarrow g_\theta(z_i)$
Case 3: $\forall i \in \Omega - \Omega \cap \Xi$	$p_i \leftarrow [y_i/\lambda_i]_{[0,1]}$	$p_i \leftarrow [y_i/\lambda_i]_{[0,1]}$

$$\bar{\mu} = \mu/|\Xi|, \; \bar{\lambda}_i = \lambda_i/|\Omega|, \; \bar{y}_i = y_i/|\Omega|$$

Algorithm 1 `UncleTC`

---

input : Observations $y_i, \forall i \in \Omega$, feature vectors $z_i, \forall i \in \Xi$, Initializations $\theta^{(0)}, p_i^{(0)}, \; \forall i \in \Xi$, and $U_k^{(0)}, \forall k$.

1 $t \leftarrow 0$;
2 **repeat**
3    $\lambda_i^{(t+1)} \leftarrow \sum_{f=1}^F \prod_{k=1}^K U_k^{(t)}(i_k, f), \; \forall i$;
4    $\tilde{p}_i^{(t+1)} \leftarrow g_{\theta^{(t)}}(z_i), \; i \in \Xi$;
5    $p_i^{(t+1)} \leftarrow \arg\min_{p_i \in [0,1]} \frac{1}{|\Omega|}(\lambda_i^{(t+1)}p_i^{(t)} - y_i \log(p_i^{(t)})) + \frac{\mu}{|\Xi|}\ell(\tilde{p}_i^{(t+1)}, p_i^{(t)}), \; \forall i \in \Omega \cap \Xi$;
6    $p_i^{(t+1)} \leftarrow \arg\min_{p_i \in [0,1]} \ell(\tilde{p}_i^{(t+1)}, p_i^{(t)}), \; \forall i \in \Xi - \Omega \cap \Xi$;
7    $p_i^{(t+1)} \leftarrow [y_i/\lambda_i^{(t+1)}]_{[0,1]}, \; \forall i \in \Omega - \Omega \cap \Xi$;
8    $\theta^{(t+1)} \leftarrow \arg\min_\theta \sum_{i \in \Xi} \ell(g_\theta(z_i), p_i^{(t+1)})$;
9   **for** $k = 1$ **to** K **do**
10      $r \leftarrow 0$;
11   **repeat**
12        $\alpha_i^{(f)} \leftarrow \frac{U_k^{(r)}(i_k, f) \prod_{j \neq k} U_j^{(t)}(i_j, f)}{\sum_{f=1}^F U_k^{(r)}(i_k, f) \prod_{j \neq k} U_j^{(t)}(i_j, f)}, \; \forall i, f$;
13        $U_k^{(r+1)}(i_k, f) \leftarrow \frac{\sum_{i \in \Omega} y_i \alpha_i^{(f)}}{\sum_{i \in \Omega} \prod_{j \neq k} U_j^{(t)}(i_j, f) p_i^{(t+1)}}, \; \forall i_k \in [I_k], f$;
14        $r \leftarrow r + 1$;
15   **until** *the stopping criterion is reached*;
16      $U_k^{(t+1)} \leftarrow U_k^{(r)}$;
17      $U_k^{(0)} \leftarrow U_k^{(r)}$;
18   **end**
19    $t \leftarrow t + 1$;
20 **until** *the stopping criterion is reached*;
   output : estimates $\{\widehat{U}_k\}_{k=1}^K$, $\widehat{\theta}$, and $\widehat{p}_i, \; \forall i \in \Omega \cup \Xi$

D. Proof of the Main Result (Theorem 4.5)

Let us first construct the following constraint sets:

$$\mathcal{L} = \left\{ \underline{\Lambda} \mid \underline{\Lambda} = \sum_{f=1}^F U_1(:, f) \circ \ldots \circ U_K(:, f), \|U_k(:, f)\|_2 \leq u, \; U_k \geq 0, \forall k, f \right\}, \tag{22a}$$

$$\mathcal{M} = \left\{ \underline{M} = \underline{\Lambda} \circledast \underline{P} \mid \beta \leq [\underline{M}]_i \leq \alpha, \forall i, \underline{\Lambda} \in \mathcal{L}, [\underline{P}]_i = g_\theta(z_i), \; g_\theta \in \mathcal{G}, z_i \in \mathbb{R}^D, \; \forall i \in \Xi \right\}. \tag{22b}$$

Under Assumption 4.1, we can choose the parameters u, α, and β as follows:

$$u = \sqrt{F}\alpha_u, \; \alpha = F\alpha_u^K p_{\max}, \; \beta = F\beta_u^K p_{\min}. \tag{23}$$

Using these notations, the MLE in (6) can be re-expressed as follows:

$$\widehat{\underline{M}} = \arg\min_{\underline{M} \in \mathcal{M}} \sum_{i \in \Omega} f(m_i; y_i, z_i), \tag{24}$$

where $f(m_i; y_i, z_i) = m_i - y_i \log m_i$, where $m_i = \lambda_i g_\theta(z_i)$.

15

Estimating $\|\underline{M}^\natural - \widehat{\underline{M}}\|_{\mathrm{F}}/\prod_k I_k$**:** Our goal is to characterize the estimation accuracies of $\lambda_{\boldsymbol{i}}^\natural$ and $p_{\boldsymbol{i}}^\natural$. To achieve this goal, we first characterize the term $\|\underline{M}^\natural - \widehat{\underline{M}}\|_{\mathrm{F}}$. We have the following result:

Theorem D.1. *Suppose that the assumptions of Theorem 4.5 hold true. Let $T = |\boldsymbol{\Omega}|$. Assume that $\{y_{\boldsymbol{i}}\}_{\boldsymbol{i} \in \boldsymbol{\Omega}}$ are i.i.d. samples under the generative model* (3) *and that $\boldsymbol{z}_{\boldsymbol{i}_1}, \ldots, \boldsymbol{z}_{\boldsymbol{i}_S}$ are the set of observed features. Then, under* (24), *with probability at least $1 - 2\delta - 3e^{-\alpha(e^2-3)}$, we have the following relation, for any $0 < \delta < 1$:*

$$\frac{\|\underline{M}^\natural - \widehat{\underline{M}}\|_{\mathrm{F}}^2}{\prod_k I_k} \leq C(\alpha, \beta) \left(2\mathfrak{R}_T(\mathcal{F}) + (5f_{\max} - f_{\min})\sqrt{\frac{2\log(4/\delta)}{T}} + F\alpha_u^K \left(1 + \frac{p_{\max}(e^2 - 2)}{p_{\min}}\right) \nu \right), \quad (25)$$

where $C(\alpha, \beta) = \frac{4\alpha\gamma}{1 - e^{-\gamma}}$, $\gamma = \frac{1}{8\beta}(\alpha - \beta)^2$, $f_{\max} \leq \alpha(1 + (e^2 - 2)\max\{|\log\beta|, \log\alpha\})$, $f_{\min} = \beta - \alpha(e^2 - 2)\log\alpha$,

$$\mathfrak{R}_T \leq \left(\frac{4}{\sqrt{T}} + \frac{12}{\sqrt{T}} \sqrt{F\sum_k I_k \log\left(6KL_f\sqrt{T}(F\alpha_u)^K\right) + \left(F\alpha_u^K L_f \sqrt{T}\|\boldsymbol{Z}\|_{\mathrm{F}}\mathscr{R}_{\mathcal{G}}\right)} \right),$$

$L_f = 1 + \frac{\alpha(e^2-2)}{\beta}$, $\boldsymbol{Z} = [\boldsymbol{z}_{\boldsymbol{i}_1}, \ldots, \boldsymbol{z}_{\boldsymbol{i}_S}] \in \mathbb{R}^{D \times S}$, and $\mathscr{R}_{\mathcal{G}}$ denotes the complexity measure of the class $\mathcal{G}$.

The proof is relegated to Appendix E.

Estimating $\|[\underline{M}^\natural - \widehat{\underline{M}}]_{\Theta}\|_{\mathrm{F}}/|\Theta|$**:** Next, we proceed to estimate $\|[\underline{M}^\natural - \widehat{\underline{M}}]_{\Theta}\|_{\mathrm{F}}/|\Theta|$ using the result in Theorem D.1.

Consider the following result:

Lemma D.2. *Assume that $\boldsymbol{i} \in \boldsymbol{\Theta}$ is drawn from $[I_1] \times \ldots \times [I_K]$ uniformly at random (without replacement). Consider that $\underline{M}_1, \underline{M}_2 \in \{\underline{M} \mid \underline{M} \in \mathbb{R}_+^{I_1 \times \cdots \times I_K}, \|\underline{M}\|_\infty \leq \alpha\}$ holds. Let us define*

$$\tau(\boldsymbol{\Theta}; \underline{M}_1, \underline{M}_2) = \left| \frac{1}{\sqrt{|\boldsymbol{\Theta}|}} \|(\underline{M}_1 - \underline{M}_2)_{\boldsymbol{\Theta}}\|_{\mathrm{F}} - \frac{1}{\sqrt{I^K}} \|\underline{M}_1 - \underline{M}_2\|_{\mathrm{F}} \right|.$$

Then, with probability greater than $1 - \delta$, we have

$$\tau(\boldsymbol{\Theta}; \underline{M}_1, \underline{M}_2) \leq \alpha \left(\log\left(\frac{2}{\delta}\right) \left(1 - \frac{(|\boldsymbol{\Theta}| - 1)}{\prod_k I_k}\right) \frac{1}{2|\boldsymbol{\Theta}|} \right)^{1/4}.$$

The proof is provided in Appendix J.

Combining Lemma D.2 (here, $\alpha = F\alpha_u^K p_{\max}$ using Assumption 4.1) with Theorem D.1 and denoting $\eta = \frac{1}{\sqrt{\prod_k I_k}} \left\|\underline{M}^\natural - \widehat{\underline{M}}\right\|_{\mathrm{F}}$, we have the following with probability at least $1 - 3\delta - 3e^{-\alpha(e^2-3)}$:

$$\frac{\left\|\left[\underline{M}^\natural - \widehat{\underline{M}}\right]_{\boldsymbol{\Theta}}\right\|_{\mathrm{F}}}{\sqrt{|\boldsymbol{\Theta}|}} \leq \eta + \tau(\boldsymbol{\Theta}; \underline{M}^\natural, \widehat{\underline{M}}), \quad (26)$$

where η^2 is given by (25) from Theorem D.1.

Recovering $\lambda_{\boldsymbol{i}}$**'s:** Next, we proceed to characterize the estimation accuracies of $\lambda_{\boldsymbol{i}}$'s using the result in (26). Under Assumptions 4.3 and 4.4, we have the following representation:

$$p_{\boldsymbol{i}}^\natural = g^\natural(\boldsymbol{z}_{\boldsymbol{i}}) = \xi + s_{\boldsymbol{i}}, \; |s_{\boldsymbol{i}}| \leq \zeta L_g, \; \forall \boldsymbol{i} \in \boldsymbol{\Theta},$$
$$\widehat{p}_{\boldsymbol{i}} = \widehat{g}_{\boldsymbol{\theta}}(\boldsymbol{z}_{\boldsymbol{i}}) = \xi' + s_{\boldsymbol{i}}', \; |s_{\boldsymbol{i}}'| \leq \zeta L_{\boldsymbol{\theta}}, \; \forall \boldsymbol{i} \in \boldsymbol{\Theta},$$

where $0 \leq \xi, \xi' \leq 1$ are certain scalars.

Hence, we get the following relation, with probability at least $1 - 3\delta - 3e^{-\alpha(e^2-3)}$:

$$
\begin{aligned}
(\eta + \tau(\boldsymbol{\Theta}; \underline{\boldsymbol{M}}^\natural, \widehat{\underline{\boldsymbol{M}}}))\sqrt{|\boldsymbol{\Theta}|} &\geq \left\| \left[\underline{\boldsymbol{M}}^\natural - \widehat{\underline{\boldsymbol{M}}} \right]_{\boldsymbol{\Theta}} \right\|_{\mathrm{F}} = \sqrt{\sum_{i \in \boldsymbol{\Theta}} (m_i^\natural - \widehat{m}_i)^2} \\
&= \sqrt{\sum_{i \in \boldsymbol{\Theta}} (\lambda_i^\natural p_i^\natural - \widehat{\lambda}_i \widehat{p}_i)^2} = \sqrt{\sum_{i \in \boldsymbol{\Theta}} (\lambda_i^\natural(\xi + s_i) - \widehat{\lambda}_i(\xi' + s_i'))^2} \\
&= \sqrt{\sum_{i \in \boldsymbol{\Theta}} (\xi \lambda_i^\natural - \xi' \widehat{\lambda}_i + s_i \lambda_i^\natural - s_i' \widehat{\lambda}_i)^2} \\
&\geq \left| \sqrt{\sum_{i \in \boldsymbol{\Theta}} (\xi \lambda_i^\natural - \xi' \widehat{\lambda}_i)^2} - \sqrt{\sum_{i \in \boldsymbol{\Theta}} (s_i \lambda_i^\natural - s_i' \widehat{\lambda}_i)^2} \right|.
\end{aligned}
$$

Since $|s_i| \leq \zeta L_g$, $|s_i'| \leq \zeta L_\theta$, and $\lambda_i \leq F\alpha_u^K$, $\forall \underline{\boldsymbol{\Lambda}} \in \mathcal{L}$ by (23), the above relation can be further expressed as

$$
\begin{aligned}
(\eta + \tau(\boldsymbol{\Theta}; \underline{\boldsymbol{M}}^\natural, \widehat{\underline{\boldsymbol{M}}}))\sqrt{|\boldsymbol{\Theta}|} + \zeta(L_g + L_\theta)F\alpha_u^K\sqrt{|\boldsymbol{\Theta}|} &\geq \xi\sqrt{\sum_{i \in \boldsymbol{\Theta}} (\lambda_i^\natural - \xi'/\xi\widehat{\lambda}_i)^2} \\
&= \xi \left\| \left[\underline{\boldsymbol{\Lambda}}^\natural - \xi'/\xi\widehat{\underline{\boldsymbol{\Lambda}}} \right]_{\boldsymbol{\Theta}} \right\|_{\mathrm{F}} \\
\implies \frac{\left\| \left[\underline{\boldsymbol{\Lambda}}^\natural - \xi'/\xi\widehat{\underline{\boldsymbol{\Lambda}}} \right]_{\boldsymbol{\Theta}} \right\|_{\mathrm{F}}}{\sqrt{|\boldsymbol{\Theta}|}} &\leq \frac{(\eta + \tau(\boldsymbol{\Theta}; \underline{\boldsymbol{M}}^\natural, \widehat{\underline{\boldsymbol{M}}})) + \zeta(L_g + L_\theta)F\alpha_u^K}{\xi} \\
&\leq \frac{(\eta + \tau(\boldsymbol{\Theta}; \underline{\boldsymbol{M}}^\natural, \widehat{\underline{\boldsymbol{M}}})) + \zeta(L_g + L_\theta)F\alpha_u^K}{p_{\min}}. \quad (27)
\end{aligned}
$$

Let $\xi'/\xi = \widehat{\xi}$. Our next goal is to characterize $\left\| \underline{\boldsymbol{\Lambda}}^\natural - \widehat{\xi}\widehat{\underline{\boldsymbol{\Lambda}}} \right\|_{\mathrm{F}}$ using the result in (27). To proceed, consider the term

$$
\tau(\boldsymbol{\Theta}; \underline{\boldsymbol{\Lambda}}^\natural, \widehat{\xi}\widehat{\underline{\boldsymbol{\Lambda}}}) = \left| \frac{\left\| \left[\underline{\boldsymbol{\Lambda}}^\natural - \widehat{\xi}\widehat{\underline{\boldsymbol{\Lambda}}} \right]_{\boldsymbol{\Theta}} \right\|_{\mathrm{F}}}{\sqrt{|\boldsymbol{\Theta}|}} - \frac{\left\| \underline{\boldsymbol{\Lambda}}^\natural - \widehat{\xi}\widehat{\underline{\boldsymbol{\Lambda}}} \right\|_{\mathrm{F}}}{\sqrt{\prod_k I_k}} \right|. \quad (28)
$$

We again invoke Lemma D.2 for characterizing the term $\tau(\boldsymbol{\Theta}; \underline{\boldsymbol{\Lambda}}^\natural, \widehat{\xi}\widehat{\underline{\boldsymbol{\Lambda}}})$, and obtain that with probability of at least $1 - \delta$,

$$
\tau(\boldsymbol{\Theta}; \underline{\boldsymbol{\Lambda}}^\natural, \widehat{\xi}\widehat{\underline{\boldsymbol{\Lambda}}}) \leq F\alpha_u^K \left(\log\left(\frac{2}{\delta}\right)\left(1 - \frac{(|\boldsymbol{\Theta}| - 1)}{\prod_k I_k}\right)\frac{1}{2|\boldsymbol{\Theta}|} \right)^{1/4}. \quad (29)
$$

Combining the result in (28) and (29) with (27), we obtain the following result with probability at least $1 - 4\delta - 3e^{-\alpha(e^2-3)}$:

$$
\frac{\left\| \underline{\boldsymbol{\Lambda}}^\natural - \widehat{\xi}\widehat{\underline{\boldsymbol{\Lambda}}} \right\|_{\mathrm{F}}}{\sqrt{\prod_k I_k}} \leq \underbrace{\tau(\boldsymbol{\Theta}; \underline{\boldsymbol{\Lambda}}^\natural, \widehat{\xi}\widehat{\underline{\boldsymbol{\Lambda}}}) + \frac{(\eta + \tau(\boldsymbol{\Theta}; \underline{\boldsymbol{M}}^\natural, \widehat{\underline{\boldsymbol{M}}})) + \zeta(L_g + L_\theta)F\alpha_u^K}{p_{\min}}}_{\eta'}.
$$

17

Recovering p_i's: Next, we proceed to bound the estimation accuracies for p_i's. Consider the following chain of inequalities:

$$
\begin{aligned}
\|\underline{\boldsymbol{M}}^{\natural} - \widehat{\underline{\boldsymbol{M}}}\|_{\mathrm{F}} &= \left\|\boldsymbol{\Lambda}^{\natural} \circledast \underline{\boldsymbol{P}}^{\natural} - \widehat{\xi}\widehat{\boldsymbol{\Lambda}} \circledast {}^{1}\!/\widehat{\xi}\underline{\widehat{\boldsymbol{P}}}\right\|_{\mathrm{F}} \\
&= \left\|\boldsymbol{\Lambda}^{\natural} \circledast \underline{\boldsymbol{P}}^{\natural} - \boldsymbol{\Lambda}^{\natural} \circledast {}^{1}\!/\widehat{\xi}\underline{\widehat{\boldsymbol{P}}} + \boldsymbol{\Lambda}^{\natural} \circledast {}^{1}\!/\widehat{\xi}\underline{\widehat{\boldsymbol{P}}} - \widehat{\xi}\widehat{\boldsymbol{\Lambda}} \circledast {}^{1}\!/\widehat{\xi}\underline{\widehat{\boldsymbol{P}}}\right\|_{\mathrm{F}} \\
&\geq \left| \left\|\boldsymbol{\Lambda}^{\natural} \circledast (\underline{\boldsymbol{P}}^{\natural} - \circledast {}^{1}\!/\widehat{\xi}\underline{\widehat{\boldsymbol{P}}})\right\|_{\mathrm{F}} - \|(\boldsymbol{\Lambda}^{\natural} - \widehat{\xi}\widehat{\boldsymbol{\Lambda}}) \circledast {}^{1}\!/\widehat{\xi}\underline{\widehat{\boldsymbol{P}}}\|_{\mathrm{F}} \right| \\
&\geq \lambda_{\min} \left\|\underline{\boldsymbol{P}}^{\natural} - {}^{1}\!/\widehat{\xi}\underline{\widehat{\boldsymbol{P}}}\right\|_{\mathrm{F}} - \|(\boldsymbol{\Lambda}^{\natural} - \widehat{\xi}\widehat{\boldsymbol{\Lambda}}) \circledast {}^{1}\!/\widehat{\xi}\underline{\widehat{\boldsymbol{P}}}\|_{\mathrm{F}} \\
&\geq \frac{\beta}{p_{\max}} \left\|\underline{\boldsymbol{P}}^{\natural} - {}^{1}\!/\widehat{\xi}\underline{\widehat{\boldsymbol{P}}}\right\|_{\mathrm{F}} - \|(\boldsymbol{\Lambda}^{\natural} - \widehat{\xi}\widehat{\boldsymbol{\Lambda}}) \circledast {}^{1}\!/\widehat{\xi}\underline{\widehat{\boldsymbol{P}}}\|_{\mathrm{F}},
\end{aligned}
$$

where $\lambda_{\min} = \min_i \lambda_i = \frac{\min_i m_i}{\max_i p_i} = \frac{\beta}{p_{\max}}$. The above inequalities imply that with probability at least $1 - 4\delta - 3e^{-\alpha(e^2-3)}$

$$
\begin{aligned}
\frac{\left\|\underline{\boldsymbol{P}}^{\natural} - {}^{1}\!/\widehat{\xi}\underline{\widehat{\boldsymbol{P}}}\right\|_{\mathrm{F}}}{\sqrt{\prod_k I_k}} &\leq \frac{p_{\max}}{\beta} \left(\frac{\left\|\boldsymbol{\Lambda}^{\natural} - \widehat{\xi}\widehat{\boldsymbol{\Lambda}}\right\|_{\mathrm{F}}}{p_{\min}\sqrt{\prod_k I_k}} + \frac{\|\underline{\boldsymbol{M}}^{\natural} - \widehat{\underline{\boldsymbol{M}}}\|_{\mathrm{F}}}{\sqrt{\prod_k I_k}} \right), \\
&= \frac{p_{\max}}{\beta} \left(\frac{\eta'}{p_{\min}} + \eta \right),
\end{aligned}
\tag{30}
$$

where we have used the result that $\|(\boldsymbol{\Lambda}^{\natural} - \widehat{\xi}\widehat{\boldsymbol{\Lambda}}) \circledast {}^{1}\!/\widehat{\xi}\underline{\widehat{\boldsymbol{P}}}\|_{\mathrm{F}} \leq \frac{1}{p_{\min}} \left\|\boldsymbol{\Lambda} - \widehat{\xi}\widehat{\boldsymbol{\Lambda}}\right\|_{\mathrm{F}}$, since $[{}^{1}\!/\widehat{\xi}\underline{\widehat{\boldsymbol{P}}}]_i = {}^{1}\!/\widehat{\xi}\widehat{p}_i = \xi/\xi'\widehat{p}_i \leq {}^{1}\!/p_{\min}$.

Next, we proceed to characterize the generalization error of the learned nonlinear function $\widehat{g}_{\boldsymbol{\theta}}$. Towards this, we first bound the term $\frac{\|[\underline{\boldsymbol{P}}^{\natural} - {}^{1}\!/\widehat{\xi}\underline{\widehat{\boldsymbol{P}}}]_{\boldsymbol{\Xi}}\|_{\mathrm{F}}}{\sqrt{S}}$ using the result in (30). Let us consider

$$
\tau(\boldsymbol{\Xi}; \underline{\boldsymbol{P}}^{\natural}, {}^{1}\!/\widehat{\xi}\underline{\widehat{\boldsymbol{P}}}) = \left| \frac{\left\|\left[\underline{\boldsymbol{P}}^{\natural} - {}^{1}\!/\widehat{\xi}\underline{\widehat{\boldsymbol{P}}}\right]_{\boldsymbol{\Xi}}\right\|_{\mathrm{F}}}{\sqrt{S}} - \frac{\left\|\underline{\boldsymbol{P}}^{\natural} - {}^{1}\!/\widehat{\xi}\underline{\widehat{\boldsymbol{P}}}\right\|_{\mathrm{F}}}{\sqrt{\prod_k I_k}} \right|.
$$

By applying Lemma D.2 combined with (30), we get that with a probability greater than $1 - 5\delta - 3e^{-\alpha(e^2-3)}$

$$
\frac{\left\|\left[\underline{\boldsymbol{P}}^{\natural} - {}^{1}\!/\widehat{\xi}\underline{\widehat{\boldsymbol{P}}}\right]_{\boldsymbol{\Xi}}\right\|_{\mathrm{F}}}{\sqrt{S}} \leq \tau(\boldsymbol{\Xi}; \underline{\boldsymbol{P}}^{\natural}, {}^{1}\!/\widehat{\xi}\underline{\widehat{\boldsymbol{P}}}) + \frac{p_{\max}}{\beta} \left(\frac{\eta'}{p_{\min}} + \eta \right),
\tag{31}
$$

where

$$
\tau(\boldsymbol{\Xi}; \underline{\boldsymbol{P}}^{\natural}, {}^{1}\!/\widehat{\xi}\underline{\widehat{\boldsymbol{P}}}) \leq p_{\max} \left(\log\left(\frac{2}{\delta}\right) \left(1 - \frac{(S-1)}{\prod_k I_k}\right) \frac{1}{2S} \right)^{1/4}.
$$

Using the result in (31), we proceed to bound the generalization performance of the learned nonlinear function $\widehat{g}_{\boldsymbol{\theta}}$. Towards this goal, consider the following result:

Lemma D.3. *Under the Assumptions given by Theorem 4.5, the following holds with probability greater than $1 - \delta$:*

$$
\mathbb{E}_{\boldsymbol{z} \sim \mathcal{D}} \left[(g^{\natural}(\boldsymbol{z}) - {}^{1}\!/\widehat{\xi}\widehat{g}_{\boldsymbol{\theta}}(\boldsymbol{z}))^2 \right] \leq \frac{\left\|\left[\underline{\boldsymbol{P}}^{\natural} - {}^{1}\!/\widehat{\xi}\underline{\widehat{\boldsymbol{P}}}\right]_{\boldsymbol{\Xi}}\right\|_{\mathrm{F}}^2}{S} + \frac{32 \left(2\|\boldsymbol{Z}\|_{\mathrm{F}} \mathscr{R}_{\mathcal{G}}\right)^{\frac{1}{4}}}{S^{5/8}} + 4\sqrt{\frac{2\log(4/\delta)}{S}},
\tag{32}
$$

where $\boldsymbol{Z} = [\boldsymbol{z}_{i_1}, \ldots, \boldsymbol{z}_{i_S}]$ and $\mathscr{R}_{\mathcal{G}}$ denotes the complexity measure as given by Assumption 4.2.

The proof is given in Appendix K.

Combining Lemma D.3 with (31), we get the final bound for $\mathbb{E}_{\boldsymbol{z} \sim \mathcal{D}} \left[(g^{\natural}(\boldsymbol{z}) - {}^{1}\!/\widehat{\xi}\widehat{g}_{\boldsymbol{\theta}}(\boldsymbol{z}))^2 \right]$ with probability greater than $1 - 6\delta - 3e^{-\alpha(e^2-3)}$.

E. Proof of Theorem D.1

For a set $\boldsymbol{\Omega} = \{\boldsymbol{i}_1, \ldots, \boldsymbol{i}_T\}$, we assume that $\boldsymbol{\Omega} \sim \Pi$ where Π is uniform, i.e., each $y_{\boldsymbol{i}}$ is observed independently at random with probability $1/(\prod_k I_k)$ [2]. Hence, we define the expected risk as follows:

$$D_\Pi(\underline{\boldsymbol{M}}; \underline{\boldsymbol{Y}}) := \mathbb{E}_{\boldsymbol{\Omega} \sim \Pi}\left[f(m_{\boldsymbol{i}}; y_{\boldsymbol{i}}, \boldsymbol{z}_{\boldsymbol{i}}) \right] = \sum_{\boldsymbol{i}} \frac{1}{\prod_k I_k} f(m_{\boldsymbol{i}}; y_{\boldsymbol{i}}, \boldsymbol{z}_{\boldsymbol{i}}). \tag{33}$$

Accordingly, we define the empirical risk as follows:

$$D_{\boldsymbol{\Omega}}(\underline{\boldsymbol{M}}; \underline{\boldsymbol{Y}}) = \frac{1}{T} \sum_{\boldsymbol{i} \in \boldsymbol{\Omega}} f(m_{\boldsymbol{i}}; y_{\boldsymbol{i}}, \boldsymbol{z}_{\boldsymbol{i}}). \tag{34}$$

Eq.(24) implies that

$$D_{\boldsymbol{\Omega}}(\widehat{\underline{\boldsymbol{M}}}; \underline{\boldsymbol{Y}}) \le D_{\boldsymbol{\Omega}}(\widetilde{\underline{\boldsymbol{M}}}; \underline{\boldsymbol{Y}}) \tag{35}$$

where $[\widetilde{\underline{\boldsymbol{M}}}]_{\boldsymbol{i}} = \lambda_{\boldsymbol{i}}^{\natural} \widetilde{g}_{\boldsymbol{\theta}}(\boldsymbol{z}_{\boldsymbol{i}})$ and $\widetilde{g}_{\boldsymbol{\theta}}$ is given by Assumption 4.2.

To proceed, consider the following chain of inequalities:

$$
\begin{aligned}
\mathbb{E}_y[D_\Pi(\widehat{\underline{\boldsymbol{M}}}; \underline{\boldsymbol{Y}}) - D_\Pi(\underline{\boldsymbol{M}}^{\natural}; \underline{\boldsymbol{Y}})] &= \mathbb{E}_y[D_\Pi(\widehat{\underline{\boldsymbol{M}}}; \underline{\boldsymbol{Y}})] - D_{\boldsymbol{\Omega}}(\widehat{\underline{\boldsymbol{M}}}; \underline{\boldsymbol{Y}}) + D_{\boldsymbol{\Omega}}(\underline{\boldsymbol{M}}^{\natural}; \underline{\boldsymbol{Y}}) - \mathbb{E}_y[D_\Pi(\underline{\boldsymbol{M}}^{\natural}; \underline{\boldsymbol{Y}})] \\
&\quad + D_{\boldsymbol{\Omega}}(\widehat{\underline{\boldsymbol{M}}}; \underline{\boldsymbol{Y}}) - D_{\boldsymbol{\Omega}}(\widetilde{\underline{\boldsymbol{M}}}; \underline{\boldsymbol{Y}}) + D_{\boldsymbol{\Omega}}(\widetilde{\underline{\boldsymbol{M}}}; \underline{\boldsymbol{Y}}) - D_{\boldsymbol{\Omega}}(\underline{\boldsymbol{M}}^{\natural}; \underline{\boldsymbol{Y}}) \\
&\le \mathbb{E}_y[D_\Pi(\widehat{\underline{\boldsymbol{M}}}; \underline{\boldsymbol{Y}})] - D_{\boldsymbol{\Omega}}(\widehat{\underline{\boldsymbol{M}}}; \underline{\boldsymbol{Y}}) + D_{\boldsymbol{\Omega}}(\underline{\boldsymbol{M}}^{\natural}; \underline{\boldsymbol{Y}}) - \mathbb{E}_y[D_\Pi(\underline{\boldsymbol{M}}^{\natural}; \underline{\boldsymbol{Y}})] \\
&\quad + |D_{\boldsymbol{\Omega}}(\widetilde{\underline{\boldsymbol{M}}}; \underline{\boldsymbol{Y}}) - D_{\boldsymbol{\Omega}}(\underline{\boldsymbol{M}}^{\natural}; \underline{\boldsymbol{Y}})| \\
&\le \sup_{\underline{\boldsymbol{M}} \in \mathcal{M}} |D_{\boldsymbol{\Omega}}(\underline{\boldsymbol{M}}; \underline{\boldsymbol{Y}}) - \mathbb{E}_y[D_\Pi(\underline{\boldsymbol{M}}; \underline{\boldsymbol{Y}})]| + |D_{\boldsymbol{\Omega}}(\underline{\boldsymbol{M}}^{\natural}; \underline{\boldsymbol{Y}}) - \mathbb{E}_y[D_\Pi(\underline{\boldsymbol{M}}^{\natural}; \underline{\boldsymbol{Y}})]| \\
&\quad + |D_{\boldsymbol{\Omega}}(\widetilde{\underline{\boldsymbol{M}}}; \underline{\boldsymbol{Y}}) - D_{\boldsymbol{\Omega}}(\underline{\boldsymbol{M}}^{\natural}; \underline{\boldsymbol{Y}})|,
\end{aligned}
\tag{36}
$$

where expectation is taken w.r.t. $y_{\boldsymbol{i}}$'s and the first inequality is by (35).

Let us consider the L.H.S. of (36):

$$
\begin{aligned}
\mathbb{E}_y[D_\Pi(\widehat{\underline{\boldsymbol{M}}}; \underline{\boldsymbol{Y}}) - D_\Pi(\underline{\boldsymbol{M}}^{\natural}; \underline{\boldsymbol{Y}})] &= \frac{1}{\prod_k I_k} \sum_{\boldsymbol{i}} \mathbb{E}_y\left[(\widehat{m}_{\boldsymbol{i}} - y_{\boldsymbol{i}} \log \widehat{m}_{\boldsymbol{i}}) - (m_{\boldsymbol{i}}^{\natural} - y_{\boldsymbol{i}} \log m_{\boldsymbol{i}}^{\natural}) \right] \\
&= \frac{1}{\prod_k I_k} \sum_{\boldsymbol{i}} \left[(\widehat{m}_{\boldsymbol{i}} - m_{\boldsymbol{i}}^{\natural} \log \widehat{m}_{\boldsymbol{i}}) - (m_{\boldsymbol{i}}^{\natural} - m_{\boldsymbol{i}}^{\natural} \log m_{\boldsymbol{i}}^{\natural}) \right] \\
&= \frac{1}{\prod_k I_k} \sum_{\boldsymbol{i}} \left[m_{\boldsymbol{i}}^{\natural} \log \frac{m_{\boldsymbol{i}}^{\natural}}{\widehat{m}_{\boldsymbol{i}}} - (m_{\boldsymbol{i}}^{\natural} - \widehat{m}_{\boldsymbol{i}}) \right] \\
&:= \mathsf{KL}(\underline{\boldsymbol{M}}^{\natural} || \widehat{\underline{\boldsymbol{M}}}),
\end{aligned}
\tag{37}
$$

where the second equality utilizes Lemma B.1 to get $\mathbb{E}[y_{\boldsymbol{i}}] = \lambda_{\boldsymbol{i}}^{\natural} p_{\boldsymbol{i}}^{\natural} = m_{\boldsymbol{i}}^{\natural}$.

Upper-bounding the first term on the R.H.S of (36). Next, we characterize the first term on the R.H.S. of (36). Towards this, we invoke the following theorem (Theorem 26.5 in (Shalev-Shwartz & Ben-David, 2014)) :

Theorem 1. (Shalev-Shwartz & Ben-David, 2014, Theorem 26.5) Assume that for all y and for all x, we have $|f(x; y)| \le f_{\max}$. Then for any $\underline{\boldsymbol{M}} \in \mathcal{M}$, the following holds with probability greater than $1 - \delta$:

$$|D_{\boldsymbol{\Omega}}(\underline{\boldsymbol{M}}; \underline{\boldsymbol{Y}}) - \mathbb{E}_y[D_\Pi(\underline{\boldsymbol{M}}; \underline{\boldsymbol{Y}})]| \le 2\mathfrak{R}_T(\mathcal{F}) + 4f_{\max}\sqrt{\frac{2\log(4/\delta)}{T}}, \tag{38}$$

---

[2]Here, the set $\boldsymbol{\Omega}$ can be a multiset, meaning it can contain multiple copies of an index since the the assumed sampling scheme is with replacement.

where $\mathcal{F}$ denotes the set

$$\mathcal{F} \triangleq \left\{ [f(m_{\boldsymbol{i_1}}; y_{\boldsymbol{i_1}}, \boldsymbol{z_{i_1}}), \ldots, f(m_{\boldsymbol{i_T}}; y_{\boldsymbol{i_T}}, \boldsymbol{z_{i_T}})]^\top \mid \underline{\boldsymbol{M}} \in \mathcal{M}, \boldsymbol{i_t} \in \boldsymbol{\Omega} \right\}$$

and $\mathfrak{R}_T(\mathcal{F})$ denotes the empirical Rademacher complexity of the set $\mathcal{F}$.

Applying Theorem 1, with probability greater than $1 - \delta$, we have

$$\sup_{\underline{\boldsymbol{M}} \in \mathcal{M}} |D_{\boldsymbol{\Omega}}(\underline{\boldsymbol{M}}; \underline{\boldsymbol{Y}}) - \mathbb{E}_y[D_\Pi(\underline{\boldsymbol{M}}; \underline{\boldsymbol{Y}})]| \le 2\mathfrak{R}_T(\mathcal{F}) + 4f_{\max}\sqrt{\frac{2\log(4/\delta)}{T}}. \tag{39}$$

Let us characterize $f_{\max}$ in (39).

Lemma E.1. *Let* $f(m_{\boldsymbol{i}}; y_{\boldsymbol{i}}, \boldsymbol{z_i}) = m_{\boldsymbol{i}} - y_{\boldsymbol{i}} \log m_{\boldsymbol{i}}$ *where* $m_{\boldsymbol{i}} = \lambda_{\boldsymbol{i}} p_{\boldsymbol{i}} = \lambda_{\boldsymbol{i}} g_{\boldsymbol{\theta}}(\boldsymbol{z_i})$. *Then, under Assumption 4.1, the maximum and minimum value taken by the function* $f(m_{\boldsymbol{i}}; y_{\boldsymbol{i}}, \boldsymbol{z_i})$, *denoted as* $f_{\max}$ *and* $f_{\min}$ *are given by*

$$f_{\max} = \alpha + c \max\{|\log\beta|, \log\alpha\}$$
$$f_{\min} = \beta - c\log\alpha,$$

with probability greater than $1 - e^{-\alpha(e^2 - 3)}$ *where* $c = \alpha(e^2 - 2)$.

The proof of the lemma is given in Appendix G.

Next, we aim to characterize the Rademacher complexity $\mathfrak{R}_T(\mathcal{F})$. Towards this, we have the following result:

Lemma E.2. *Suppose that the assumptions of Theorem 4.5 hold true. Let* $\boldsymbol{\Omega} = \{\boldsymbol{i_1}, \ldots, \boldsymbol{i_T}\}$ *and* $T = |\boldsymbol{\Omega}|$. *Assume that the observations* $y_{\boldsymbol{i_t}}$*'s are i.i.d. samples under the generative model* (3) *and that* $\boldsymbol{z_{i_1}}, \ldots, \boldsymbol{z_{i_S}}$ *are the set of observed features. Consider the set*

$$\mathcal{F} = \{\boldsymbol{f} \in \mathbb{R}^T \mid \boldsymbol{f} = [f(m_{\boldsymbol{i_1}}; y_{\boldsymbol{i_1}}, \boldsymbol{z_{i_1}}), \ldots, f(m_{\boldsymbol{i_T}}; y_{\boldsymbol{i_T}}, \boldsymbol{z_{i_T}})]^\top \mid m_{\boldsymbol{i_t}} = [\underline{\boldsymbol{M}}]_{\boldsymbol{i_t}}, \boldsymbol{i_t} \in \boldsymbol{\Omega}, \underline{\boldsymbol{M}} \in \mathcal{M}\}.$$

With probability greater than $1 - e^{-\alpha(e^2 - 3)}$, *the empirical Rademacher complexity of* $\mathcal{F}$ *is bounded by*

$$\mathfrak{R}_T(\mathcal{F}) \le \left(\frac{4}{\sqrt{T}} + \frac{12}{\sqrt{T}} \sqrt{F \sum_k I_k \log\left(6KL_f\sqrt{T}(F\alpha_u)^K\right) + \left(F\alpha_u^K L_f \sqrt{T}\|\boldsymbol{Z}\|_F \mathscr{R}_{\mathcal{G}}\right)} \right),$$

where $L_f = 1 + \frac{\alpha(e^2 - 2)}{\beta}$, $\boldsymbol{Z} = [\boldsymbol{z_{i_1}}, \ldots, \boldsymbol{z_{i_S}}] \in \mathbb{R}^{D \times S}$ *and and* $\mathscr{R}_{\mathcal{G}}$ *denotes the complexity measure of the class* $\mathcal{G}$.

The proof is provided in Appendix F. By combining Lemma E.1 and Lemma E.2, we completely characterize (39).

Upper-bounding the second term on the R.H.S of (36). Next, we upper bound the second term on the R.H.S of (36). Let us consider the Hoeffding's inequality

Lemma E.3. *Let* $F_1, \ldots, F_T$ *be independent bounded random variables with* $F_t \in [f_{\min}, f_{\max}]$ *for all* t *where* $-\infty < f_{\min} \le f_{\max} < \infty$. *Then for all* $t \ge 0$,

$$\Pr\left(\frac{1}{T} \sum_{t=1}^T (F_t - \mathbb{E}[F_t]) \ge q \right) \le \exp\left(-\frac{2Tq^2}{(f_{\max} - f_{\min})^2} \right).$$

To use Lemma E.3, let us define the random variable F_t as follows:

$$F_t \triangleq f(m_{\boldsymbol{i_t}}; y_{\boldsymbol{i_t}}, \boldsymbol{z_{i_t}}).$$

Then, invoking Lemma E.3, one can obtain

$$\Pr\left(\left| D_{\boldsymbol{\Omega}}(\underline{\boldsymbol{M}}^\natural; \underline{\boldsymbol{Y}}) - \mathbb{E}_y[D_\Pi(\underline{\boldsymbol{M}}^\natural; \underline{\boldsymbol{Y}})] \right| \ge q \right) \le \exp\left(-\frac{2Tq^2}{(f_{\max} - f_{\min})^2} \right). \tag{40}$$

where $f_{\max}$ and $f_{\min}$ are given by Lemma E.1. Hence, by substituting $q = (f_{\max} - f_{\min})\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2T}}$, where $\delta \in (0, 1)$ in (40), we get that with probability greater than $1 - \delta$

$$\left| D_{\boldsymbol{\Omega}}(\underline{\boldsymbol{M}}^\natural; \underline{\boldsymbol{Y}}) - \mathbb{E}_y[D_\Pi(\underline{\boldsymbol{M}}^\natural; \underline{\boldsymbol{Y}})] \right| \le (f_{\max} - f_{\min})\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2T}}. \tag{41}$$

Upper-bounding the third term on the R.H.S of (36).   Consider the following chain of equations:

$$
\begin{aligned}
\left| D_{\boldsymbol{\Omega}}(\widetilde{\boldsymbol{M}}; \underline{\boldsymbol{Y}}) - D_{\boldsymbol{\Omega}}(\underline{\boldsymbol{M}}^{\natural}; \underline{\boldsymbol{Y}}) \right| &= \left| \frac{1}{T} \sum_{i \in \boldsymbol{\Omega}} f(\widetilde{m}_{\boldsymbol{i}}; y_{\boldsymbol{i}}, \boldsymbol{z}_{\boldsymbol{i}}) - \frac{1}{T} \sum_{i \in \boldsymbol{\Omega}} f(m_{\boldsymbol{i}}^{\natural}; y_{\boldsymbol{i}}, \boldsymbol{z}_{\boldsymbol{i}}) \right| \\
&= \frac{1}{T} \left| \sum_{i \in \boldsymbol{\Omega}} (\widetilde{m}_{\boldsymbol{i}} - y_{\boldsymbol{i}} \log \widetilde{m}_{\boldsymbol{i}}) - (m_{\boldsymbol{i}}^{\natural} - y_{\boldsymbol{i}} \log m_{\boldsymbol{i}}^{\natural}) \right| \\
&= \frac{1}{T} \left| \sum_{i \in \boldsymbol{\Omega}} (\lambda_{\boldsymbol{i}}^{\natural}(\widetilde{g}_{\boldsymbol{\theta}}(\boldsymbol{z}_{\boldsymbol{i}}) - g_{\boldsymbol{\theta}}^{\natural}(\boldsymbol{z}_{\boldsymbol{i}})) - y_{\boldsymbol{i}} \left(\log \lambda_{\boldsymbol{i}}^{\natural} \widetilde{g}_{\boldsymbol{\theta}}(\boldsymbol{z}_{\boldsymbol{i}}) - \log \lambda_{\boldsymbol{i}}^{\natural} g_{\boldsymbol{\theta}}^{\natural}(\boldsymbol{z}_{\boldsymbol{i}}) \right) \right| \\
&= \frac{1}{T} \left| \sum_{i \in \boldsymbol{\Omega}} (\lambda_{\boldsymbol{i}}^{\natural}(\widetilde{g}_{\boldsymbol{\theta}}(\boldsymbol{z}_{\boldsymbol{i}}) - g_{\boldsymbol{\theta}}^{\natural}(\boldsymbol{z}_{\boldsymbol{i}})) - y_{\boldsymbol{i}} \left(\log \widetilde{g}_{\boldsymbol{\theta}}(\boldsymbol{z}_{\boldsymbol{i}}) - \log g_{\boldsymbol{\theta}}^{\natural}(\boldsymbol{z}_{\boldsymbol{i}}) \right) \right| \\
&\leq \max_{\boldsymbol{i}} \lambda_{\boldsymbol{i}}^{\natural} |\widetilde{g}_{\boldsymbol{\theta}}(\boldsymbol{z}_{\boldsymbol{i}}) - g_{\boldsymbol{\theta}}^{\natural}(\boldsymbol{z}_{\boldsymbol{i}})| + \max_{\boldsymbol{i}} y_{\boldsymbol{i}} \frac{|\widetilde{g}_{\boldsymbol{\theta}}(\boldsymbol{z}_{\boldsymbol{i}}) - g_{\boldsymbol{\theta}}^{\natural}(\boldsymbol{z}_{\boldsymbol{i}})|}{p_{\min}}, \\
&\leq (c_{\lambda} + c/p_{\min})\nu, \\
&= F \alpha_u^K (1 + p_{\max}(e^2 - 2)/p_{\min})\nu,
\end{aligned}
\tag{42}
$$

where the first inequality employs the triangle inequality and the Lipschitz continuity of the log function. The first inequality also employs the Assumptions 4.1 that the lowerbound of both $\widetilde{g}_{\boldsymbol{\theta}}(\boldsymbol{z}_{\boldsymbol{i}})$ and $g_{\boldsymbol{\theta}}^{\natural}(\boldsymbol{z}_{\boldsymbol{i}})$ are $p_{\min}$. The second inequality is obtained by applying $c_{\lambda} = F\alpha_u^K$ from Assumption 4.1, $c = \alpha + \alpha(e^2 - 3)$ from Lemma H.1 with probability greater than $1 - e^{-\alpha(e^2 - 3)}$ and $\alpha = F\alpha_u^K p_{\max}$.

Putting Together.   Combining (36), (37) with the upperbounds (39), (41), and (42), we get with probability greater than $1 - 2\delta - 3e^{-\alpha(e^2 - 3)}$.

$$
\mathsf{KL}(\underline{\boldsymbol{M}}^{\natural} \| \widehat{\underline{\boldsymbol{M}}}) \leq 2\mathfrak{R}_T(\mathcal{F}) + (5f_{\max} - f_{\min})\sqrt{\frac{2\log(4/\delta)}{T}} + F\alpha_u^K \left(1 + \frac{p_{\max}(e^2 - 2)}{p_{\min}} \right) \nu,
\tag{43}
$$

where $f_{\max} = \alpha + c \max\{|\log \beta|, \log \alpha\}$ and $f_{\min} = \beta - c \log \alpha$ given by Lemma E.1 .

By (Cao & Xie, 2015, Lemma 8), we have

$$
\mathsf{KL}(\underline{\boldsymbol{M}}^{\natural} \| \widehat{\underline{\boldsymbol{M}}}) \geq \frac{\|\underline{\boldsymbol{M}}^{\natural} - \widehat{\underline{\boldsymbol{M}}}\|_{\mathrm{F}}^2}{C(\alpha, \beta) \prod_k I_k},
\tag{44}
$$

where $C(\alpha, \beta) = \frac{4\alpha\gamma}{1 - e^{-\gamma}}$ and $\gamma = \frac{1}{8\beta}(\alpha - \beta)^2$. Hence, combining (43) and (44), we have the following with probability greater than $1 - 2\delta - 3e^{-\alpha(e^2 - 3)}$:

$$
\frac{\|\underline{\boldsymbol{M}}^{\natural} - \widehat{\underline{\boldsymbol{M}}}\|_{\mathrm{F}}^2}{\prod_k I_k} \leq C(\alpha, \beta) \left(2\mathfrak{R}_T(\mathcal{F}) + (5f_{\max} - f_{\min})\sqrt{\frac{2\log(4/\delta)}{T}} + F\alpha_u^K \left(1 + \frac{p_{\max}(e^2 - 2)}{p_{\min}} \right) \nu \right).
\tag{45}
$$

F. Proof of Lemma E.2

Let us consider the following vector $\boldsymbol{f}$:

$$
\boldsymbol{f} = [f(m_{\boldsymbol{i}_1}; y_{\boldsymbol{i}_1}, \boldsymbol{z}_{\boldsymbol{i}_1}), \ldots, f(m_{\boldsymbol{i}_T}; y_{\boldsymbol{i}_T}, \boldsymbol{z}_{\boldsymbol{i}_T})]^{\top} \in \mathcal{F},
$$

where $i_t \in \boldsymbol{\Omega}$, $\forall t$ and

$$
\mathcal{F} = \{\boldsymbol{f} \in \mathbb{R}^T | \boldsymbol{f} = [f(m_{\boldsymbol{i}_1}; y_{\boldsymbol{i}_1}, \boldsymbol{z}_{\boldsymbol{i}_1}), \ldots, f(m_{\boldsymbol{i}_T}; y_{\boldsymbol{i}_T}, \boldsymbol{z}_{\boldsymbol{i}_T})]^{\top} | m_{\boldsymbol{i}_t} = [\underline{\boldsymbol{M}}]_{\boldsymbol{i}_t}, \underline{\boldsymbol{M}} \in \mathcal{M}\}.
\tag{46}
$$

The Rademacher complexity of set $\mathcal{F}$, denoted as $\mathfrak{R}_T(\mathcal{F})$, is defined as follows:

Definition F.1. (Shalev-Shwartz & Ben-David, 2014) The empirical Rademacher complexity of a set of vectors $\mathcal{F} \subset \mathbb{R}^T$ is defined as follows:

$$\mathfrak{R}_T(\mathcal{F}) = \frac{1}{T}\mathbb{E}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\sum_{i=1}^T \epsilon_i f_i\right], \tag{47}$$

where expectation is w.r.t. ϵ_i's which are i.i.d. Rademacher random variables taking values from $\{-1, 1\}$.

By Definition F.1, the empirical Rademacher complexity of the set $\mathcal{F}$ defined in (46) is given by

$$\mathfrak{R}_T(\mathcal{F}) = \frac{1}{T}\mathbb{E}_\sigma\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\sum_{i=1}^T \sigma_i f(m_{\boldsymbol{i}}; y_{\boldsymbol{i}}, \boldsymbol{z_i})\right]. \tag{48}$$

To characterize the Rademacher complexity $\mathfrak{R}_T(\mathcal{F})$, we proceed to characterize the covering number of $\mathcal{F}$ which is defined as follows:

Definition F.2. (Vershynin, 2012) The ε-net of a set $\mathcal{F}$ represented by $\overline{\mathcal{F}}_\varepsilon$ is a finite set such that for any $\boldsymbol{f} \in \mathcal{F}$, there is a $\overline{\boldsymbol{f}} \in \overline{\mathcal{F}}_\varepsilon$ satisfying

$$\|\boldsymbol{f} - \overline{\boldsymbol{f}}\|_2 \le \varepsilon.$$

The covering number of $\mathcal{F}$ is $\mathsf{N}(\mathcal{F}, \varepsilon) = |\overline{\mathcal{F}}_\varepsilon|$.

Consider the following:

$$\begin{aligned}
\|\boldsymbol{f} - \overline{\boldsymbol{f}}\|_2^2 &= \sum_{\boldsymbol{i}\in\boldsymbol{\Omega}}(f(m_{\boldsymbol{i}}, y_{\boldsymbol{i}}, \boldsymbol{z_i}) - f(\overline{m}_{\boldsymbol{i}}; y_{\boldsymbol{i}}, \boldsymbol{z_i}))^2 \\
&\le \sum_{\boldsymbol{i}\in\boldsymbol{\Omega}} L_f^2 (m_{\boldsymbol{i}} - \overline{m}_{\boldsymbol{i}})^2 \\
&\le L_f^2 T \|[\boldsymbol{M} - \overline{\boldsymbol{M}}]_{\boldsymbol{\Omega}}\|_\mathrm{F}^2 \\
&\le L_f^2 T \|\left[\boldsymbol{M} - \overline{\boldsymbol{M}}\right]_{\boldsymbol{\Xi}}\|_\mathrm{F}^2.
\end{aligned} \tag{49}$$

where $\overline{\boldsymbol{f}} \in \overline{\mathcal{F}}_\varepsilon$ belongs to an ε-net for $\mathcal{F}$, $\overline{\boldsymbol{M}} \in \mathcal{M}_\varepsilon$ belongs to an ε-net of $\mathcal{M}$, $\overline{m}_{\boldsymbol{i}} = [\overline{\boldsymbol{M}}]_{\boldsymbol{i}}$, and $T = |\boldsymbol{\Omega}|$. The last inequality is by applying the assumption that $\boldsymbol{\Omega} \subseteq \boldsymbol{\Xi}$. To characterize the term L_f in (49), we have the following result:

Fact F.3. *Assume that there exist two real numbers $\beta > 0$ and $\alpha < \infty$ such that $\beta \le m_{\boldsymbol{i}} \le \alpha$ for all $\boldsymbol{i} \in \boldsymbol{\Omega}$. Then, the following hold for all $\boldsymbol{i} \in \boldsymbol{\Omega}$ with probability greater than $1 - e^{-\alpha(e^2-3)}$:*

$$|f_{\boldsymbol{i}}'(m_{\boldsymbol{i}}; y_{\boldsymbol{i}}, \boldsymbol{z_i})| \le 1 + \frac{\alpha(e^2 - 2)}{\beta}.$$

The proof is provided in Appendix H.

Hence, applying Fact F.3, in (49), we have the following relation with probability greater than $1 - e^{-\alpha(e^2-3)}$,

$$L_f = 1 + (\alpha(e^2 - 2))/\beta. \tag{50}$$

Next, we consider the term $\|[\boldsymbol{M} - \overline{\boldsymbol{M}}]_{\boldsymbol{\Xi}}\|_\mathrm{F}^2$ in (49):

$$\begin{aligned}
\|[\boldsymbol{M} - \overline{\boldsymbol{M}}]_{\boldsymbol{\Xi}}\|_\mathrm{F} &= \left\|\left[\boldsymbol{\Lambda}\circledast\boldsymbol{P} - \overline{\boldsymbol{\Lambda}}\circledast\overline{\boldsymbol{P}}\right]_{\boldsymbol{\Xi}}\right\|_\mathrm{F} \\
&= \left\|\left[\boldsymbol{\Lambda}\circledast\boldsymbol{P} + \overline{\boldsymbol{\Lambda}}\circledast\boldsymbol{P} - \overline{\boldsymbol{\Lambda}}\circledast\boldsymbol{P} - \overline{\boldsymbol{\Lambda}}\circledast\overline{\boldsymbol{P}}\right]_{\boldsymbol{\Xi}}\right\|_\mathrm{F} \\
&\le \left\|\left[(\boldsymbol{\Lambda} - \overline{\boldsymbol{\Lambda}})\circledast\boldsymbol{P}\right]_{\boldsymbol{\Xi}}\right\|_\mathrm{F} + \left\|\left[\overline{\boldsymbol{\Lambda}}\circledast(\boldsymbol{P} - \overline{\boldsymbol{P}})\right]_{\boldsymbol{\Xi}}\right\|_\mathrm{F} \\
&\le \max_{\boldsymbol{i}\in\boldsymbol{\Xi}} p_{\boldsymbol{i}} \|[\boldsymbol{\Lambda} - \overline{\boldsymbol{\Lambda}}]_{\boldsymbol{\Xi}}\|_\mathrm{F} + \max_{\boldsymbol{i}\in\boldsymbol{\Xi}} \overline{\lambda}_{\boldsymbol{i}} \|[\boldsymbol{P} - \overline{\boldsymbol{P}}]_{\boldsymbol{\Xi}}\|_\mathrm{F} \\
&\le \max_{\boldsymbol{i}} p_{\boldsymbol{i}} \|\boldsymbol{\Lambda} - \overline{\boldsymbol{\Lambda}}\|_\mathrm{F} + \max_{\boldsymbol{i}} \overline{\lambda}_{\boldsymbol{i}} \|g_{\boldsymbol{\theta}}([\boldsymbol{Z}]_{\boldsymbol{\Xi}}) - \overline{g}_{\boldsymbol{\theta}}([\boldsymbol{Z}]_{\boldsymbol{\Xi}})\|_\mathrm{F} \\
&\le \|\boldsymbol{\Lambda} - \overline{\boldsymbol{\Lambda}}\|_\mathrm{F} + c_\lambda \|g_{\boldsymbol{\theta}}([\boldsymbol{Z}]_{\boldsymbol{\Xi}}) - \overline{g}_{\boldsymbol{\theta}}([\boldsymbol{Z}]_{\boldsymbol{\Xi}})\|_\mathrm{F},
\end{aligned} \tag{51}$$

22

where $g_{\boldsymbol{\theta}}([\boldsymbol{Z}]_{\boldsymbol{\Xi}}) = [g_{\boldsymbol{\theta}}(\boldsymbol{z}_{i_1}), \ldots, g_{\boldsymbol{\theta}}(\boldsymbol{z}_{i_S})]^{\top}$, $g_{\boldsymbol{\theta}} \in \mathcal{G}$, $\overline{g}_{\boldsymbol{\theta}}([\boldsymbol{Z}]_{\boldsymbol{\Xi}}) = [\overline{g}_{\boldsymbol{\theta}}(\boldsymbol{z}_{i_1}), \ldots, \overline{g}_{\boldsymbol{\theta}}(\boldsymbol{z}_{i_S})]^{\top}$, $\overline{g}_{\boldsymbol{\theta}} \in \overline{\mathcal{G}}_{\varepsilon}$ belongs to the ε-net of $\mathcal{G}$, $\boldsymbol{z}_{i_1}, \ldots, \boldsymbol{z}_{i_S}$ are the set of observed features, and $\underline{\boldsymbol{\Lambda}} \in \overline{\mathcal{L}}_{\varepsilon}$ belongs to an ε-net of $\mathcal{L}$. The first inequality uses the triangle inequality and the last inequality uses the facts that $\max_i p_i \leq 1$ and $\max_i \overline{\lambda}_i \leq c_{\lambda}$ where we have $c_{\lambda} = F\alpha_u^K$ under Assumption 4.1.

Eq. (49) and (51) imply that that if there exists an $\varepsilon/2c_{\lambda}L_f\sqrt{T}$-net covering for $\mathcal{G} \circ \boldsymbol{\Xi}$, where

$$\mathcal{G} \circ \boldsymbol{\Xi} = \{g_{\boldsymbol{\theta}}([\boldsymbol{Z}]_{\boldsymbol{\Xi}}) = [g(\boldsymbol{z}_{i_1}), \ldots, g(\boldsymbol{z}_{i_S})]^{\top}, \, g_{\boldsymbol{\theta}} \in \mathcal{G}\}, \tag{52}$$

and an $\varepsilon/2L_f\sqrt{T}$-net covering for $\mathcal{L}$, then we can construct an ε-net to cover $\mathcal{F}$. Since $\underline{\boldsymbol{\Lambda}}$ which is a low-rank tensor, we invoke the following result to get the covering number of for $\mathcal{L}$:

Lemma F.4. *(Fan et al., 2020) Let* $\mathcal{L} = \{\boldsymbol{X} \in \mathbb{R}^{I_1 \times \ldots \times I_K} \mid \boldsymbol{X} = \sum_{f=1}^{F} \boldsymbol{U}_1(:,f) \circ \ldots \circ \boldsymbol{U}_K(:,f), \, \|\boldsymbol{U}_k(:,f)\|_2 \leq u\}$. *Then, the covering number of* $\mathcal{L}$ *with respect to the Frobenius norm satisfies*

$$\mathsf{N}(\mathcal{L}, \varepsilon) \leq \left(\frac{3K}{\varepsilon}(Fu^2)^{K/2}\right)^{F\sum_k I_k}. \tag{53}$$

The proof of Lemma F.4 is also detailed in Sec. I.

Invoking Lemma F.4, we have:

$$\mathsf{N}(\mathcal{L}, \varepsilon/2L_f\sqrt{T}) \leq \left(\frac{6KL_f\sqrt{T}(F\alpha_u)^K}{\varepsilon}\right)^{F\sum_k I_k}, \tag{54}$$

where we have applied $\|\boldsymbol{U}_k(:,f)\|_2 \leq \sqrt{F}\alpha_u$. The covering number of the set $\mathcal{G} \circ \boldsymbol{\Xi}$ is given by Lemma 14 of (Lin & Zhang, 2019) (cf. Lemma M.1 in Sec. M) as below:

$$\mathsf{N}(\mathcal{G} \circ \boldsymbol{\Xi}, \varepsilon) \leq \exp\left(\frac{\|\boldsymbol{Z}\|_F \mathscr{R}_{\mathcal{G}}}{\varepsilon}\right)^{1/2}, \tag{55}$$

where $\boldsymbol{Z} = [\boldsymbol{z}_{i_1}, \ldots, \boldsymbol{z}_{i_S}] \in \mathbb{R}^{D \times S}$ and $\mathscr{R}_{\mathcal{G}}$ denotes the complexity measure of the class $\mathcal{G}$. Applying (55), we get

$$\mathsf{N}(\mathcal{G} \circ \boldsymbol{\Xi}, \varepsilon/2c_{\lambda}L_f\sqrt{T}) \leq \exp\left(\frac{2c_{\lambda}L_f\sqrt{T}\|\boldsymbol{Z}\|_F \mathscr{R}_{\mathcal{G}}}{\varepsilon}\right)^{1/2}. \tag{56}$$

From (54) and (56), combined with (49) and (51), one can obtain that

$$\begin{aligned}
\mathsf{N}(\mathcal{F}, \varepsilon) &\leq \mathsf{N}(\mathcal{L}, \varepsilon/2L_f\sqrt{T}) \times \mathsf{N}(\mathcal{G}, \varepsilon/2c_{\lambda}L_f\sqrt{T}) \\
&\leq \left(\frac{6KL_f\sqrt{T}(F\alpha_u)^K}{\varepsilon}\right)^{F\sum_k I_k} \times \exp\left(\frac{2c_{\lambda}L_f\sqrt{T}\|\boldsymbol{Z}\|_F \mathscr{R}_{\mathcal{G}}}{\varepsilon}\right)^{1/2}.
\end{aligned} \tag{57}$$

To characterize the Rademacher complexity $\mathfrak{R}_T(\mathcal{F})$ using the covering number $\mathsf{N}(\mathcal{F}, \varepsilon)$, we invoke the following lemma:

Lemma F.5. *(Bartlett et al., 2017, Lemma A.5) The empirical Rademacher complexity of the set* $\mathcal{F} \subset \mathbb{R}^T$ *is upper bounded as follows:*

$$\mathfrak{R}_T(\mathcal{F}) \leq \inf_{a>0} \left(\frac{4a}{\sqrt{T}} + \frac{12}{T}\int_a^{\sqrt{T}} \sqrt{\log \mathsf{N}(\mathcal{F}, \mu)}d\mu\right). \tag{58}$$

Applying Lemma F.5, we have the following set of relations:

$$
\begin{aligned}
\mathfrak{R}_T(\mathcal{F}) &\leq \inf_{a>0} \left(\frac{4a}{\sqrt{T}} + \frac{12}{T} \int_a^{\sqrt{T}} \sqrt{\log \mathsf{N}(\mathcal{F}, \mu)} d\mu \right) \\
&\leq \inf_{a>0} \left(\frac{4a}{\sqrt{T}} + \frac{12}{\sqrt{T}} \sqrt{\log \mathsf{N}(\mathcal{F}, a)} \right) \\
&\leq \inf_{a>0} \left(\frac{4a}{\sqrt{T}} + \frac{12}{\sqrt{T}} \sqrt{F \sum_k I_k \log \left(\frac{6KL_f \sqrt{T}(F\alpha_u)^K}{a} \right) + \left(\frac{c_\lambda L_f \sqrt{T} \|\boldsymbol{Z}\|_{\mathrm{F}} \mathscr{R}_{\mathcal{G}}}{a} \right)} \right) \\
&\leq \left(\frac{4}{\sqrt{T}} + \frac{12}{\sqrt{T}} \sqrt{F \sum_k I_k \log \left(6KL_f \sqrt{T}(F\alpha_u)^K \right) + \left(F\alpha_u^K L_f \sqrt{T} \|\boldsymbol{Z}\|_{\mathrm{F}} \mathscr{R}_{\mathcal{G}} \right)} \right),
\end{aligned}
\tag{59}
$$

where the second inequality is obtained since $\sqrt{\log \mathsf{N}(\mathcal{F}, \mu)}$ increases monotonically with the decrease of μ and hence

$$
\int_a^{\sqrt{T}} \sqrt{\log \mathsf{N}(\mathcal{F}, \mu)} d\mu \;\leq\; \sqrt{T} \sqrt{\log \mathsf{N}(\mathcal{F}, a)},
$$

the third inequality is by applying (57), the last inequality is by setting $a = 1$, applying $c_\lambda = F\alpha_u^K$ and L_f is given by (50) with probability greater than $1 - e^{-\alpha(e^2-3)}$.

G. Proof of Lemma E.1

We start by noting that $\beta \leq m_i \leq \alpha$ (see (22) and (23)). By Lemma H.1, we have

$$
\Pr(y_i \leq c) \geq 1 - e^{-\alpha(e^2-3)}, \quad c = \alpha(e^2 - 2). \tag{60}
$$

Also, if $\beta < 1$, we have $f_i = m_i - y_i \log m_i \leq \alpha + c|\log \beta|$. Suppose $\beta \geq 1$, we have $f_i = m_i - y \log m_i \leq m_i + y_i \log m_i \leq \alpha + c \log \alpha$. Similarly, we get that $f_i \geq \beta - c \log \alpha$. Hence, we get

$$
\begin{aligned}
f_{\max} &\leq \alpha + c \max\{|\log \beta|, \log \alpha\} \\
f_{\min} &\leq \beta - c \log \alpha,
\end{aligned}
$$

which hold with probability greater than $1 - e^{-\alpha(e^2-3)}$.

H. Proof of Fact F.3

We have

$$
f_i(m_i) = m_i - y_i \log m_i.
$$

Hence, we get

$$
f_i'(m_i) = 1 - \frac{y_i}{m_i}. \tag{61}
$$

To proceed, we invoke the following lemma:

Lemma H.1. *(Cao & Xie, 2015), Lemma 9 (Tail Bound of Poisson) For $y \sim \mathsf{Poisson}(m)$ with $m \leq \alpha$, we have*

$$
\Pr(y - m \geq t) \leq e^{-t} \tag{62}
$$

for all $t \geq t_0$ where $t_0 = \alpha(e^2 - 3)$, where e is the Euler's number.

Utilizing Lemma H.1, it can be seen that

$$\mathsf{Pr}(y_i - \alpha \geq \alpha(e^2 - 3)) \leq e^{-\alpha(e^2 - 3)}$$
$$\implies \mathsf{Pr}(y_i \leq \alpha(e^2 - 3) + \alpha) \geq 1 - e^{-\alpha(e^2 - 3)}.$$

Combining this result with (61), we get that with probability greater than $1 - e^{-\alpha(e^2 - 3)}$,

$$|f'_i(m_i)| \leq 1 + \frac{y_i}{m_i} \leq 1 + \frac{\alpha(e^2 - 2)}{\beta},$$

where the last inequality uses the assumption that $m_i \geq \beta$.

I. Proof of Lemma F.4

Let $\overline{\mathcal{L}}_\varepsilon$ denote the ε-net of $\mathcal{L}$. Consider $\underline{\overline{\boldsymbol{X}}} \in \mathcal{L}_\varepsilon$. In addition, let us define the following set:

$$\mathcal{R}_k = \{\boldsymbol{U} \in \mathbb{R}^{I_k \times F} \mid \|\boldsymbol{U}(:, f)\|_2 \leq u\}.$$

Let $\overline{\boldsymbol{U}}_k$ belongs to ε-net of $\mathcal{R}_k$ such that $\|\boldsymbol{U}_k - \overline{\boldsymbol{U}}_k\|_{\mathrm{F}} \leq \varepsilon$. Then, the cardinality of the ε-net of $\mathcal{R}_k$ is given by (Wainwright, 2019)

$$\mathsf{N}(\mathcal{R}_k, \varepsilon) \leq \left(\frac{3\sqrt{Fu^2}}{\varepsilon}\right)^{I_k F}. \tag{63}$$

In order to derive the covering number of the set $\mathcal{L}$, we consider the following chain of relations:

$$\|\underline{\boldsymbol{X}} - \underline{\overline{\boldsymbol{X}}}\|_{\mathrm{F}} = \left\|\sum_{f=1}^{F} \boldsymbol{U}_1(:, f) \circ \ldots \circ \boldsymbol{U}_K(:, f)\|_{\mathrm{F}} - \sum_{f=1}^{F} \overline{\boldsymbol{U}}_1(:, f) \circ \ldots \circ \overline{\boldsymbol{U}}_K(:, f)\right\|_{\mathrm{F}} \tag{64a}$$

$$\leq \left\|\sum_{f=1}^{F} \left(\boldsymbol{U}_1(:, f) - \overline{\boldsymbol{U}}_1(:, f)\right) \circ \boldsymbol{U}_2(:, f) \ldots \circ \boldsymbol{U}_K(:, f)\right\|_{\mathrm{F}}$$
$$+ \left\|\sum_{f=1}^{F} \overline{\boldsymbol{U}}_1(:, f) \circ \left(\sum_{f=1}^{F} \boldsymbol{U}_2(:, f) \ldots \circ \boldsymbol{U}_K(:, f) - \overline{\boldsymbol{U}}_2(:, f) \ldots \circ \overline{\boldsymbol{U}}_K(:, f)\right)\right\|_{\mathrm{F}} \tag{64b}$$

$$\leq (\sqrt{Fu^2})^{K-1} \left\|\sum_{f=1}^{F} \left(\boldsymbol{U}_1(:, f) - \overline{\boldsymbol{U}}_1(:, f)\right)\right\|_{\mathrm{F}}$$
$$+ \sqrt{Fu^2} \underbrace{\left\|\sum_{f=1}^{F} \boldsymbol{U}_2(:, f) \circ \ldots \circ \boldsymbol{U}_K(:, f)\|_{\mathrm{F}} - \sum_{f=1}^{F} \overline{\boldsymbol{U}}_2(:, f) \circ \ldots \circ \overline{\boldsymbol{U}}_K(:, f)\right\|_{\mathrm{F}}}_{Q_{\boldsymbol{U}_1}} \tag{64c}$$

$$\leq (\sqrt{Fu^2})^{K-1} \|\boldsymbol{U}_1 - \overline{\boldsymbol{U}}_1\|_{\mathrm{F}} + \sqrt{Fu^2} Q_{\boldsymbol{U}_1}. \tag{64d}$$

In a similar way as followed in (64), we can derive upper-bounds for every $Q_{\boldsymbol{U}_k}, k \in [K]$. Then, we can finally establish the below relationship:

$$\|\underline{\boldsymbol{X}} - \underline{\overline{\boldsymbol{X}}}\|_{\mathrm{F}} \leq (\sqrt{Fu^2})^{K-1} \sum_{k=1}^{K} \|\boldsymbol{U}_k - \overline{\boldsymbol{U}}_k\|_{\mathrm{F}}. \tag{65}$$

The result in (65) implies that if there exists an $\varepsilon/K(\sqrt{Fu^2})^{K-1}$-net to cover each $\mathcal{R}_k$, then we can construct an ε-net to cover $\mathcal{L}$. Then, the cardinality of the ε-net of $\mathcal{L}$ is given by

$$\mathsf{N}(\mathcal{L}, \varepsilon) \leq \prod_{k=1}^{K} \mathsf{N}\left(\mathcal{R}_k, \varepsilon/K(\sqrt{Fu^2})^{K-1}\right)$$

$$\leq \left(\frac{3K}{\varepsilon}(Fu^2)^{K/2}\right)^{F\sum_k I_k},$$

where the inequality is by applying (63). Hence the proof.

J. Proof of Lemma D.2

Let us define

$$\widehat{u} = \frac{1}{|\boldsymbol{\Theta}|} \left\|[\boldsymbol{M}_1 - \boldsymbol{M}_2]_{\boldsymbol{\Theta}}\right\|_{\mathrm{F}}^2, \tag{66a}$$

$$u = \frac{1}{\prod_k I_k} \left\|\boldsymbol{M}_1 - \boldsymbol{M}_2\right\|_{\mathrm{F}}^2. \tag{66b}$$

Next, we consider the following lemma which is the Serfling's sampling-without-replacement extension of the Hoeffding's inequality (Serfling, 1974):

Lemma J.1. *Let $X_1, X_2, \ldots, X_M$ be a set of samples taken without replacement from $\{x_1, x_2, \ldots, x_N\}$ of mean μ. Denote $a = \min_i x_i$ and $b = \max_i x_i$. Then*

$$\mathsf{Pr}\left[\left|\frac{1}{M}\sum_{i=1}^{M} X_i - \mu\right| \geq t\right]$$

$$\leq 2\exp\left(-\frac{2Mt^2}{(1 - (M-1)/N)(b-a)^2}\right).$$

One can see that u in (66b) denotes the mean of $\prod_k I_k$ terms of $(\widehat{m}_{\boldsymbol{i}} - m_{\boldsymbol{i}}^{\natural})^2$, $\widehat{u}$ denotes the mean of $|\boldsymbol{\Theta}|$ samples randomly drawn from $\{(\widehat{m}_{\boldsymbol{i}} - m^{\natural})^2\}$ without replacement. Hence, using the assumption that $\boldsymbol{i} \in \boldsymbol{\Theta}$ is drawn uniformly at random from $[I_1] \times \ldots \times [I_K]$ and invoking Lemma J.1, we have

$$\mathsf{Pr}\left[|\widehat{u} - u| \geq t\right] \leq 2\exp\left(-\frac{2|\boldsymbol{\Theta}|t^2}{(1 - (|\boldsymbol{\Theta}| - 1)/\prod_k I_k)\alpha^4}\right).$$

where we also applied $(\widehat{m}_{\boldsymbol{i}} - m_{\boldsymbol{i}}^{\natural})^2 \leq \alpha^2$.

Let $\delta = 2\exp\left(-\frac{2|\boldsymbol{\Theta}|t^2}{(1-(|\boldsymbol{\Theta}|-1)/\prod_k I_k)\alpha^4}\right)$, we have the following result with probability $1 - \delta$:

$$|\widehat{u} - u| \leq \alpha^2 \sqrt{\log\left(\frac{2}{\delta}\right)\left(1 - \frac{(|\boldsymbol{\Theta}| - 1)}{\prod_k I_k}\right)\frac{1}{2|\boldsymbol{\Theta}|}}. \tag{67}$$

Using $|\sqrt{a} - \sqrt{b}| \leq \sqrt{|a - b|}$ for nonnegative a and b, we have

$$\left|\sqrt{\widehat{u}} - \sqrt{u}\right| \leq \sqrt{|\widehat{u} - u|},$$

which implies that

$$\left|\sqrt{\widehat{u}} - \sqrt{u}\right| \leq \alpha \left(\log\left(\frac{2}{\delta}\right)\left(1 - \frac{(|\boldsymbol{\Theta}| - 1)}{\prod_k I_k}\right)\frac{1}{2|\boldsymbol{\Theta}|}\right)^{1/4}, \tag{68}$$

holds with probability greater than $1 - \delta$.

K. Proof of Lemma D.3

First, let us define the following notations w.r.t the loss function $\ell(p_1, p_2) = (p_1 - p_2)^2$, $p_1, p_2 \in [0, 1]$ and the observed features $\boldsymbol{Z} = [\boldsymbol{z}_{\boldsymbol{i}_1}, \ldots, \boldsymbol{z}_{\boldsymbol{i}_S}]$ where each $\boldsymbol{z}_{\boldsymbol{i}_s}$ is sampled i.i.d. from the distribution $\mathcal{D}$:

$$L_{\boldsymbol{Z}}(\widehat{g}_{\boldsymbol{\theta}}) \triangleq \frac{1}{S} \sum_{s=1}^{S} \ell(g^{\natural}(\boldsymbol{z}_{\boldsymbol{i}_s}), {}^1\!/\widehat{\varepsilon}\widehat{g}_{\boldsymbol{\theta}}(\boldsymbol{z}_{\boldsymbol{i}_s})) = \frac{1}{S} \sum_{s=1}^{S} (g^{\natural}(\boldsymbol{z}_{\boldsymbol{i}_s}) - {}^1\!/\widehat{\varepsilon}\widehat{g}_{\boldsymbol{\theta}}(\boldsymbol{z}_{\boldsymbol{i}_s}))^2 = \frac{\left\| \left[\boldsymbol{P}^{\natural} - {}^1\!/\widehat{\varepsilon}\widehat{\boldsymbol{P}} \right]_{\boldsymbol{\Xi}} \right\|_{\mathrm{F}}^2}{S} \tag{69a}$$

$$L_{\mathcal{D}}(\widehat{g}_{\boldsymbol{\theta}}) \triangleq \mathbb{E}_{\boldsymbol{z} \sim \mathcal{D}} \left[\ell(g^{\natural}(\boldsymbol{z}), {}^1\!/\widehat{\varepsilon}\widehat{g}_{\boldsymbol{\theta}}(\boldsymbol{z})) \right] = \mathbb{E}_{\boldsymbol{z} \sim \mathcal{D}} \left[(g^{\natural}(\boldsymbol{z}) - {}^1\!/\widehat{\varepsilon}\widehat{g}_{\boldsymbol{\theta}}(\boldsymbol{z}))^2 \right]. \tag{69b}$$

We invoke Theorem 26.5 in (Shalev-Shwartz & Ben-David, 2014) and get that with probability greater than $1 - \delta$:

$$L_{\mathcal{D}}(\widehat{g}_{\boldsymbol{\theta}}) \leq L_{\boldsymbol{Z}}(\widehat{g}_{\boldsymbol{\theta}}) + 2\mathfrak{R}_S(\ell \circ \mathcal{G} \circ \boldsymbol{\Xi}) + 4\bar{c}\sqrt{\frac{2\log(4/\delta)}{S}}$$

$$\implies \mathbb{E}_{\boldsymbol{z} \sim \mathcal{D}} \left[(g^{\natural}(\boldsymbol{z}) - {}^1\!/\widehat{\varepsilon}\widehat{g}_{\boldsymbol{\theta}}(\boldsymbol{z}))^2 \right] \leq \frac{\left\| \left[\boldsymbol{P}^{\natural} - {}^1\!/\widehat{\varepsilon}\widehat{\boldsymbol{P}} \right]_{\boldsymbol{\Xi}} \right\|_{\mathrm{F}}^2}{S} + 2\mathfrak{R}_S(\mathcal{G} \circ \boldsymbol{\Xi}) + 4\sqrt{\frac{2\log(4/\delta)}{S}} \tag{70}$$

where the last inequality utilizes the definitions in (69), the contraction lemma (Lemma 26.9 from (Shalev-Shwartz & Ben-David, 2014)) and also applied $\bar{c} = 1$ since $|\ell(\boldsymbol{f}, \boldsymbol{y})| \leq 1$ in our case. The term $\mathfrak{R}_S(\mathcal{G} \circ \boldsymbol{\Xi})$ denotes the empirical Rademacher complexity of the function class $\mathcal{G} \circ \boldsymbol{\Xi}$ (see its definition in (52)) which is upperbounded via the sensitive complexity parameter $\mathscr{R}_{\mathcal{G}}$ as follows (Lin & Zhang, 2019):

$$\mathfrak{R}_S(\mathcal{G} \circ \boldsymbol{\Xi}) \leq 16S^{-5/8} \left(2\|\boldsymbol{Z}\|_{\mathrm{F}} \mathscr{R}_{\mathcal{G}} \right)^{\frac{1}{4}}.$$

This completes the proof.

L. Proof of Lemma C.1

Consider the following fact due to the cocavity of the $\log$ function and Jensen's inequality (Boyd & Vandenberghe, 2004):

Fact L.1. *Assume that α_f's are certain scalars such that $\alpha_f > 0, \sum_f \alpha_f = 1$. Then, for any $u_f > 0$,*

$$\log \sum_f \alpha_f u_f \geq \sum_f \alpha_f \log u_f.$$

To proceed, we define the following scalars for our case:

$$\alpha_{\boldsymbol{i}}^{(f)} := \frac{\overline{\boldsymbol{U}}_k(i_k, f) \prod_{j \neq k} \boldsymbol{U}_j(i_j, f)}{\sum_{f=1}^{F} \overline{\boldsymbol{U}}_k(i_k, f) \prod_{j \neq k} \boldsymbol{U}_j(i_j, f)}, \ \forall f$$

One can see that $\sum_{f=1}^{F} \alpha_{\boldsymbol{i}}^{(f)} = 1$, $\alpha_{\boldsymbol{i}}^{(f)} > 0$, $\forall f$. Hence, we can use Fact L.1 to have the following inequality:

$$\log \left(\sum_{f=1}^{F} \boldsymbol{U}_k(i_k, f) \prod_{j \neq k} \boldsymbol{U}_j(i_j, f) \right) \geq \sum_{f=1}^{F} \alpha_{\boldsymbol{i}}^{(f)} \log \left(\frac{\boldsymbol{U}_k(i_k, f) \prod_{j \neq k} \boldsymbol{U}_j(i_j, f)}{\alpha_{\boldsymbol{i}}^{(f)}} \right).$$

Applying the above relation, we immediately have

$$w(\boldsymbol{U}_k) \leq s(\boldsymbol{U}_k; \overline{\boldsymbol{U}}_k), \ \ \forall \boldsymbol{U}_k. \tag{71}$$

By substituting $\boldsymbol{U}_k = \overline{\boldsymbol{U}}_k$ in $s(\boldsymbol{U}_k; \overline{\boldsymbol{U}}_k)$, we have

$$s(\overline{\boldsymbol{U}}_k; \overline{\boldsymbol{U}}_k) = \sum_{\boldsymbol{i} \in \boldsymbol{\Omega}} \left[\sum_{f=1}^{F} \overline{\boldsymbol{U}}_k(i_k, f) \prod_{j \neq k} \boldsymbol{U}_j(i_j, f) p_{\boldsymbol{i}} - y_{\boldsymbol{i}} \sum_{f=1}^{F} \alpha_{\boldsymbol{i}}^{(f)} \log \left(\sum_{f=1}^{F} \overline{\boldsymbol{U}}_k(i_k, f) \prod_{j \neq k} \boldsymbol{U}_j(i_j, f) \right) \right]$$

$$= \sum_{\boldsymbol{i} \in \boldsymbol{\Omega}} \left[\left(\sum_{f=1}^{F} \overline{\boldsymbol{U}}_k(i_k, f) \prod_{j \neq k} \boldsymbol{U}_j(i_j, f) \right) p_{\boldsymbol{i}} - y_{\boldsymbol{i}} \log \left(\sum_{f=1}^{F} \overline{\boldsymbol{U}}_k(i_k, f) \prod_{j \neq k} \boldsymbol{U}_j(i_j, f) \right) \right]$$

$$= w(\overline{\boldsymbol{U}}_k), \tag{72}$$

where the second equality is due to $\sum_f \alpha_{\boldsymbol{i}}^{(f)} = 1$.

By taking the derivative of $s(\boldsymbol{U}_k; \overline{\boldsymbol{U}}_k)$ w.r.t. $\boldsymbol{U}_k(i_k, f)$, we have

$$
\begin{aligned}
\nabla_{\boldsymbol{U}_k(i_k,f)} s(\boldsymbol{U}_k; \overline{\boldsymbol{U}}_k) &= \sum_{\boldsymbol{i}_{-k} \in \boldsymbol{\Omega}_{-k}} \left[\prod_{j \neq k} \boldsymbol{U}_j(i_j, f) p_{\boldsymbol{i}} \right. \\
&\left. - y_{\boldsymbol{i}} \alpha_{\boldsymbol{i}}^{(f)} \left(\frac{\alpha_{\boldsymbol{i}}^{(f)}}{\boldsymbol{U}_k(i_k, f) \prod_{j \neq k} \boldsymbol{U}_j(i_j, f)} \right) \left(\frac{\prod_{j \neq k} \boldsymbol{U}_j(i_j, f)}{\alpha_{\boldsymbol{i}}^{(f)}} \right) \right] \\
&= \sum_{\boldsymbol{i}_{-k} \in \boldsymbol{\Omega}_{-k}} \left[\prod_{j \neq k} \boldsymbol{U}_j(i_j, f) p_{\boldsymbol{i}} - y_{\boldsymbol{i}} \left(\frac{\alpha_{\boldsymbol{i}}^{(f)}}{\boldsymbol{U}_k(i_k, f)} \right) \right] \\
&= \sum_{\boldsymbol{i}_{-k} \in \boldsymbol{\Omega}_{-k}} \left[\prod_{j \neq k} \boldsymbol{U}_j(i_j, f) p_{\boldsymbol{i}} \right. \\
&\left. - y_{\boldsymbol{i}} \left(\frac{\overline{\boldsymbol{U}}_k(i_k, f) \prod_{j \neq k} \boldsymbol{U}_j(i_j, f)}{\boldsymbol{U}_k(i_k, f) (\sum_{f=1}^{F} \overline{\boldsymbol{U}}_k(i_k, f) \prod_{j \neq k} \boldsymbol{U}_j(i_j, f))} \right) \right],
\end{aligned}
\tag{73}
$$

where $\boldsymbol{i}_{-k} = (i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_K)$, $\boldsymbol{\Omega}_{-k}$ collects all $\boldsymbol{i}_{-k}$'s such that the corresponding $y_{\boldsymbol{i}}$'s are observed and the last equality (73) is obtained by substituting the definition of $\alpha_{\boldsymbol{i}}^{(f)}$. Further, by taking the derivative of $w(\boldsymbol{U}_k)$ w.r.t. $\boldsymbol{U}_k(i_k, f)$, we get

$$
\nabla_{\boldsymbol{U}_k(i_k,f)} w(\boldsymbol{U}_k) = \sum_{\boldsymbol{i}_{-k} \in \boldsymbol{\Omega}_{-k}} \left[\prod_{j \neq k} \boldsymbol{U}_j(i_j, f) p_{\boldsymbol{i}} - y_{\boldsymbol{i}} \left(\frac{\prod_{j \neq k} \boldsymbol{U}_j(i_j, f)}{(\sum_{f=1}^{F} \overline{\boldsymbol{U}}_k(i_k, f) \prod_{j \neq k} \boldsymbol{U}_j(i_j, f))} \right) \right].
\tag{74}
$$

From (73) and (74), one can see that

$$
\nabla_{\boldsymbol{U}_k(i_k,f)} w(\overline{\boldsymbol{U}}_k) = \nabla_{\boldsymbol{U}_k(i_k,f)} s(\overline{\boldsymbol{U}}_k; \overline{\boldsymbol{U}}_k).
\tag{75}
$$

M. Characterization of the Complexity Measure $\mathscr{R}_{\mathcal{G}}$

The work in (Lin & Zhang, 2019) introduces a complexity measure, called the sensitive complexity and denoted as $\mathscr{R}_{\mathcal{G}}$, to assess the expressive power of various families of neural network classes. For example, regrading the fully connected neural network function classes, the following result is presented:

Lemma M.1. *(Lin & Zhang, 2019, Lemma 14) Consider the following function class:*

$$
\mathcal{G} = \{ g_{\boldsymbol{\theta}} : \mathbb{R}^D \to [0, 1] \mid g_{\boldsymbol{\theta}}(\boldsymbol{z}) = \sigma(\boldsymbol{w}_L^\top \boldsymbol{\sigma}(\boldsymbol{W}_{L-1} \boldsymbol{\sigma}(\ldots \boldsymbol{\sigma}(\boldsymbol{W}_1 \boldsymbol{z} + \boldsymbol{b}_1) + \ldots) + b_L)), \ \forall \boldsymbol{z} \in \mathbb{R}^d \},
$$

where $\boldsymbol{\sigma}(\cdot) = [\sigma(\cdot), \ldots, \sigma(\cdot)]^\top$ denote the activation function, L denotes the number of layers, $\boldsymbol{W}_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$, $d_0 = D$, $\boldsymbol{b}_\ell \in \mathbb{R}^{d_\ell}$, and $\boldsymbol{\theta}$ represents the parameters such that $\boldsymbol{\theta} = (\{\boldsymbol{W}_\ell, \boldsymbol{b}_\ell\}_{\ell=1}^{L-1}, \boldsymbol{w}_L, b_L)$. Then, the covering number of $\mathcal{G} \circ \boldsymbol{\Xi}$ with respect to the Frobenius norm satisfies

$$
\mathsf{N}(\mathcal{G} \circ \boldsymbol{\Xi}, \varepsilon) \leq \exp \left(\frac{\|\boldsymbol{Z}\|_{\mathrm{F}} \mathscr{R}_{\mathcal{G}}}{\varepsilon} \right)^{1/2},
\tag{76}
$$

where

$$
\mathcal{G} \circ \boldsymbol{\Xi} = \{ g_{\boldsymbol{\theta}}([\boldsymbol{Z}]_{\boldsymbol{\Xi}}) = [g(\boldsymbol{z}_{\boldsymbol{i}_1}), \ldots, g(\boldsymbol{z}_{\boldsymbol{i}_S})]^\top, \ g_{\boldsymbol{\theta}} \in \mathcal{G} \},
$$

$\boldsymbol{Z} = [\boldsymbol{z}_{\boldsymbol{i}_1}, \ldots, \boldsymbol{z}_{\boldsymbol{i}_S}] \in \mathbb{R}^{D \times S}$ *represents the training data and*

$$
\mathscr{R}_{\mathcal{G}} = \left(2 \prod_{\ell=1}^{L} \rho_\ell s_\ell \right) \left(\sum_{\ell=1}^{L} \frac{d_\ell^2 d_{\ell-1}^2 a_\ell}{s_\ell} \right) L^2,
$$

where $\|\boldsymbol{W}_\ell\|_{\mathrm{F}} \leq a_\ell$, $\|\boldsymbol{W}_\ell\|_2 \leq s_\ell$, the activation function σ_ℓ is ρ_ℓ-Lipschitz and $\sigma_\ell(0) = 0$.
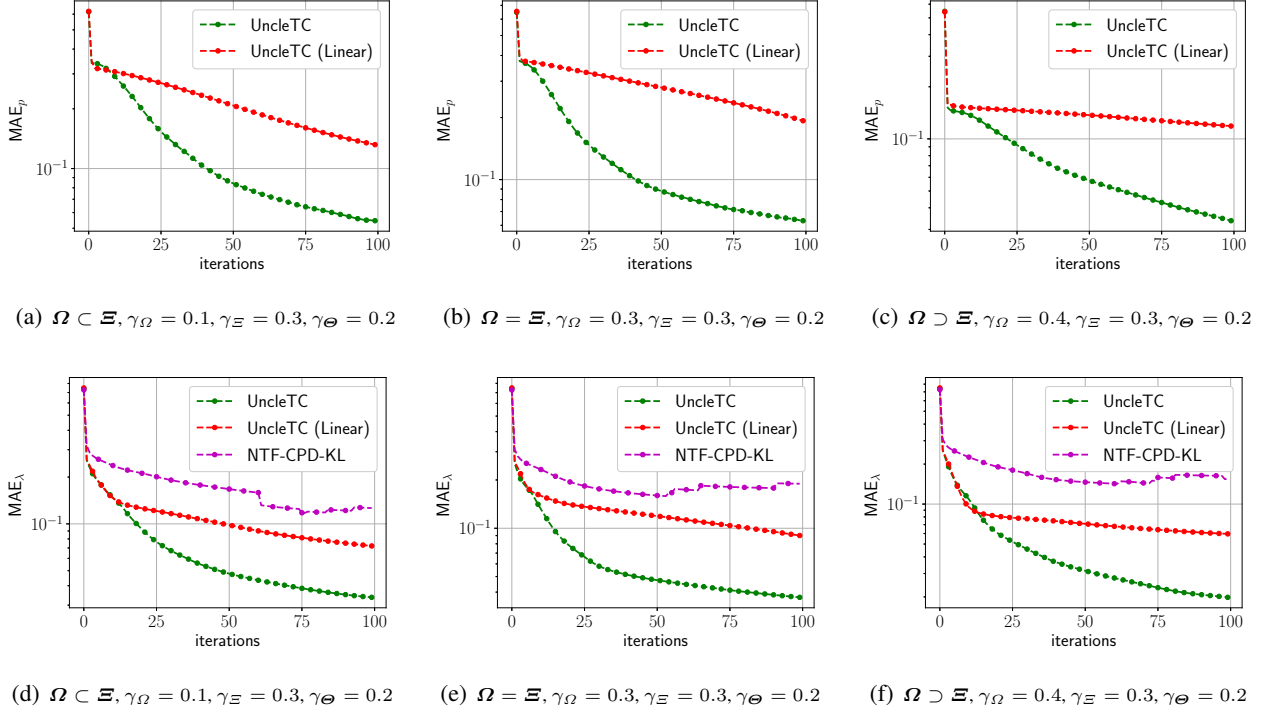
Figure 2. Average MAE_p (top) and MAE_λ (bottom) over 20 random trials plotted against iterations for different settings $\boldsymbol{\Omega} \subset \boldsymbol{\Xi}, \boldsymbol{\Omega} = \boldsymbol{\Xi}$, and $\boldsymbol{\Omega} \supset \boldsymbol{\Xi}$,. $K = 3, I_k = 20, \forall k, F = 3, D = 10, \text{SNR} = 40\text{dB}, g(\boldsymbol{z}) = \text{sigmoid}(\boldsymbol{\nu}^\top(3\boldsymbol{z}^3 + 0.2\boldsymbol{z}))$, where the vector $\boldsymbol{\nu} \in \mathbb{R}^D$ is generated by randomly sampling its entries a uniform distribution between 0 and 1.

N. More Details on Experiments

N.1. Synthetic Data Experiments - Implementation Settings.

To learn $g(\cdot)$, we use a neural network $g_{\boldsymbol{\theta}}(\cdot)$ with 3 hidden layers and 20 ReLU activation functions in each hidden layer. The optimizer for handling the subproblems of $\boldsymbol{\theta}$ is Adam (Kingma & Ba, 2015) with an initial learning rate of 10^{-3}. The optimizer uses a batch size of 1024. The subproblems are stopped when the relative change in the corresponding objective functions is smaller than 10^{-6}. Also, the algorithm is stopped if the relative change in the overall objective function is less than 10^{-6} or if 100 BCD iterations are completed. We fix the regularization term in the loss function to be $\ell(x, y) = (x - y)^2$. The regularization parameter μ is chosen to be a high value, i.e., $\mu = 2000$ across all experiments.

N.2. Additional Synthetic Data Experiments

In Fig. 2, we compare the performance of the proposed approach and two baselines by plotting MAE_p and MAE_λ over iterations, under different conditions such as $\boldsymbol{\Omega} \subset \boldsymbol{\Xi}, \boldsymbol{\Omega} \supset \boldsymbol{\Xi}$, and $\boldsymbol{\Omega} = \boldsymbol{\Xi}$. The baseline NTF-CPD-KL does not consider the detection probability in their model. Hence, we only plot its MAE_λ. It can be observed that UncleTC outperforms the baselines after around 10 iterations under all the settings under test.

Table 7 shows the results on a different synthetic dataset. Here, we consider another nonlinear function for the data generation, i.e.,

$$g(\boldsymbol{z}) = \text{sigmoid}\left(\boldsymbol{\nu}^\top(0.1\log(\boldsymbol{z}^2) + 0.1\boldsymbol{z}^2)\right),$$

where the vector $\boldsymbol{\nu} \in \mathbb{R}^D$ is generated by randomly sampling its entries from the uniform distribution between 0 and 1. We fix $\gamma_\Omega = 0.2, \gamma_\Xi = 0.3, \gamma_{\boldsymbol{\Theta}} = 0.2, \text{SNR} = 40\text{dB}$, and vary the rank F of the ground-truth tensor $\underline{\boldsymbol{\Lambda}}^\natural$. One can note that under these settings, the proposed algorithm still consistently gives better performance relative to the baselines.

Table 8 shows the results where wrong tensor ranks $\widehat{F}$'s were used in our algorithms. In practice, the underlying data generation process is unknown and hence, the true rank F is hard to know or estimate. Hence, it is of interest to understand

Table 7. Average MAE$_p$ and MAE$_\lambda$ over 20 random trials under different values of F. $K = 3, I_k = 20, \forall k, D = 10, \gamma_\Omega = 0.2, \gamma_\Xi = 0.3, \gamma_\Theta = 0.2$, SNR $= 40$dB, $g(\boldsymbol{z}) = \text{sigmoid}\left(\boldsymbol{\nu}^\top (0.1 \log(\boldsymbol{z}^2) + 0.1\boldsymbol{z}^2)\right)$.

Algorithm	Metric	$F = 5$	$F = 7$	$F = 10$
UncleTC	MAE$_p$	0.059 ± 0.001	0.049 ± 0.029	0.030 ± 0.002
	MAE$_\lambda$	0.058 ± 0.007	0.052 ± 0.021	0.074 ± 0.001
UncleTC (Linear)	MAE$_p$	0.148 ± 0.018	0.187 ± 0.026	0.077 ± 0.005
	MAE$_\lambda$	0.102 ± 0.024	0.202 ± 0.024	0.131 ± 0.014
NTF-CPD-KL	MAE$_\lambda$	0.166 ± 0.050	0.214 ± 0.085	0.120 ± 0.001

Table 8. Average MAE$_p$ and MAE$_\lambda$ over 20 random trials for different $\widehat{F}$ with true rank $F = 7$. $K = 3, I_k = 20, \forall k, D = 10, \gamma_\Omega = 0.2, \gamma_\Xi = 0.3, \gamma_\Theta = 0.2$, SNR $= 40$dB, $g(\boldsymbol{z}) = \text{sigmoid}\left(\boldsymbol{\nu}^\top (0.1 \log(\boldsymbol{z}^2) + 0.1\boldsymbol{z}^2)\right)$.

Algorithm	Metric	$\widehat{F} = 3$	$\widehat{F} = 5$	$\widehat{F} = 7$	$\widehat{F} = 10$	$\widehat{F} = 12$
UncleTC	MAE$_p$	0.052 ± 0.015	0.042 ± 0.014	0.046 ± 0.024	0.051 ± 0.018	0.046 ± 0.019
	MAE$_\lambda$	0.125 ± 0.008	0.084 ± 0.011	0.060 ± 0.018	0.075 ± 0.019	0.076 ± 0.019
UncleTC (Linear)	MAE$_p$	0.110 ± 0.041	0.108 ± 0.057	0.130 ± 0.045	0.112 ± 0.041	0.116 ± 0.046
	MAE$_\lambda$	0.141 ± 0.017	0.120 ± 0.039	0.140 ± 0.046	0.155 ± 0.064	0.197 ± 0.143
NTF-CPD-KL	MAE$_p$	0.151 ± 0.029	0.134 ± 0.046	0.159 ± 0.071	0.157 ± 0.072	0.151 ± 0.069

the algorithm's robustness to using a wrong $\widehat{F}$. One can see that the MAE values are worse when $\widehat{F} < F$ but become better when $\widehat{F} \geq F$. This makes sense since underestimating the rank means that less information of the true model is captured by the algorithms.

Table 9 shows the performance under various μ's, i.e., the regularization parameter of ℓ in UncleTC. The settings follow those of Table 7 with rank $F = 3$. The results indicate that using reasonably large μ ($\mu \geq 1000$) may be preferable, as we hope that the regularization term can enforce equality. In practice, one may also use a validation set to choose μ.

N.3. Real Data Experiments - Dataset Details.

COVID-19 Data. The dataset (Zhang et al., 2020b) includes the number of reported COVID-19 cases of about 2270 US counties during 462 days in 2020-2021. The dataset has 25 attributes, including social distancing index, percentage of people staying home, trips per person, percentage of out-of-county trips, percentage of out-of-state trips, transit mode share, miles per person, work trips per person, non-work trips per person, COVID exposure per 1000 people, percentage of people older than 60, median income, percentage of African Americans, percentage of Hispanic Americans, population density, number of hot spots per 1000 people, total population, percentage of people working from home, imported COVID cases, hospital beds per 1000 people, testing capacity gap, tests done per 1000 people, unemployment rate, cumulative inflation rate, and unemployment claims per 1000 people.

In order to extract a count-type tensor from the COVID-19 dataset, we adopt the following procedure: Each county is represented using two coordinates, i.e., the rounded value of latitude and longitude of its geographic center (we collected this information from a publicly available website[3]). The rounded value of latitude points ranges between 20 and 69, out of which 43 values correspond to the latitudes of at least one county. On the other hand, the discretized longitude points ranges between -165 and -69, out of which 80 discrete points correspond to the longitudes of at least one county. The third coordinate of the tensor is the dates on which the COVID-19 cases were reported. Hence, each entry of the tensor represents the number of reported COVID-19 cases associated with a latitude point and a longitude point on a particular day. We use the data corresponding to 60 days from January 1, 2021 to March 1, 2021. We form a $43 \times 80 \times 60$ tensor, which has 19.73% of observed entries, out of which 15.43% are nonzeros entries and 4.30% are zeros.

Pollinator-Plant Interaction (PPI) Data. The Plant-Pollinator Interaction (PPI) dataset is a publicly available dataset, collected by the researchers at the H.J. Andrews (HJA) Long-term Ecological Research site in Oregon, USA (Seo et al., 2022; Daly et al., 2019). The dataset comprises plant-pollinator interactions observed at about 12 meadows over a period of seven years. Each meadow was visited 5-7 times per year by student researchers. The dataset contains interactions between 69 plant species and 215 pollinator species observed during 37 visits. Only 1.07% of the entries are recorded. The counts in

---

[3]https://public.opendatasoft.com/explore/dataset/us-county-boundaries

Table 9. Average MAE$_p$ and MAE$_\lambda$ for `UncleTC` over 20 random trials under different μ values. We fix $K = 3, F = 3, D = 10, \gamma_\Omega = 0.2, \gamma_\Xi = 0.3, \gamma_\Theta = 0.2, \mathrm{SNR} = 40\mathrm{dB}, g(\boldsymbol{z}) = \mathrm{sigmoid}\left(\boldsymbol{\nu}^\top(0.1\log(\boldsymbol{z}^2) + 0.1\boldsymbol{z}^2)\right)$

| Settings	Metric	$\mu = 100$	$\mu = 500$	$\mu = 1000$	$\mu = 2000$
$I_k = 20, \gamma_\Omega = 0.3, \gamma_\Xi = 0.3$	MAE$_p$	0.184 ± 0.025	0.074 ± 0.024	0.052 ± 0.011	0.042 ± 0.004
	MAE$_\lambda$	0.171 ± 0.031	0.067 ± 0.029	0.042 ± 0.014	0.028 ± 0.004
$I_k = 25, \gamma_\Omega = 0.3, \gamma_\Xi = 0.3$	MAE$_p$	0.198 ± 0.054	0.083 ± 0.030	0.056 ± 0.015	0.054 ± 0.013
	MAE$_\lambda$	0.184 ± 0.049	0.069 ± 0.029	0.039 ± 0.013	0.033 ± 0.010
$I_k = 20, \gamma_\Omega = 0.1, \gamma_\Xi = 0.3$	MAE$_p$	0.238 ± 0.069	0.151 ± 0.085	0.108 ± 0.074	0.080 ± 0.044
	MAE$_\lambda$	0.256 ± 0.087	0.166 ± 0.097	0.113 ± 0.084	0.078 ± 0.047
$I_k = 25, \gamma_\Omega = 0.1, \gamma_\Xi = 0.3$	MAE$_p$	0.260 ± 0.038	0.156 ± 0.073	0.141 ± 0.079	0.121 ± 0.079
	MAE$_\lambda$	0.305 ± 0.045	0.203 ± 0.111	0.194 ± 0.126	0.201 ± 0.134
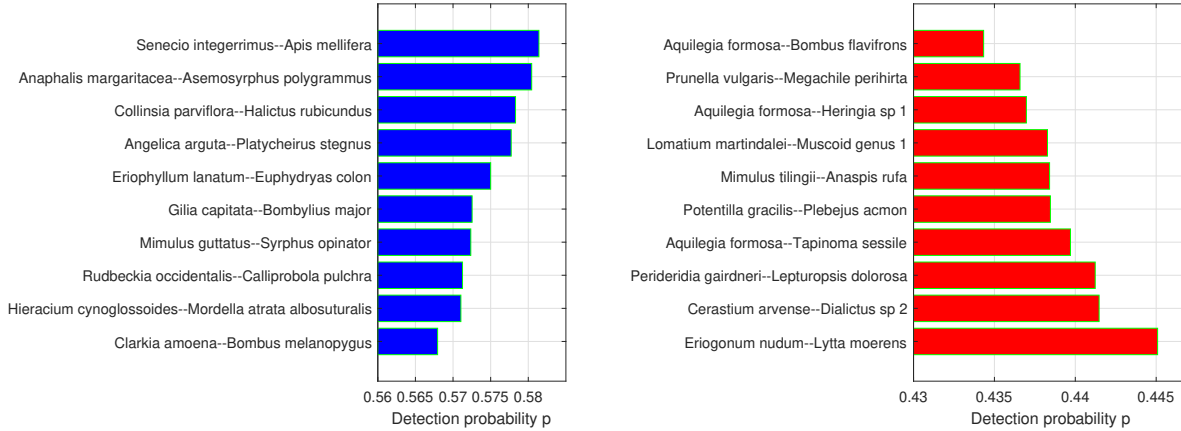
Figure 3. The top 10 plant-pollinator interactions having highest average detection probabilities (left) and lowest average average detection probabilities (right) as identified by the proposed `UncleTC`.

this dataset are widely believed to be under-counted (Fu et al., 2021; Seo & Hutchinson, 2018). We consider a subset of the dataset by selecting 50 plants and 50 pollinators that exhibit the highest number of interactions. This leads to a $50 \times 50 \times 37$ integer tensor with 22.06% observed entries, out of which 4.01% are nonzero entries and 18.05% are zero entries. The zero entries indicate that the plant and pollinator were both observed by the researcher during the visit, but they were not observed to be interacting with one another.

The dataset has 37 attributes:

• 12 plant species static traits: the estimate of reward per flower, the basic ecological life form of a plant, floral structure, peduncle feebleness, the level of exclusion, exclusion by flower color, exclusion by pendant position, time of diel availability, size of pollen grain, the ability of the inflorescence to support the arrival of the flower visitor, the possible affinity of plant species to edaphic characteristics, and dispersal mechanism.

• 11 pollinator species static traits: the length of a species body size, the width of a species body size, the depth of a species body size, individual biomass, size, energy requirement, a species behavior characterized by active time, the level of exclusion, an indicator showing whether the platform is required to approach flowers, tongue length, and tube size.

• 1 plant temporal feature: flower abundance

• 4 temporal features: month, day, wind, and cloud

• 7 temperature features: mean, max, min, max time, min time, cumulative degree days with 5 degrees, and cumulative degree days with 10 degrees.

• 2 precipitation features: daily precipitation, and antecedent with $k = 0.9$
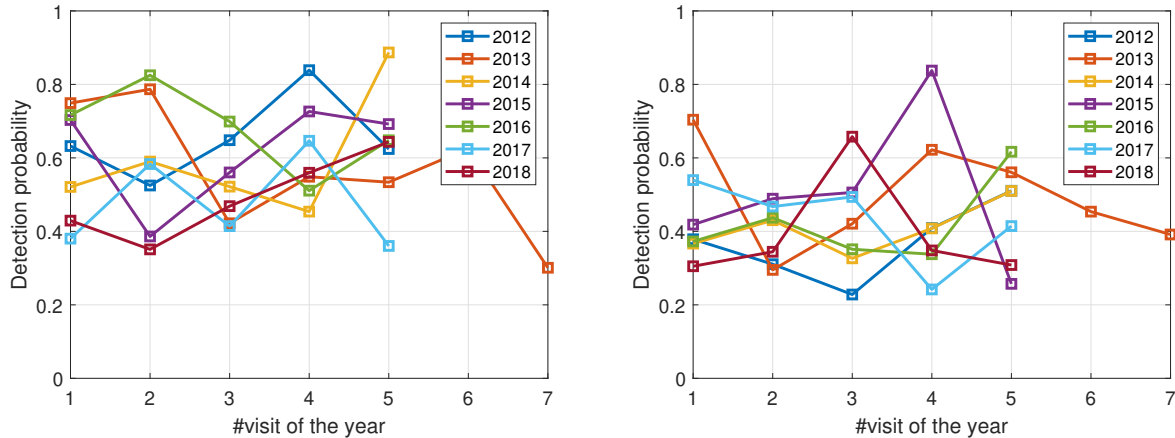
31

Figure 4. (left) The detection probabilities of the interaction pair *Senecio integerrimus–Apis mellifera*, which is reported by our algorithm to have the highest average detection probability. (right) The detection probabilities of the pair *Aquilegia formosa–Bombus flavifrons*, which is reported by our algorithm to have the lowest average detection probability.

N.4. Real Data Experiments - Implementation Settings.

For the proposed `UncleTC`, we fix `ReLU` as the neural network activation function and `Adam` as the the optimizer. Other parameters such as rank F, the regularization parameter μ, batch size, learning rates, and neural network size are selected via the grid search hyper-parameter tuning strategy on the validation set where the `rRMSE` is used to measure the validation score. For the proposed `UncleTC`, we choose the rank F of the tensor from $[5, 10, 15, 20]$, the regularization parameter μ from $[500, 1000, 2000, 3000]$, the number of hidden units from $[10, 20, 30]$, the number of hidden layers from $[1, 3, 5]$, the learning rate from $[0.001, 0.005, 0.01]$ and the batch size from $[256, 512, 1024]$. For all the low-rank tensor decomposition-based baselines, we select the rank from $[5, 10, 15, 20]$. In addition, the sparsity parameter of the `BPTF-CPD` (Schein et al., 2015) is selected from $[0.02, 0.05, 0.1, 0.2]$ and the regularization parameter of the `HaLRTC` (Liu et al., 2013) is selected from $[10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}]$, based on the recommended range of values from the respective papers. The number of iterations to stop training all the iterative algorithms is also chosen using the validation sets.

In the experiments, we perform 5-fold cross-validation to characterize the prediction performance with k chosen to be 5, where 3 splits serve for training, and the other two for validation and testing, respectively. The roles of the sets are switched in a cyclical way in different trials.

N.5. Real Data Experiments - Additional Results.

Fig. 3 displays the top 10 interactions with the highest and lowest average detection probabilities, as discovered by our proposed `UncleTC` method. Generally, interactions involving larger, more common species are more likely to have higher detection probabilities, while smaller, rarer species tend to have lower detection probabilities. The species with the highest average detection probabilities have a larger number of observed interactions compared to those with the lowest average detection probabilities. However, it is worth noting that not all species follow this trend. Other factors such as species rareness, environmental conditions, or seasonal variability may contribute to these differences (Dorado et al., 2011; Chacoff et al., 2012).

In Fig. 4, we present the periodic variation of detection probabilities for specific species interactions over the years. Both figures illustrate temporal variation across visits and years for two different interactions—one with the highest average detection probability and the other with the lowest. Despite the variations, the overall range remains relatively consistent, with probabilities ranging from approximately 0.4 to 0.8 in the left figure and from 0.2 to 0.6 in the right figure, ultimately leading to the highest and lowest average detection probabilities.

Considering the exceptions where detection probability does not align with species activity levels, we further examined interactions that show the most and least significant variations in detection probabilities over time. Fig. 5 plots the detection probabilities over visits for interactions with the highest (left) and lowest (right) standard deviation of detection probabilities
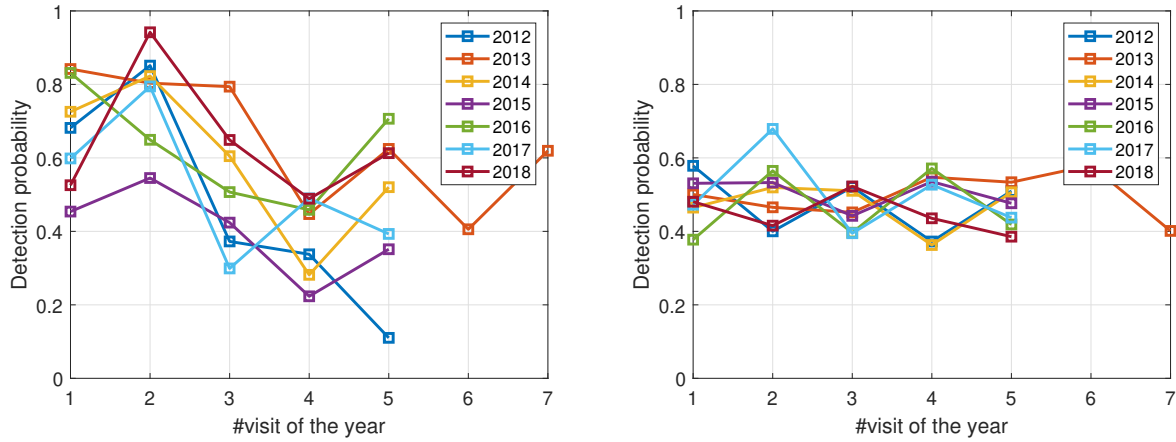
Figure 5. (left) The detection probability of the interaction *Gilia capitata–Apis mellifera* over different years, which is reported by our algorithm to have the highest standard deviation of the detection probability across 37 visits. (right) The detection probability pattern for the interaction *Eriophyllum lanatum–Bombus mixtus*, which is reported by our method to have the lowest standard deviation of the detection probability across 37 visits.

across visits and annually. The left figure confirms that the detection probability of the interaction between two well-known common species (*Gilia capitata* and *Apis mellifera*) is significantly influenced by the time of year, resulting in a relatively lower average detection probability. The right figure provides another example of an interaction between another plant and pollinator species (*Eriophyllum lanatum* and *Bombus mixtus*), with relatively lower detection probabilities over time. Such variation may be a product of the phenology of the species (i.e., when flowers are blooming and when insects are flying).