The Bethe and Sinkhorn Permanents of Low Rank Matrices and Implications for Profile Maximum Likelihood

Nima Anari ANARI@CS.STANFORD.EDU
Stanford University

Moses Charikar MOSES@CS.STANFORD.EDU
Stanford University

Kirankumar Shiragur SHIRAGUR@STANFORD.EDU
Stanford University

Aaron Sidford SIDFORD@STANFORD.EDU
Stanford University

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

In this paper we consider the problem of computing the likelihood of the profile of a discrete distribution, i.e., the probability of observing the multiset of element frequencies, and computing a profile maximum likelihood (PML) distribution, i.e., a distribution with the maximum profile likelihood. For each problem we provide polynomial time algorithms that given n i.i.d. samples from a discrete distribution, achieve an approximation factor of $\exp\left(-O(\sqrt{n}\log n)\right)$, improving upon the previous best-known bound achievable in polynomial time of $\exp(-O(n^{2/3}\log n))$ (Charikar, Shiragur and Sidford, 2019). Through the work of Acharya, Das, Orlitsky and Suresh (2016), this implies a polynomial time universal estimator for symmetric properties of discrete distributions in a broader range of error parameter.

To obtain our results on PML we establish new connections between PML and the well-studied Bethe and Sinkhorn approximations to the permanent (Vontobel, 2012 and 2014). It is known that the PML objective is proportional to the permanent of a certain Vandermonde matrix (Vontobel, 2012) with $\sqrt{n}$ distinct columns, i.e. with non-negative rank at most $\sqrt{n}$. This allows us to show that the convex approximation to computing PML distributions studied in (Charikar, Shiragur and Sidford, 2019) is governed, in part, by the quality of Sinkhorn approximations to the permanent. We show that both Bethe and Sinkhorn permanents are $\exp(O(k\log(N/k)))$ approximations to the permanent of $N \times N$ matrices with non-negative rank at most k. This improves upon the previous known bounds of $\exp(O(N))$ and combining these insights with careful rounding of the convex relaxation yields our results.

Keywords: symmetric property estimation, profile maximum likelihood, permanent approximation

1. Introduction

Symmetric property estimation is an important and well-studied problem in statistics and theoretical computer science. Given access to n i.i.d samples from a hidden discrete distribution $\mathbf{p}$, the goal is to estimate $\mathbf{f}(\mathbf{p})$, for a symmetric property $\mathbf{f}(\cdot)$. Formally, a property is symmetric if it is invariant to permutating the labels, i.e. it is a function of the multiset of probabilities and does not depend on the symbol labels. There are many well-known well-studied such properties, including support size and coverage, entropy, distance to uniformity, Renyi entropy, and sorted ℓ_1 distance. Understanding the computational and sample complexity for estimating these symmetric properties has led to an extensive line of interesting research over the past decade.

Symmetric property estimation spans applications in many different fields. For instance, entropy estimation has found applications in neuroscience Rieke et al. (1999), physics Vinck et al. (2012) and others Plotkin and Wyner (1996); Porta et al. (2001). Support size and coverage estimation were initially used in estimating ecological diversity Chao (1984); Chao and Lee (1992); Bunge and Fitzpatrick (1993); Colwell et al. (2012) and subsequently applied to many different applications Efron and Thisted (1976); Thisted and Efron (1987); Fürnkranz (2005); Kroes et al. (1999); Paster et al. (2001); Daley and Smith (2013); Robins et al. (2009); Gao et al. (2007); Hughes et al. (2001). For applications of other symmetric properties we refer the reader to Han et al. (2017b,a); Acharya et al. (2014); Raghunathan et al. (2017); Zou et al. (2016); Wu and Yang (2016); Raskhodnikova et al. (2007); Wu and Yang (2015); Orlitsky et al. (2016); Valiant and Valiant (2011b); Wu and Yang (2016); Jiao et al. (2015, 2016); Valiant and Valiant (2011a).

Universal estimators: Early work on symmetric property estimation developed estimators tailored to the particular property of interest. Consequently, a fundamental and important open question was to come up with an estimator that is *universal*, i.e. the same estimator could be used for all symmetric properties. A natural approach for constructing universal estimators is a plug-in approach, where given samples we first compute a distribution independent of the property and later we output the (value of this) property for the computed distribution as our estimate.

Recently, Acharya et al. (2016) provided an approach for constructing universal plug-in estimators. This approach leveraged the observation that a sufficient statistic for estimating a symmetric property from a sequence of samples is the profile, i.e. the multiset of frequencies of symbols in the sequence, e.g. the profile of sequence $abbc$ is $\{2, 1, 1\}$. In their approach, they used the *profile maximum likelihood (PML)* distribution introduced by Orlitsky et al. (2004) as a plug-in distribution: given a sequence of n samples, PML is the distribution that maximizes the likelihood of the observed profile. The authors in Acharya et al. (2016) showed that a plug-in estimator using a optimal PML distribution is universal in estimating various symmetric properties of distributions. In fact it suffices to compute a β-approximate PML distribution (i.e. a distribution that approximates the PML objective to within a factor of β) for $\beta > \exp(-n^{1-\delta})$ for constant $\delta > 0$. The parameter β in β-approximate PML effects the error parameter regime under which the estimator is sample complexity optimal. Larger values of β yield a universal estimator that is sample optimal over broader parameter regime.

Previous work of the authors in Charikar et al. (2019a), gave the first polynomial time algorithm to compute a β-approximate PML for some non-trivial β. In particular, Charikar et al. (2019a) gave a nearly linear running time algorithm to compute an $\exp(-O(n^{2/3} \log n))$-approximate PML distribution. Leveraging Acharya et al. (2016), this yields a universal estimator that is sample complexity optimal for estimating certain symmetric properties within accuracy for $\epsilon > n^{-0.16666}$, where ϵ is the desired accuracy of the estimation.

Motivating questions: Given the possible utility of universal estimators based on PML and the recent progress of Charikar et al. (2019a), a key motivating question for this paper is:

What is the smallest β for which we can compute β-approximate distributions in polynomial time?

Recently, Charikar et al. (2019a) showed that $\exp(-O(n^{2/3} \log n))$-approximate PML distributions can be computed in polynomial time and in this paper, we seek to improve this approximation quality. Acharya et al. (2016) implies that such improvements yield efficient universal estimators for certain symmetric properties in broader accuracy regime.

Beyond seeking improved approximation guarantees for PML, this paper seek to better understand the algorithmic machinery which underlies computing approximate PML. The algorithm and analysis in Charikar et al. (2019a) were somewhat specialized and it is unclear how this work relates to the design of approximation algorithms more broadly.

Towards this goal, we note that computing approximate PML corresponds to solving a non-convex optimization problem, where the objective function is a permanent of generalized Vandermonde matrix. There is rich literature Yedidia et al. (2005); Vontobel (2013); Linial et al. (1998); Schrijver (1998); Gurvits (2011); Vontobel (2013); Gurvits and Samorodnitsky (2014); Schrijver (1978); Alon and Spencer (2004) on approximating the permanent of a fixed non-negative matrix using continuous optimization problems. For example, the Bethe and Sinkhorn permanent approximations Gurvits and Samorodnitsky (2014); Grier and Schaeffer (2018); Vontobel (2014); Yedidia et al. (2005); Vontobel (2012) are few such optimization problems that were originally studied in statistical physics as popular tools for providing deterministic approximations to the permanent of non-negative matrices. Although these prior works provide a way to approximate the permanent of a fixed matrix, approximately maximizing the permanent over the entries of the matrices remains largely unknown. Nevertheless, we ask:

Can we draw connections between permanent approximation and approximate PML computation?

Unfortunately, establishing such a connection faces an immediate barrier. The current analysis of the Sinkhorn and Bethe permanents show that the ratio between the permanent and these approximations is upper bounded by c^N Gurvits and Samorodnitsky (2014); Anari and Rezaei (2018) for some constant $c > 0$, where N is the dimension of the matrix. However, in most of our setting we seek to compute $e^{O(N^{1-\delta})}$ approximate PML distribution for constant $\delta > 0$ and it therefore seems that such results may be insufficient for our purposes.

Towards overcoming this issue, we note that these previous results on the Bethe and Sinkhorn permanent, do not exploit many structural properties of the matrix being approximated. For example, the PML matrix has low non-negative rank. Consequently, towards enabling Bethe and Sinkhorn permanent approximations to yield non-trivial guarantees for PML we ask:

Can we efficiently provide efficient deterministic approximations to the permanent of a broad class of structured matrices, where the approximation factor depends on the structural parameter?

We believe this question is interesting on its own.

Our contribution: In this paper we make progress on addressing the preceding motivating questions. Our main results are two fold: (1) we show that it is possible to compute a $\exp(-O(\sqrt{n}\log n))$-approximate PML distribution in polynomial time through a continuous optimization problem closely related to Sinkhorn approximations to the permanent (in fact this is the same as the one in Charikar et al. (2019a)) and (2) we provide new bounds on the quality of Bethe and Sinkhorn approximations to the permanent in the case of low rank matrices (the special case of this to matrices with a bounded number of distinct columns we leverage to prove (1)).

Our first result has two conceptual steps: First, we use the idea of probability discretization and apply it to the Sinkhorn permanent to study a different convex optimization problem. We show that this new convex program approximates the probability of a given profile with respect to a fixed distribution $\mathbf{p}$; the approximation factor is upper bounded by $\exp(-O(k\log n))$, where n is the number of samples and k is the number of distinct frequencies in the profile. Given n samples as the number of distinct frequencies k is always less than $\sqrt{n}$, we immediately get a deterministic $\exp\left(-O(\sqrt{n}\log n)\right)$ approximation to the probability of a profile.

In the second step, we study a variant of the convex program from the first part to encode the problem of maximizing over all distributions. This new optimization problem is also convex and its feasible solutions represent fractional distributions[1]. To return a valid approximate PML distribution with the desired guarantee, we provide a new efficient rounding algorithm that rounds the fractional distributions while incurring a loss of $\exp(O(\sqrt{n}\log n))$ in the objective.

Recall that the previous best known result for efficiently computing approximate PML distribution is due to Charikar et al. (2019a), where the authors using combinatorial techniques provide a convex optimization problem that helps them compute an $\exp(-O(n^{2/3}\log n))$-approximate PML distribution in nearly linear time. The convex program provided in Charikar et al. (2019a) is the same as ours, which is quite surprising as the prior derivation of this relaxation in Charikar et al. (2019a) was purely combinatorial and was not directly derived from Sinkhorn approximation. We also remark that in a follow up work to ours Anari et al. (2020), using the connection we established between the Sinkhorn and PML from the first part, and several other key steps in our rounding algorithm from the second part, the authors in Anari et al. (2020) provided an improved efficient rounding algorithm that returns an instance based $\exp(-O(k\log n))$ approximate PML distribution. Anari et al. (2020) further used the instance based approximation to efficiently implement the PseudoPML Charikar et al. (2019b) and profile entropy results Hao and Orlitsky (2020) (See Section 4.1 for further details).

Leveraging the result from Acharya et al. (2016), such an improved $\exp(-O(\sqrt{n}\log n))$ approximate PML distribution provides us an universal estimator that is sample optimal in the regime $\epsilon > n^{-0.249}$, while the previous work Charikar et al. (2019a) provided analogous result for $\epsilon > n^{-0.166}$.

We now describe our second result (a crucial ingredient used in the improved PML approximation described above): we show that the approximation ratio between the permanent and the scaled Sinkhorn permanent is upper bounded by $\exp(-O(k\log(N/k)))$, where N, k are the dimension and the non-negative rank of the matrix respectively. This result implies the same approximation guarantee for the Bethe permanent, an alternative to the Sinkhorn permanent with a tighter worst-case multiplicative approximation. We also give an explicit construction of a matrix to show that our result for this structural parameter is asymptotically tight. As described earlier, the main application of the improved upper bound for the Sinkhorn and Bethe permanents are in PML. Recall that the scaled Sinkhorn permanent was used to provide the convex programs studied in our first result and the analysis of the scaled Sinkhorn was the key ingredient to prove the desired approximation guarantees for the PML.

Organization of the paper: In Section 2, we provide a overview of our techniques. In Section 3 we present preliminaries. In Section 4, we provide the main results of the paper. In Section 5, we demonstrate the process of computing approximate PML through an illustrative example. In Appendix A, we analyze the scaled Sinkhorn permanent of structured matrices. In Appendix A.2, we prove an upper bound for the approximation ratio of the scaled Sinkhorn permanent to the permanent as a function of the number of distinct columns. In Appendix A.3, we prove the generalized result of the scaled Sinkhorn permanent for the low non-negative rank matrices. In Appendix B, we prove the lower bound for the Bethe and scaled Sinkhorn approximations of the permanent. In Appendix C,

[1] The variable of this new optimization problem is a matrix whose rows correspond to probability values and the ith row sum denotes the number of domain elements in the distribution with that probability value.

we combine the result for the scaled Sinkhorn permanent with the idea of probability discretization to provide the convex program that returns a fractional representation of an approximate PML distribution. In the same section, we provide the rounding algorithm to return a valid approximate PML distribution.

2. Overview of Techniques

Here we provide a broad overview of our approach to compute an approximate PML distribution. In Section 2.1, we outline the key ideas in obtaining an efficient algorithm to compute an $\exp(-O(\sqrt{n}\log n))$ approximate PML. A crucial ingredient in this result is a bound on the quality of the Sinkhorn permanent approximation for low non-negative rank matrices. We give an overview of this proof in Section 2.2.

2.1. Efficient computation of approximate PML distribution

Here we provide a proof overview of our primary result, where we draw a connection between the previous known permanent approximations and PML to provide an efficient algorithm to compute $\exp(-O(\sqrt{n}\log n))$ approximate PML distributions. Our approach leverages that we can obtain improved bounds on the approximation ratio of the Bethe and scaled Sinkhorn permanent approximations of low non-negative matrices, which we discuss in greater detail in Section 2.2.

The idea of using these permanent approximations for computing an approximate PML distribution comes from the fact that the likelihood of a profile with respect to a distribution can be written as the permanent of a non-negative Vandermonde matrix (which we call the profile probability matrix) Vontobel (2012). The number of distinct rows and columns of this profile probability matrix correspond to the number of distinct frequencies in the profile and distinct probability values in the distribution respectively.

As the non-negative rank of a matrix is always upper bounded by the minimum of the number of its distinct rows and columns, through our analysis (outlined in the next section) we get that the Bethe and scaled Sinkhorn permanents are within a factor $\exp\left(-O(k\log n)\right)$ of the PML objective with respect to a fixed distribution, where k is the number of distinct frequencies in the profile. Given n samples, as the number of distinct frequencies is always upper bounded by $\sqrt{n}$, our analysis of the scaled Sinkhorn permanent immediately implies an $\exp\left(-O(\sqrt{n}\log n)\right)$ approximation to the PML objective with respect to a fixed distribution.

Even with this improved bound on the quality of the Bethe and scaled Sinkhorn approximations as applied to the PML objective, challenges remain in obtaining an improved approximate PML distribution. In particular, we do not know of an efficient algorithm to maximize the Bethe or the scaled Sinkhorn permanent of the profile probability matrix over a family of distributions as it would be needed to compute the Bethe or the scaled Sinkhorn approximation to the optimum of the PML objective. Prior work by Vontobel suggests an alternating maximization approach, but this is only guaranteed to produce a local optimum. To address this, we apply the idea of probability discretization to rewrite the scaled Sinkhorn optimization problem. Our new optimization problem is convex and its variables form a matrix. The rows of this variable matrix are indexed by a fixed set of probability values and the columns are indexed by the distinct frequencies. Further the row and column sums of this matrix are equal to the number of domain elements with their corresponding probability values and frequencies respectively. One nice property of this new optimization problem is that a slight modification to it helps us encode the part of maximizing over all the distributions. The modification results in a new convex program whose optimal solution is a fractional representation of an approximate PML distribution. Surprisingly, the resulting convex program is exactly the same as the one in Charikar et al. (2019a), where a completely different (combinatorial) technique was used to arrive at the convex program.

The final challenge towards obtaining our PML results is to round the fractional solution produced so that the approximation guarantee is preserved. The rounding procedure from Charikar et al. (2019a) does not immediately suffice, but we present a more sophisticated and delicate rounding procedure that does indeed give us the required approximation guarantee. A main task of the algorithm is to round the fractional solution matrix such that all row sums are integral while preserving the column constraints. Our rounding algorithm proceeds in three steps, where in the first step we first apply a procedure analogous to Charikar et al. (2019a) to handle large probability values and in the later steps we provide a new procedure to the smaller probability values; in each step, we ensure that the objective function does not drop significantly. We create rows corresponding to new probability values in the course of the rounding algorithm, maintain column sums and eventually ensure that all row sums are integral, and ensure that the objective function has not dropped significantly. We provide a more detailed proof overview of the rounding algorithm in Appendix C.2.

2.2. Permanent approximations of low non-negative-rank matrices

Here we provide a proof overview of one of our main technical results where we show that the approximation ratio between the permanent and the Bethe and scaled Sinkhorn permanent approximations are upper bounded by an exponential in the non-negative rank of the matrix (up to a logarithmic factor). The Bethe and scaled Sinkhorn permanents of a non-negative matrix A are optimum solutions to maximization problems over doubly stochastic matrices Q where the objective functions have entropy-like terms involving the entries of A and Q. Our analysis here exploits the non-trivial fact that the Bethe and scaled Sinkhorn approximations are lower bounds for the permanent of a non-negative matrix. In order to obtain an upper bound on the Bethe and scaled Sinkhorn approximation as a function of the non-negative rank, we show the existence of a doubly stochastic matrix Q as a witness such that the objective of the Bethe and scaled Sinkhorn w.r.t. Q upper bounds the permanent of A within the desired factor.

We first work with a simpler setting of matrices A with at most k distinct columns.[2] Here we consider a modified matrix $\hat{A}$ that contains the k distinct columns of A. We define a distribution μ on permutations of the domain where the probability of a permutation σ is proportional to its contribution to the permanent of A. There is a many-to-one mapping from such permutations to 0-1, $N \times k$ matrices with row sums 1 and column sums ϕ_j, the number of times the j'th column of $\hat{A}$ appears in A. We next define an $N \times k$ real-valued, non-negative matrix P with row sums 1 and column sums ϕ_j, in terms of the marginals of the distribution μ. We also define a different distribution ν on 0-1, $N \times k$ row-stochastic matrices by independent sampling from P. Finally, we use the fact that the KL-divergence between μ and ν is non-negative to get the required upper bound on the scaled Sinkhorn approximation with a doubly stochastic witness Q (obtained from P). This proof technique is inspired by the recent work of Anari and Rezaei Anari and Rezaei (2018) that

[2] In response to an initial submission of this paper, an anonymous reviewer showed that a simpler proof for the distinct column case can be derived using Corollary 3.4.5 of Barvinok's book Barvinok (2017). We thank the anonymous reviewer for this and include the derivation in Appendix D. The proof of the Corollary 3.4.5 further uses the famous Bregman–Minc inequality, a relatively heavy hammer. In contrast, our proof is self-contained and we believe it provides further insight into the structure of the Sinkhorn/Bethe approximations. See Section 4.1 for further details.

gives a tight $\sqrt{2}^N$ bound on the approximation ratio of the Bethe approximation for the permanent of an $N \times N$ non-negative matrix.

Both our work and Anari and Rezaei (2018) use entropy based methods but they differ at key places. In most of the prior entropy based methods Schrijver (1978); Radhakrishnan (1997); Anari et al. (2018), μ is the distribution of interest and is straightforward to construct. On the other hand, the distribution ν differs across various approaches and is the crucial part of the analysis. For instance, we exploit the structure of repetitive columns and work with a marginal distribution ν defined over $0-1$, $N \times k$ matrices while the proof of Anari and Rezaei (2018) analyzes a distribution defined over permutations. The idea of working with marginal distributions reduces the dimension of the problem and helps us derive bounds in terms of k instead of N. Another key difference between our work and Anari and Rezaei (2018) is in the procedure to derive distribution ν. While Anari and Rezaei (2018) used dependent sampling procedure to define the distribution ν, we use a simple independent sampling procedure.

Though this bound on the quality of the Bethe and scaled Sinkhorn approximations for non-negative matrices with k distinct columns suffices for our PML applications, interestingly we show that it can be extended to non-negative matrices with bounded rank. In order to obtain an upper bound on the Bethe and scaled Sinkhorn approximation as a function of the non-negative rank of A, recall that we need to show the existence of a suitable doubly stochastic witness Q which certifies the required guarantee. We express the permanent of A as the sum of $O(\exp(k\log(N/k)))$ terms of the form $\mathrm{perm}(U)\mathrm{perm}(V)$ where matrices U and V have at most k distinct columns. We focus on the largest of these terms, and construct a doubly stochastic witness Q for matrix A from the witnesses for matrices U and V in this largest term. This doubly stochastic witness Q certifies the required guarantee and we get an upper bound on the scaled Sinkhorn approximation as a function of the non-negative rank. This result for the scaled Sinkhorn approximation further implies an upper bound for the Bethe approximation.

3. Preliminaries

Let $[a,b]$ and $[a,b]_{\mathbb{R}}$ denote the interval of integers and reals $\geq a$ and $\leq b$ respectively. Let $\mathcal{D}$ be the domain of elements and $N \stackrel{\mathrm{def}}{=} |\mathcal{D}|$ be its size. Let $\mathbf{A} \in \mathbb{R}^{\mathcal{D}\times\mathcal{D}}$ be a non-negative matrix, where its (x,y)'th entry is denoted by $\mathbf{A}_{x,y}$. We further use $\mathbf{A}_{x:}$ and $\mathbf{A}_{:y}$ to denote the row and column corresponding to x and y respectively. The non-negative rank of a non-negative matrix $\mathbf{A} \in \mathbb{R}^{\mathcal{D}\times\mathcal{D}}$ is equal to the smallest number k such there exist non-negative vectors $\mathbf{v}_j, \mathbf{u}_j \in \mathbb{R}^{\mathcal{D}}$ for $j \in [1,k]$ such that $\mathbf{A} = \sum_{j\in[1,k]} \mathbf{v}_j \mathbf{u}_j^{\top}$. Let $S_{\mathcal{D}}$ be the set of all permutations of domain $\mathcal{D}$ and we denote a permutation σ in the following way $\sigma = \{(x,\sigma(x)) \text{ for all } x \in \mathcal{D}\}$. The permanent of a matrix $\mathbf{A}$ denoted by $\mathrm{perm}(\mathbf{A})$ is defined as: $\mathrm{perm}(\mathbf{A}) \stackrel{\mathrm{def}}{=} \sum_{\sigma\in S_{\mathcal{D}}} \prod_{e\in\sigma} \mathbf{A}_e$. Let $\mathbf{K}_{rc} \subseteq \mathbb{R}_{\geq 0}^{\mathcal{D}\times\mathcal{D}}$ be the set of all non-negative matrices that are doubly stochastic. For any matrix $\mathbf{A} \in \mathbb{R}_{\geq 0}^{\mathcal{D}\times\mathcal{D}}$ and $\mathbf{Q} \in \mathbf{K}_{rc}$, we define the following set of functions:

$$\mathrm{U}(\mathbf{A},\mathbf{Q}) \stackrel{\mathrm{def}}{=} \sum_{(x,y)\in\mathcal{D}\times\mathcal{D}} \mathbf{Q}_{x,y} \log\left(\frac{\mathbf{A}_{x,y}}{\mathbf{Q}_{x,y}}\right) \quad \text{and} \quad \mathrm{V}(\mathbf{Q}) = \sum_{(x,y)\in\mathcal{D}\times\mathcal{D}} (1-\mathbf{Q}_{x,y})\log\left(1-\mathbf{Q}_{x,y}\right) . \tag{1}$$

For a matrix $\mathbf{A} \in \mathbb{R}_{\geq 0}^{\mathcal{D}\times\mathcal{D}}$, the Bethe and Sinkhorn permanent of $\mathbf{A}$ are defined as follows,

$$\mathrm{bethe}(\mathbf{A}) \stackrel{\mathrm{def}}{=} \max_{\mathbf{Q}\in\mathbf{K}_{rc}} \exp\left(\mathrm{U}(\mathbf{A},\mathbf{Q}) + \mathrm{V}(\mathbf{Q})\right) \quad \mathrm{sinkhorn}(\mathbf{A}) \stackrel{\mathrm{def}}{=} \max_{\mathbf{Q}\in\mathbf{K}_{rc}} \exp\left(\mathrm{U}(\mathbf{A},\mathbf{Q})\right) .$$

Later we will see that it is convenient to work with $\exp(-N)\mathrm{sinkhorn}(\mathbf{A})$ than $\mathrm{sinkhorn}(\mathbf{A})$ itself; we define this expression to be scaled Sinkhorn and we formally define it as follows.

$$\mathrm{scaledsinkhorn}(\mathbf{A}) \stackrel{\mathrm{def}}{=} \max_{\mathbf{Q}\in\mathbf{K}_{rc}} \exp\left(\mathrm{U}(\mathbf{A},\mathbf{Q}) - N\right) \ .$$

The maximization objectives in the definitions of $\mathrm{bethe}(\mathbf{A})$, $\mathrm{sinkhorn}(\mathbf{A})$ and $\mathrm{scaledsinkhorn}(\mathbf{A})$ are all concave functions of $\mathbf{Q}$ after taking log. The log concavity of $\mathrm{sinkhorn}(\mathbf{A})$ and $\mathrm{scaledsinkhorn}(\mathbf{A})$ objectives is immediate due to the use of the entropy function. However, showing that the objective of $\mathrm{bethe}(\mathbf{A})$ is log concave is nontrivial and shown in Vontobel (2013).

Lemma 3.1 (Stirling's approximation) *For all $n \in \mathbb{Z}_+$, the following inequalities hold:* $\exp(n\log n - n) \leq n! \leq O(\sqrt{n})\exp(n\log n - n)$.

3.1. Profile maximum likelihood

Let $\Delta^{\mathcal{D}} \subset [0,1]^{\mathcal{D}}_{\mathbb{R}}$ be the set of all discrete distributions supported on domain $\mathcal{D}$. Here on we use the word distribution to refer to discrete distributions. Throughout this paper we assume that we receive a sequence of n independent samples from an underlying distribution $\mathbf{p} \in \Delta^{\mathcal{D}}$. Let $\mathcal{D}^n$ be the set of all length n sequences and $y^n \in \mathcal{D}^n$ be one such sequence with y^n_i denoting its i'th element. The probability of observing sequence y^n is:

$$\mathbb{P}(\mathbf{p}, y^n) \stackrel{\mathrm{def}}{=} \prod_{x\in\mathcal{D}} \mathbf{p}_x^{\mathbf{f}(y^n,x)}$$

where $\mathbf{f}(y^n,x) = |\{i \in [n] \mid y^n_i = x\}|$ is the frequency/multiplicity of symbol x in sequence y^n and $\mathbf{p}_x$ is the probability of domain element $x \in \mathcal{D}$. For any given sequence one could define its profile (histogram of a histogram or fingerprint) that is sufficient statistic for symmetric property estimation.

Definition 3.2 (Profile) *For any sequence $y^n \in \mathcal{D}^n$, let $\boldsymbol{M} = \{\boldsymbol{f}(y^n,x)\}_{x\in\mathcal{D}}\backslash\{0\}$ be the set of all its non-zero distinct frequencies and $\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_{|\boldsymbol{M}|}$ be elements of the set $\boldsymbol{M}$. The* profile *of a sequence $y^n \in \mathcal{D}^n$ denoted by $\phi = \Phi(y^n) \in \mathbb{Z}_+^{|\boldsymbol{M}|}$ is $\phi \stackrel{\mathrm{def}}{=} (\phi_j)_{j\in[1,|\boldsymbol{M}|]}$, where $\phi_j = \phi_j(y^n) \stackrel{\mathrm{def}}{=} |\{x \in \mathcal{D} \mid \boldsymbol{f}(y^n,x) = \boldsymbol{m}_j\}|$[3]. We call n the length of profile ϕ and let Φ^n denote the set of all profiles of length n. We use k to denote the number of distinct frequencies and $k = |\boldsymbol{M}|$.*[4]

For any distribution $\mathbf{p} \in \Delta^{\mathcal{D}}$, the probability of a profile $\phi \in \Phi^n$ is defined as:

$$\mathbb{P}(\mathbf{p}, \phi) \stackrel{\mathrm{def}}{=} \sum_{\{y^n\in\mathcal{D}^n \mid \Phi(y^n)=\phi\}} \mathbb{P}(\mathbf{p}, y^n) \qquad (2)$$

The profile maximum likelihood and approximate profile maximum likelihood distributions are defined as follows.

Definition 3.3 (Profile maximum likelihood) *For any profile $\phi \in \Phi^n$, a* profile maximum likelihood *(PML) distribution $\boldsymbol{p}_{\mathrm{pml},\phi} \in \Delta^{\mathcal{D}}$ is: $\boldsymbol{p}_{\mathrm{pml},\phi} \in \arg\max_{\boldsymbol{p}\in\Delta^{\mathcal{D}}} \mathbb{P}(\boldsymbol{p},\phi)$ and $\mathbb{P}(\boldsymbol{p}_{\mathrm{pml},\phi},\phi)$ is the maximum PML objective value.*

Definition 3.4 (Approximate PML) *For any profile $\phi \in \Phi^n$, a distribution $\boldsymbol{p}^{\beta}_{\mathrm{pml},\phi} \in \Delta^{\mathcal{D}}$ is a* β-approximate PML *distribution if $\mathbb{P}(\boldsymbol{p}^{\beta}_{\mathrm{pml},\phi},\phi) \geq \beta \cdot \mathbb{P}(\boldsymbol{p}_{\mathrm{pml},\phi},\phi)$.*

[3] The profile does not contain information about the number of unseen domain elements.

[4] Note that the number of distinct frequencies denoted by k in a length n sequence is always upper bounded by $\sqrt{n}$.

4. Results

Here we state the main results of this paper. In our first result, we draw a connection between the Sinkhorn and Bethe permanent approximations to the PML, and provide an efficient algorithm to compute an $\exp(-O(\sqrt{n}\log n))$ approximate PML distribution. We defer the proof of the following theorem to Appendix C.

Theorem 4.1 ($\exp(\sqrt{n}\log n)$**-approximate PML)** *For any given profile $\phi \in \Phi^n$, Algorithm 4 computes an $\exp(-O(\sqrt{n}\log n))$-approximate PML distribution in $\widetilde{O}(n^{1.5})$ time.*

Previously the best known result by Charikar et al. (2019a) gave an efficient algorithm to compute $\exp(-O(n^{2/3}\log n))$-approximate PML distribution. One important application of approximate PML is in symmetric property estimation. In Acharya et al. (2016), the authors showed that a β-approximate PML distribution based plug-in estimator is sample complexity optimal for estimating certain symmetric properties; the approximation factor β affects the error parameter regime under which the estimator is sample complexity optimal. Combining their result with our Theorem 4.1, we get an efficient version of Theorem 2 in Acharya et al. (2016); we summarize this result next.

Theorem 4.2 (Efficient universal estimator using approximate PML) *Let n be the optimal sample complexity of estimating entropy, support, support coverage and distance to uniformity. If $\epsilon \geq \frac{c}{n^{0.2499}}$ for some constant $c > 0$, then there exists a PML based universal plug-in estimator that runs in time $\widetilde{O}(n^{1.5})$ and is sample complexity optimal for estimating entropy, support, support coverage and distance to uniformity to accuracy ϵ.*

Note that the dependency on ϵ in Theorem 4.2 and the approximation factor in Theorem 4.1 are strictly better than Charikar et al. (2019a), which is the previous efficient PML based approach for universal symmetric property estimation; Charikar et al. (2019a) works when $\epsilon \geq \frac{1}{n^{0.166}}$.

Recent work Hao and Orlitsky (2019) shows the optimality of an approximate PML distribution based estimator for other symmetric properties, such as sorted distribution estimation (under ℓ_1 distance), α-Renyi entropy for non-integer $\alpha > 3/4$, and other broad class of additive properties that are Lipschitz. Hao and Orlitsky (2019) also provides a PML-based tester to test whether an unknown distribution is $\geq \epsilon$ far from a given distribution in ℓ_1 distance and achieves the optimal sample complexity up to logarithmic factors. Our result further implies an efficient version of all these results (for a broader range of error ϵ than could be achieved by using Charikar et al. (2019a)).

As mentioned in Sections 1 and 2, we achieve the above results through the improved analysis of the Bethe and scaled Sinkhorn permanent approximations of low non-negative matrices. Recall, that for any fixed distribution $\mathbf{p}$ and profile ϕ, $\mathbb{P}(\mathbf{p}, \phi)$ is proportional to the permanent of the non-negative matrix $\mathbf{A}^{\mathbf{p},\phi}$ (See Equation (67) for the definition of $\mathbf{A}^{\mathbf{p},\phi}$). Note that the number of distinct columns in the profile probability matrix $\mathbf{A}^{\mathbf{p},\phi}$ is upper bounded by the number of distinct frequencies plus one, which further is always less than $\sqrt{n}+1$. Therefore the non-negative rank of the profile probability matrix $\mathbf{A}^{\mathbf{p},\phi}$ is always upper bounded by $\sqrt{n}+1$. In our next result, we show that the scaled Sinkhorn permanent approximates the permanent of any non-negative matrix $\mathbf{A}$, where the approximation factor (up to log factors) depends exponentially on the non-negative rank of the matrix $\mathbf{A}$. Since $\mathrm{scaledsinkhorn}(\mathbf{A})$ can be computed in polynomial time Charikar et al. (2019a)[5], our next

[5] $\mathrm{scaledsinkhorn}(\mathbf{A})$ corresponds to a convex optimization problem and a minor modification of the approach in Charikar et al. (2019a) to solve a related, but slightly different optimization problem, yields a polynomial time algorithm to compute $\mathrm{scaledsinkhorn}(\mathbf{A})$ up to high accuracy.

theorem implies an efficient algorithm to approximate the value $\mathbb{P}(\mathbf{p}, \phi)$ for a fixed distribution **p** up to multiplicative $\exp(O(-k\log n))$ factor, where $k \leq \sqrt{n}$ is the number of distinct frequencies in the profile and is the best known approximation factor achieved by a deterministic algorithm. We formally state the scaled Sinkhorn result next and defer its proof to Appendix A.

Theorem 4.3 (Scaled Sinkhorn permanent approximation for low non-negative rank matrices) *For any matrix $\boldsymbol{A} \in \mathbb{R}_{\geq 0}^{\mathcal{D}\times\mathcal{D}}$ with non-negative rank at most k, the following inequality holds,*

$$\mathrm{scaledsinkhorn}(\boldsymbol{A}) \leq \mathrm{perm}(\boldsymbol{A}) \leq \exp\left(O\left(k\log\frac{N}{k}\right)\right)\mathrm{scaledsinkhorn}(\boldsymbol{A})\ . \tag{3}$$

Further using $\mathrm{scaledsinkhorn}(\mathbf{A}) \leq \mathrm{bethe}(\mathbf{A})$ (See Corollary A.5) and $\mathrm{bethe}(\mathbf{A}) \leq \mathrm{perm}(\mathbf{A})$ (See Lemma A.2) we immediately get the same result for the Bethe permanent.

Corollary 4.4 (Bethe permanent approximation for low non-negative rank matrices) *For any matrix $\boldsymbol{A} \in \mathbb{R}_{\geq 0}^{\mathcal{D}\times\mathcal{D}}$ with non-negative rank at most k, the following inequality holds,*

$$\mathrm{bethe}(\boldsymbol{A}) \leq \mathrm{perm}(\boldsymbol{A}) \leq \exp\left(O\left(k\log\frac{N}{k}\right)\right)\mathrm{bethe}(\boldsymbol{A})\ . \tag{4}$$

Interestingly, in the worst case, Sinkhorn is an e^N approximation to the permanent of $\mathbf{A} \in \mathbb{R}_{\geq 0}^{\mathcal{D}\times\mathcal{D}}$, even when **A** has at most 1 distinct column (consider the all 1's matrix). Consequently, for matrices with non-negative rank at most k, whenever $k = o(N/\log N)$, scaled Sinkhorn is a compelling alternative to Sinkhorn, with a tighter worst-case multiplicative approximation to the permanent.

Our results improve the analysis of the Bethe permanent for such structured matrices. Previously, the best known analysis of the Bethe permanent showed an $\sqrt{2}^N$-approximation factor to the permanent Anari and Rezaei (2018). The analysis in Anari and Rezaei (2018) is tight for general non-negative matrices and the authors showed that this bound cannot be improved without leveraging further structure. Our next result is of a similar flavor, and we provide an asymptotically tight example for Theorem 4.3 and Corollary 4.4. Refer Appendix B for the proof of the following theorem.

Theorem 4.5 (Lower bound for the Bethe and the scaled Sinkhorn permanents approximation) *There exists a matrix $\boldsymbol{A} \in \mathbb{R}_{\geq 0}^{\mathcal{D}\times\mathcal{D}}$ with non-negative rank k, that satisfies*

$$\mathrm{perm}(\boldsymbol{A}) \geq \exp\left(\Omega\left(k\log\frac{N}{k}\right)\right)\mathrm{bethe}(\boldsymbol{A})\ , \tag{5}$$

which further implies,

$$\mathrm{perm}(\boldsymbol{A}) \geq \exp\left(\Omega\left(k\log\frac{N}{k}\right)\right)\mathrm{scaledsinkhorn}(\boldsymbol{A})\ . \tag{6}$$

4.1. Related work

We divide this section into two parts: profile maximum likelihood and permanent approximations.

Profile maximum likelihood: As discussed in the introduction, PML was introduced by Orlitsky et al. (2004). Many heuristic approaches such as the EM algorithm Orlitsky et al. (2004), algebraic approaches Acharya et al. (2010) and a dynamic programming approach Pavlichin et al. (2017) have been proposed to compute approximations to PML. Further Vontobel (2012, 2014) used the Bethe permanent as a heuristic to compute the PML distribution. All these approaches don't provide theoretical guarantees for the quality of the approximate PML distribution and it was an open question to efficiently compute a non-trivial approximate PML distribution. Charikar et al. (2019a) gave the first efficient algorithm to compute the $\exp(-n^{2/3}\log n)$ approximate PML distribution.

The connection between PML and universal estimators was first studied in Acharya et al. (2016). Acharya et al. (2016) showed that an approximate PML distribution can be used as an universal estimator that is sample optimal for estimating symmetric properties, namely entropy, distance to uniformity, support size and coverage when error $\epsilon > n^{-0.249}$. See Hao and Orlitsky (2019) for broad applicability of approximate PML in property testing and estimating other symmetric properties such as sorted ℓ_1 distance, Renyi entropy, and other broad class of additive properties. Very recently, authors in Han and Shiragur (2020) provide an improved competitive analysis of the PML, where they show that the PML based plug-in approach is sample complexity optimal in estimating sorted ℓ_1 distance and various other symmetric properties when $\epsilon > n^{-0.333}$. In a follow up work Han (2020), one of the authors from the previous work, further show that the condition on the accuracy parameter $\epsilon > n^{-0.333}$ is actually tight for PML and other broad class of reasonable universal estimators.

Charikar et al. (2019a) combined with Acharya et al. (2016), gave the first efficient PML based universal estimator for symmetric property estimation. There have been several other approaches for designing universal estimators for symmetric properties. Valiant and Valiant (2011b) adopted and rigorously analyzed a linear programming based approach for universal estimators proposed by Efron and Thisted (1976) and showed that it is sample complexity optimal in the constant error regime for estimating certain symmetric properties (namely, entropy, support size, support coverage, and distance to uniformity). Recent work of Han et al. (2018) applied a local moment matching based approach in designing efficient universal symmetric property estimators for a single distribution. Han et al. (2018) achieves the optimal sample complexity in a broader error regimes for estimating the power sum function, support and entropy.

In Charikar et al. (2019b); Hao and Orlitsky (2019) it was shown that variants of PML called *PseudoPML* and *truncated PML* respectively, which compute an approximate PML distribution on a subset of the coordinates, yield sample optimal estimators in broader error regime for a wide range of symmetric properties. Further, in Hao and Orlitsky (2020) an instance dependent quantity known as *profile entropy* was shown to govern the accuracy achievable by PML and their analysis holds for all symmetric properties with no additional assumption on the structure of the property.

Estimating symmetric properties of a distribution is a rich field and extensive work has been dedicated to studying their optimal sample complexity for estimating each of these properties. Optimal sample complexities for estimating many symmetric properties were resolved in the past few years; support Valiant and Valiant (2011b); Wu and Yang (2015), support coverage Orlitsky et al. (2016); Zou et al. (2016), entropy Valiant and Valiant (2011b); Wu and Yang (2016), distance to uniformity Valiant and Valiant (2011a); Jiao et al. (2016), sorted ℓ_1 distance Valiant and Valiant (2011a); Han et al. (2018), Renyi entropy Acharya et al. (2014, 2017), KL divergence Bu et al. (2016); Han et al. (2016) and many others.

Comparison to Charikar et al. (2019a): As discussed earlier, Charikar et al. (2019a) provides an efficient algorithm to compute an $\exp(-n^{2/3}\log n)$-approximate PML distribution. Suppose

ℓ and k are the number of distinct probability values and frequencies respectively, then Charikar et al. (2019a) provides a convex program that using *combinatorial techniques* they analyze and show that it approximates the PML objective up to $\exp(-\widetilde{O}(\ell \times k))$ multiplicative factor. Further this convex program outputs a fractional solution and Charikar et al. (2019a) provide a rounding algorithm that outputs a valid integral solution (that corresponds to a valid distribution). Charikar et al. (2019a) further incur a $\exp(-\widetilde{O}(\ell \times k))$ multiplicative loss in the rounding procedure. Using the discretization results, up to $\exp(-n^{2/3} \log n)$-multiplicative loss one can assume $\ell, k \leq n^{1/3}$ and therefore Charikar et al. (2019a) output a $\exp(-n^{2/3} \log n)$-approximate PML distribution.

However in our current work, using results for the scaled Sinkhorn permanent, we show that the same convex program in Charikar et al. (2019a) approximates the PML objective up to $\exp(-\widetilde{O}(\ell + k))$ multiplicative factor. Further we also provide a better rounding algorithm that outputs a valid distribution and incur a $\exp(-\widetilde{O}(\ell + k))$ multiplicative loss. Further using the discretization results, up to $\exp(-\sqrt{n} \log n)$-multiplicative loss one can assume $\ell, k \leq \sqrt{n}$ and therefore our work provides a $\exp(-\sqrt{n} \log n)$-approximate PML distribution.

Permanent approximations: Valiant (1979) showed that computing the permanent of matrices even when it has entries in 0, 1 is #P-Hard. This led to the study of computing approximations to the permanent. For fixed rank the permanent can be computed in polynomial time Barvinok (1996). Additive approximation to the permanent for arbitrary $\mathbf{A}$ was given by Gurvits (2005). On the other hand, multiplicative approximation to the permanent (or even determining the sign) is hard for general $\mathbf{A}$ Aaronson and Arkhipov (2011); Grier and Schaeffer (2018). This hardness result led to the study of the multiplicative approximation to the permanent for special class of matrices and one such class is the set of non-negative matrices. In this direction, Jerrum et al. (2004) gave the first efficient randomized algorithm to approximate the permanent within $(1 + \epsilon)$ multiplicative accuracy. There has also been a rich literature on coming up with deterministic approximation to the permanent of non-negative matrices. Linial et al. (1998) gave the first deterministic algorithm to the permanent of $N \times N$ non-negative matrices with approximation ratio $\leq e^N$. Gurvits (2011) using an inequality from Schrijver (1998) showed that the Bethe permanent lower bounds the value of the permanent of non-negative matrices. Bethe approximation is based on the Bethe free energy approximation and is very closely connected to the belief propagation algorithm Yedidia et al. (2005); Vontobel (2013). We refer the reader to Vontobel (2013); Gurvits and Samorodnitsky (2014) for the polynomial computability of the Bethe permanent and Anari and Rezaei (2018) for a more rigorous literature survey on the Bethe permanent and other related work.

As discussed in the footnote of the introduction, an anonymous reviewer showed us an alternative and simpler proof for the upper bound on the Sinkhorn approximation to the permanent of matrices with at most k distinct columns (Lemma A.1). This proof is deferred to Appendix D and is derived using Corollary 3.4.5. in Barvinok's book Barvinok (2017). The result in turn, is proved using the Bregman-Minc inequality conjectured by Minc, cf. Spence (1982) and later proved by Bregman Brègman (1973). The Bregman-Minc inequality is well-known and there are many different proofs Schrijver (1978); Radhakrishnan (1997); Alon and Spencer (2004) known. In comparison to this alternative proof for matrices with k distinct columns, our proof is self contained and intuitive. We believe it could help provide further insights into the Sinkhorn/Bethe approximations.

5. Illustrative example

The algorithm to compute an approximate PML includes several steps and we provide an overview of these steps through an illustrative example. Let $aabc$ be the observed sample and $\phi = \{\{2, 1, 1\}\}$[6] be its corresponding profile. Although the number of unseen domain elements is unknown while computing PML, for simplicity we assume it to be 1 and let d be this unseen domain element. Let $\mathcal{D} = \{a, b, c, d\}$ denote the complete domain and $\Delta^{\mathcal{D}}$ be the set of all distributions supported on $\mathcal{D}$.

Probability of a profile For a distribution $\mathbf{p} = (\mathbf{p}_a, \mathbf{p}_b, \mathbf{p}_c, \mathbf{p}_d) \in \Delta^{\mathcal{D}}$, the probability of the profile $\{\{2, 1, 1\}\}$ with respect to $\mathbf{p}$ is given by,

$$\begin{aligned}\mathbb{P}(\mathbf{p}, \phi) = 12[&\mathbf{p}_a\mathbf{p}_b(\mathbf{p}_c^2\mathbf{p}_d^0 + \mathbf{p}_c^0\mathbf{p}_d^2) + \mathbf{p}_a\mathbf{p}_c(\mathbf{p}_b^2\mathbf{p}_d^0 + \mathbf{p}_b^0\mathbf{p}_d^2) + \mathbf{p}_a\mathbf{p}_d(\mathbf{p}_b^2\mathbf{p}_c^0 + \mathbf{p}_b^0\mathbf{p}_c^2) \\ &+ \mathbf{p}_b\mathbf{p}_c(\mathbf{p}_a^2\mathbf{p}_d^0 + \mathbf{p}_a^0\mathbf{p}_d^2) + \mathbf{p}_b\mathbf{p}_d(\mathbf{p}_a^2\mathbf{p}_c^0 + \mathbf{p}_a^0\mathbf{p}_c^2) + \mathbf{p}_c\mathbf{p}_d(\mathbf{p}_a^2\mathbf{p}_b^0 + \mathbf{p}_a^0\mathbf{p}_b^2)] \,.\end{aligned} \tag{7}$$

The first two terms in the summation correspond to the sequences $abcc$ and $abdd$ respectively. The factor 12 further counts the permutations of those sequences. The other summation terms similarly correspond to other sequences whose profile is $\{\{2, 1, 1\}\}$.

Permanent formulation and PML It is often convenient to think of the Equation (7) in terms of the permanent of a matrix. For any $\mathbf{p} \in \Delta^{\mathcal{D}}$ and profile ϕ, it is known that $\mathbb{P}(\mathbf{p}, \phi)$ (Equation (7)) is proportional to the permanent of the following generalized Vandermonde matrix (Vontobel (2012)),

$$\mathbf{A}_{\mathbf{p}} = \begin{bmatrix} \mathbf{p}_a^2 & \mathbf{p}_a & \mathbf{p}_a & \mathbf{p}_a^0 \\ \mathbf{p}_b^2 & \mathbf{p}_b & \mathbf{p}_b & \mathbf{p}_b^0 \\ \mathbf{p}_c^2 & \mathbf{p}_c & \mathbf{p}_c & \mathbf{p}_c^0 \\ \mathbf{p}_d^2 & \mathbf{p}_d & \mathbf{p}_d & \mathbf{p}_d^0 \end{bmatrix} .$$

Consequently, computing a PML distribution corresponds to maximizing $\mathrm{perm}(\mathbf{A}_{\mathbf{p}})$ over all $\mathbf{p} \in \Delta^{\mathcal{D}}$.

Sinkhorn permanent approximation In our work, we consider the Sinkhorn permanent to approximate the permanent. For a fixed distribution $\mathbf{p} \in \Delta^{\mathcal{D}}$, the Sinkhorn permanent (Section 3) approximation to $\mathrm{perm}(\mathbf{A}_{\mathbf{p}})$ is given by the following optimization problem,

$$\begin{gathered}\mathrm{scaledsinkhorn}(\mathbf{A}_{\mathbf{p}}) = \max_{\mathbf{Q} \in \mathbb{R}^{\mathcal{D} \times \mathcal{D}}} \sum_{x,y \in \mathcal{D}} \mathbf{Q}_{x,y} \log \frac{\mathbf{p}_x^{f_y}}{\mathbf{Q}_{x,y}} \,, \\ \text{such that } \sum_{x \in \mathcal{D}} \mathbf{Q}_{x,y} = 1 \;\; \forall\, y \in \mathcal{D}, \text{ and } \sum_{y \in \mathcal{D}} \mathbf{Q}_{x,y} = 1 \;\; \forall\, x \in \mathcal{D} \,,\end{gathered} \tag{8}$$

where f_y is the frequency of $y \in \mathcal{D}$ ($f_a = 2, f_b = 1, f_c = 1, f_d = 0$) and 4 is the dimension of $\mathbf{A}_{\mathbf{p}}$. For a fixed $\mathbf{p} \in \Delta^{\mathcal{D}}$, the above problem can be solved by convex optimization methods as the constraints are linear and the objective is log concave. However recall that our goal is to maximize over all $\mathbf{p} \in \Delta^{\mathcal{D}}$. Unfortunately, the objective is not jointly log concave in Q and p in general.

New formulation To handle the issue of maximizing over all distributions, we rewrite the optimization problem 8. To illustrate this step and for simplicity, we consider a distribution $\mathbf{p}$ that takes all its probability values in the set $\{r_1, r_2\}$ and let $\mathbf{p}_a = \mathbf{p}_b = \mathbf{p}_c = r_1$, $\mathbf{p}_d = r_2$. Note that the rows

[6] We use double brackets to denote the multiset

in $\mathbf{A_p}$ corresponding to elements a, b, c are all the same as they share the same probability value and due to symmetry, it is immediate that there exists an optimum solution $\mathbf{Q}^*$ to problem 8 such that $\mathbf{Q}^*_{a.} = \mathbf{Q}^*_{b.} = \mathbf{Q}^*_{c.}$. Therefore it is sufficient to maximize 8 over doubly stochastic matrices $\mathbf{Q}$ that satisfy $\mathbf{Q}_{a.} = \mathbf{Q}_{b.} = \mathbf{Q}_{c.}$. Further, note that all of these matrices have at most two distinct rows denoted by $\mathbf{Q}_{1.}$ and $\mathbf{Q}_{2.}$ corresponding to probability values $\{r_1, r_2\}$ respectively. These structured matrices $\mathbf{Q}$ have one to one correspondence to low dimensional matrices $\mathbf{P}$ that consist of two rows, where $\mathbf{P}_{1.} = 3\mathbf{Q}_{1.}$, $\mathbf{P}_{2.} = \mathbf{Q}_{2.}$ and the optimization problem 8 can be rewritten as follows,

$$\max_{\mathbf{P}\in\mathbb{R}^{[1,2]\times\mathcal{D}}} \sum_{i\in[1,2],y\in\mathcal{D}} \mathbf{P}_{i,y} \log \frac{r_i^{f_y}}{\mathbf{P}_{i,y}} + \sum_{i\in[1,2]} \Big(\sum_{y\in\mathcal{D}} \mathbf{P}_{i,y}\Big) \log \Big(\sum_{y\in\mathcal{D}} \mathbf{P}_{i,y}\Big) , \tag{9}$$

$$\text{such that} \sum_{i\in[1,2]} \mathbf{P}_{i,y} = 1 \text{ for all } y \in \mathcal{D}, \sum_{y\in\mathcal{D}} \mathbf{P}_{1,y} = 3 \text{ and} \sum_{y\in\mathcal{D}} \mathbf{P}_{2,y} = 1 \ .$$

Note that the values 3 and 1 on the right hand side of row constraints correspond to the number of domain elements with probability values $\{r_1, r_2\}$ respectively. A similar argument as the one we applied to rows/probabilities can also be applied to columns/frequencies and we can further compress columns of matrix $\mathbf{Q}$ (equivalently $\mathbf{P}$) to get a new variable matrix $\mathbf{S}$ that has dimension $\ell \times k$, where ℓ denotes the distinct probabilities and k the distinct columns in $\mathbf{Q}$ (same as distinct columns in $\mathbf{P}$) which is further equal to the number of distinct frequencies. Now, rewriting the above optimization problem in terms of $\mathbf{S}_{ij}$ yields the following optimization problem,

$$\max_{\mathbf{S}\in\mathbb{R}^{[1,2]\times[0,2]}} \sum_{i\in[1,2],j\in[0,2]} \mathbf{S}_{i,j} \log \frac{r_i^j}{\mathbf{S}_{i,j}} + \sum_{i\in[1,2]} \Big(\sum_{j\in[0,2]} \mathbf{S}_{i,j}\Big) \log \Big(\sum_{j\in[0,2]} \mathbf{S}_{i,j}\Big) + \sum_{j\in[0,2]} \phi_j \log \phi_j \ ,$$

$$\text{such that} \sum_{i\in[1,2]} \mathbf{S}_{i,j} = \phi_j \text{ for all } j \in [0.2], \sum_{j\in[0,2]} \mathbf{S}_{1,j} = 3 \text{ and} \sum_{j\in[0,2]} \mathbf{S}_{2,j} = 1 \ , \tag{10}$$

where ϕ_j denotes the number of domain elements with frequency j ($\phi_0 = 1, \phi_1 = 2, \phi_2 = 1$). In general, the optimization problem 8 always exhibits an optimum solution that assigns same values to rows and columns that share equal probability and frequency values respectively. Therefore for any other distribution $\mathbf{p}'$ that has probabilities in set $\{r_1, r_2\}$, the probability of profile ϕ with respect to $\mathbf{p}'$ can be approximated by the above optimization problem by just replacing the right hand side of row sum constraints with values x_1 and x_2 that count the number of domain elements in $\mathbf{p}'$ with probability r_1 and r_2 respectively. Further, any distribution that takes all its probability values in some set $\mathbf{R}$ can be approximated by extending the above optimization problem by including a row constraint for each probability value in $\mathbf{R}$.

Maximizing over all distributions, convex program and rounding As discussed, the i'th row sum of the variable matrix $\mathbf{S}$ indicates the number of domain elements with probability r_i. To handle the maximization over all distributions, we remove all the row constraints in optimization problem (10) and replace them by just one constraint $\sum_{i\in\mathbf{R}} r_i \left(\sum_j \mathbf{S}_{ij}\right) \leq 1$ which serves as a proxy to capture all pseudo-distributions. This new optimization problem with the above constraint is our final convex program with variable matrix $\mathbf{S}$. The optimum solution to this new optimization problem may not necessarily have integral row sums and therefore might not correspond to a valid distribution. Therefore, in our final step we design a rounding algorithm that rounds these fractional row sums to integral while not losing much in the objective. Such a rounded solution then corresponds to an approximate PML distribution.

Acknowledgments

We thank reviewers on earlier versions of this paper for their thoughtful comments. Nima Anari was supported in part by an NSF CAREER Award CCF-2045354 and a Google Faculty Research Award. Moses Charikar was supported by a Simons Investigator award. Kirankumar Shiragur was supported by a Stanford Data Science Scholarship and a Dantzig-Lieberman Operations Research Fellowship. Aaron Sidford was supported in part by a Microsoft Research Faculty Fellowship, NSF CAREER Award CCF-1844855, NSF Grant CCF-1955039, a PayPal research award, and a Sloan Research Fellowship.

References

Scott Aaronson and Alex Arkhipov. The computational complexity of linear optics. In *Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing*, STOC '11, pages 333–342, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0691-1. doi: 10.1145/1993636.1993682. URL http://doi.acm.org/10.1145/1993636.1993682.

J. Acharya, H. Das, H. Mohimani, A. Orlitsky, and S. Pan. Exact calculation of pattern probabilities. In *2010 IEEE International Symposium on Information Theory*, pages 1498–1502, June 2010. doi: 10.1109/ISIT.2010.5513565.

Jayadev Acharya, Alon Orlitsky, Ananda Theertha Suresh, and Himanshu Tyagi. The complexity of estimating rényi entropy. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1855–1869, 2014. doi: 10.1137/1.9781611973730.124. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611973730.124.

Jayadev Acharya, Hirakendu Das, Alon Orlitsky, and Ananda Theertha Suresh. A unified maximum likelihood approach for optimal distribution property estimation. *CoRR*, abs/1611.02960, 2016. URL http://arxiv.org/abs/1611.02960.

Jayadev Acharya, Alon Orlitsky, Ananda Theertha Suresh, and Himanshu Tyagi. Estimating renyi entropy of discrete distributions. *IEEE Trans. Inf. Theor.*, 63(1):38–56, January 2017. ISSN 0018-9448. doi: 10.1109/TIT.2016.2620435. URL https://doi.org/10.1109/TIT.2016.2620435.

Noga Alon and Joel H Spencer. *The probabilistic method.* John Wiley & Sons, 2004.

N. Anari, S. Oveis Gharan, and C. Vinzant. Log-concave polynomials, entropy, and a deterministic approximation algorithm for counting bases of matroids. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 35–46, Los Alamitos, CA, USA, oct 2018. IEEE Computer Society. doi: 10.1109/FOCS.2018.00013. URL https://doi.ieeecomputersociety.org/10.1109/FOCS.2018.00013.

Nima Anari and Alireza Rezaei. A tight analysis of bethe approximation for permanent. *CoRR*, abs/1811.02933, 2018. URL http://arxiv.org/abs/1811.02933.

Nima Anari, Moses Charikar, Kirankumar Shiragur, and Aaron Sidford. Instance based approximations to profile maximum likelihood, 2020.

Alexander Barvinok. *Combinatorics and Complexity of Partition Functions*. Springer Publishing Company, Incorporated, 1st edition, 2017. ISBN 3319518283.

Alexander I Barvinok. Two algorithmic results for the traveling salesman problem. *Mathematics of Operations Research*, 21(1):65–84, 1996.

Lev Meerovich Brègman. Some properties of nonnegative matrices and their permanents. *Doklady Akademii Nauk*, 211(1):27–30, 1973.

Y. Bu, S. Zou, Y. Liang, and V. V. Veeravalli. Estimation of kl divergence between large-alphabet distributions. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 1118–1122, July 2016. doi: 10.1109/ISIT.2016.7541473.

John Bunge and Michael Fitzpatrick. Estimating the number of species: a review. *Journal of the American Statistical Association*, 88(421):364–373, 1993.

A Chao. Nonparametric estimation of the number of classes in a population. scandinavianjournal of statistics11, 265-270. *Chao26511Scandinavian Journal of Statistics1984*, 1984.

Anne Chao and Shen-Ming Lee. Estimating the number of classes via sample coverage. *Journal of the American statistical Association*, 87(417):210–217, 1992.

Moses Charikar, Kirankumar Shiragur, and Aaron Sidford. Efficient profile maximum likelihood for universal symmetric property estimation. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2019, pages 780–791, New York, NY, USA, 2019a. ACM. ISBN 978-1-4503-6705-9. doi: 10.1145/3313276.3316398. URL http://doi.acm.org/10.1145/3313276.3316398.

Moses Charikar, Kirankumar Shiragur, and Aaron Sidford. A general framework for symmetric property estimation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 12426–12436, 2019b.

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

Timothy Daley and Andrew D Smith. Predicting the molecular complexity of sequencing libraries. *Nature methods*, 10(4):325, 2013.

Bradley Efron and Ronald Thisted. Estimating the number of unseen species: How many words did shakespeare know? *Biometrika*, 63(3):435–447, 1976.

Johannes Fürnkranz. Web mining. In *Data mining and knowledge discovery handbook*, pages 899–920. Springer, 2005.

Zhan Gao, Chi-hong Tseng, Zhiheng Pei, and Martin J Blaser. Molecular analysis of human forearm superficial skin bacterial biota. *Proceedings of the National Academy of Sciences*, 104(8): 2927–2932, 2007.

Daniel Grier and Luke Schaeffer. New hardness results for the permanent using linear optics. In *Proceedings of the 33rd Computational Complexity Conference*, CCC '18, pages 19:1–19:29, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 978-3-95977-069-9. doi: 10.4230/LIPIcs.CCC.2018.19. URL https://doi.org/10.4230/LIPIcs.CCC.2018.19.

Leonid Gurvits. On the complexity of mixed discriminants and related problems. In *Proceedings of the 30th International Conference on Mathematical Foundations of Computer Science*, MFCS'05, pages 447–458, Berlin, Heidelberg, 2005. Springer-Verlag. ISBN 3-540-28702-7, 978-3-540-28702-5. doi: 10.1007/11549345_39. URL http://dx.doi.org/10.1007/11549345_39.

Leonid Gurvits. Unleashing the power of Schrijver's permanental inequality with the help of the Bethe Approximation. *arXiv e-prints*, art. arXiv:1106.2844, Jun 2011.

Leonid Gurvits and Alex Samorodnitsky. Bounds on the permanent and some applications. *arXiv e-prints*, art. arXiv:1408.0976, Aug 2014.

Yanjun Han. On the high accuracy limitation of adaptive property estimation, 2020.

Yanjun Han and Kirankumar Shiragur. On the competitive analysis and high accuracy optimality of profile maximum likelihood, 2020.

Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax estimation of KL divergence between discrete distributions. *CoRR*, abs/1605.09124, 2016. URL http://arxiv.org/abs/1605.09124.

Yanjun Han, Jiantao Jiao, and Rajarshi Mukherjee. On Estimation of L_{r}-Norms in Gaussian White Noise Models. *arXiv e-prints*, art. arXiv:1710.03863, Oct 2017a.

Yanjun Han, Jiantao Jiao, Tsachy Weissman, and Yihong Wu. Optimal rates of entropy estimation over Lipschitz balls. *arXiv e-prints*, art. arXiv:1711.02141, Nov 2017b.

Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Local moment matching: A unified methodology for symmetric functional estimation and distribution estimation under wasserstein distance. *arXiv preprint arXiv:1802.08405*, 2018.

Yi Hao and Alon Orlitsky. The Broad Optimality of Profile Maximum Likelihood. *arXiv e-prints*, art. arXiv:1906.03794, Jun 2019.

Yi Hao and Alon Orlitsky. Profile entropy: A fundamental measure for the learnability and compressibility of discrete distributions, 2020.

Jennifer B Hughes, Jessica J Hellmann, Taylor H Ricketts, and Brendan JM Bohannan. Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl. Environ. Microbiol.*, 67(10):4399–4406, 2001.

Mark Jerrum, Alistair Sinclair, and Eric Vigoda. A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. *J. ACM*, 51(4):671–697, July 2004. ISSN 0004-5411. doi: 10.1145/1008731.1008738. URL http://doi.acm.org/10.1145/1008731.1008738.

J. Jiao, K. Venkat, Y. Han, and T. Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, May 2015. ISSN 0018-9448. doi: 10.1109/TIT.2015.2412945.

J. Jiao, Y. Han, and T. Weissman. Minimax estimation of the l1 distance. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 750–754, July 2016. doi: 10.1109/ISIT.2016.7541399.

Ian Kroes, Paul W Lepp, and David A Relman. Bacterial diversity within the human subgingival crevice. *Proceedings of the National Academy of Sciences*, 96(25):14547–14552, 1999.

Nathan Linial, Alex Samorodnitsky, and Avi Wigderson. A deterministic strongly polynomial algorithm for matrix scaling and approximate permanents. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, pages 644–652, New York, NY, USA, 1998. ACM. ISBN 0-89791-962-9. doi: 10.1145/276698.276880. URL http://doi.acm.org/10.1145/276698.276880.

A. Orlitsky, N. P. Santhanam, and J. Zhang. Always good turing: asymptotically optimal probability estimation. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 179–188, Oct 2003. doi: 10.1109/SFCS.2003.1238192.

A. Orlitsky, S. Sajama, N. P. Santhanam, K. Viswanathan, and Junan Zhang. Algorithms for modeling distributions over large alphabets. In *International Symposium on Information Theory, 2004. ISIT 2004. Proceedings.*, pages 304–304, 2004. doi: 10.1109/ISIT.2004.1365341.

Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu. Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences*, 113(47):13283–13288, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1607774113. URL http://www.pnas.org/content/113/47/13283.

Bruce J Paster, Susan K Boches, Jamie L Galvin, Rebecca E Ericson, Carol N Lau, Valerie A Levanos, Ashish Sahasrabudhe, and Floyd E Dewhirst. Bacterial diversity in human subgingival plaque. *Journal of bacteriology*, 183(12):3770–3783, 2001.

D. S. Pavlichin, J. Jiao, and T. Weissman. Approximate Profile Maximum Likelihood. *ArXiv e-prints*, December 2017.

Nina T Plotkin and Abraham J Wyner. An entropy estimator algorithm and telecommunications applications. In *Maximum Entropy and Bayesian Methods*, pages 351–363. Springer, 1996.

A. Porta, S. Guzzetti, N. Montano, R. Furlan, M. Pagani, A. Malliani, and S. Cerutti. Entropy, entropy rate, and pattern classification as tools to typify complexity in short heart period variability series. *IEEE Transactions on Biomedical Engineering*, 48(11):1282–1291, Nov 2001. ISSN 0018-9294. doi: 10.1109/10.959324.

Jaikumar Radhakrishnan. An entropy proof of bregman's theorem. *Journal of Combinatorial Theory, Series A*, 77(1):161 – 164, 1997. ISSN 0097-3165. doi: https://doi.org/10.1006/jcta.1996.2727. URL http://www.sciencedirect.com/science/article/pii/S0097316596927272.

Aditi Raghunathan, Gregory Valiant, and James Zou. Estimating the unseen from multiple populations. *CoRR*, abs/1707.03854, 2017. URL http://arxiv.org/abs/1707.03854.

S. Raskhodnikova, D. Ron, A. Shpilka, and A. Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 559–569, Oct 2007. doi: 10.1109/FOCS.2007.47.

Fred Rieke, Davd Warland, Rob de Ruyter van Steveninck, and William Bialek. *Spikes: Exploring the Neural Code*. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-18174-6.

Harlan S Robins, Paulo V Campregher, Santosh K Srivastava, Abigail Wacher, Cameron J Turtle, Orsalem Kahsai, Stanley R Riddell, Edus H Warren, and Christopher S Carlson. Comprehensive assessment of t-cell receptor β-chain diversity in $\alpha\beta$ t cells. *Blood*, 114(19):4099–4107, 2009.

A Schrijver. A short proof of minc's conjecture. *Journal of Combinatorial Theory, Series A*, 25(1):80 – 83, 1978. ISSN 0097-3165. doi: https://doi.org/10.1016/0097-3165(78)90036-5. URL http://www.sciencedirect.com/science/article/pii/0097316578900365.

Alexander Schrijver. Counting 1-factors in regular bipartite graphs. *Journal of Combinatorial Theory, Series B*, 72(1):122 – 135, 1998. ISSN 0095-8956. doi: https://doi.org/10.1006/jctb.1997.1798. URL http://www.sciencedirect.com/science/article/pii/S0095895697917986.

E. Spence. H. minc, permanents (encyclopedia of mathematics and its applications, vol. 6, addison-wesley advanced book programme, 1978), xviii 205 pp., 21.50. *Proceedings of the Edinburgh Mathematical Society*, 25(1):110–110, 1982. doi: 10.1017/S0013091500004284.

Ronald Thisted and Bradley Efron. Did shakespeare write a newly-discovered poem? *Biometrika*, 74(3):445–455, 1987.

G. Valiant and P. Valiant. The power of linear estimators. In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*, pages 403–412, Oct 2011a. doi: 10.1109/FOCS.2011.81.

Gregory Valiant and Paul Valiant. Estimating the unseen: An n/log(n)-sample estimator for entropy and support size, shown optimal via new clts. In *Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing*, STOC '11, pages 685–694, New York, NY, USA, 2011b. ACM. ISBN 978-1-4503-0691-1. doi: 10.1145/1993636.1993727. URL http://doi.acm.org/10.1145/1993636.1993727.

L.G. Valiant. The complexity of computing the permanent. *Theoretical Computer Science*, 8(2):189 – 201, 1979. ISSN 0304-3975. doi: https://doi.org/10.1016/0304-3975(79)90044-6. URL http://www.sciencedirect.com/science/article/pii/0304397579900446.

Martin Vinck, Francesco P. Battaglia, Vladimir B. Balakirsky, A. J. Han Vinck, and Cyriel M. A. Pennartz. Estimation of the entropy based on its polynomial representation. *Phys. Rev. E*, 85: 051139, May 2012. doi: 10.1103/PhysRevE.85.051139. URL https://link.aps.org/doi/10.1103/PhysRevE.85.051139.

P. O. Vontobel. The bethe permanent of a nonnegative matrix. *IEEE Transactions on Information Theory*, 59(3):1866–1901, March 2013. ISSN 0018-9448. doi: 10.1109/TIT.2012.2227109.

P. O. Vontobel. The bethe and sinkhorn approximations of the pattern maximum likelihood estimate and their connections to the valiant-valiant estimate. In *2014 Information Theory and Applications Workshop (ITA)*, pages 1–10, Feb 2014. doi: 10.1109/ITA.2014.6804280.

Pascal O Vontobel. The bethe approximation of the pattern maximum likelihood distribution. In *2012 IEEE International Symposium on Information Theory Proceedings*. IEEE, 2012.

Y. Wu and P. Yang. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *ArXiv e-prints*, April 2015.

Y. Wu and P. Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, June 2016. ISSN 0018-9448. doi: 10.1109/TIT.2016.2548468.

Yihong Wu and Pengkun Yang. Sample complexity of the distinct elements problem. *arXiv e-prints*, art. arXiv:1612.03375, Dec 2016.

J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51(7):2282–2312, July 2005. doi: 10.1109/TIT.2005.850085.

James Zou, Gregory Valiant, Paul Valiant, Konrad Karczewski, Siu On Chan, Kaitlin Samocha, Monkol Lek, Shamil Sunyaev, Mark Daly, and Daniel G. MacArthur. Quantifying unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects. *Nature Communications*, 7:13293 EP –, 10 2016. URL https://doi.org/10.1038/ncomms13293.

Appendix A. The Sinkhorn permanent for structured matrices.

In this section, we provide the proof for our first main theorem (Theorem 4.3). We show that the scaled Sinkhorn permanent of a non-negative matrix approximates its permanent, where the approximation factor is exponential in the non-negative rank of the matrix (up to log factors). Our proof is divided into two parts. First in Appendix A.2, we work with a simpler setting of matrices A with at most k distinct columns and prove the following lemma.

Lemma A.1 (Scaled Sinkhorn permanent approximation) *For any matrix $\boldsymbol{A} \in \mathbb{R}_{\geq 0}^{\mathcal{D}\times\mathcal{D}}$ with at most k distinct columns, the following holds,*

$$\mathrm{scaledsinkhorn}(\boldsymbol{A}) \leq \mathrm{perm}(\boldsymbol{A}) \leq \exp\left(O\left(k\log\frac{N}{k}\right)\right)\mathrm{scaledsinkhorn}(\boldsymbol{A})\ . \tag{11}$$

Further using the above result, in Appendix A.3 we prove our main theorem (Theorem 4.3) for low non-negative rank matrices. We start by providing some basic inequalities related to the Bethe and scaled Sinkhorn permanents.

A.1. Basic inequalities

A well known and important result about the Bethe permanent is that it lower bounds the value of permanent of a non-negative matrix and we state this result next.

Lemma A.2 (Gurvits (2011) based on Schrijver (1998)) *For any non-negative $\boldsymbol{A} \in \mathbb{R}_{\geq 0}^{\mathcal{D}\times\mathcal{D}}$, the following holds,*

$$\mathrm{bethe}(\boldsymbol{A}) \leq \mathrm{perm}(\boldsymbol{A})$$

To establish the relationship between the Bethe and the Sinkhorn permanent we need the following lemma from Gurvits and Samorodnitsky (2014).

Lemma A.3 (Proposition 3.1 in Gurvits and Samorodnitsky (2014)) *For any distribution $\boldsymbol{p} \in \mathbb{R}_{\geq 0}^{\mathcal{D}}$, the following holds,*

$$\sum_{x\in\mathcal{D}}(1-\boldsymbol{p}_x)\log(1-\boldsymbol{p}_x) \geq -1\ .$$

For any matrix $\mathbf{Q} \in \mathbf{K}_{rc}$, each row of $\mathbf{Q}$ is a distribution; therefore the following holds,

$$\mathrm{V}(\mathbf{Q}) \geq -N\ .$$

As a corollary of the above inequality we have,

Corollary A.4 *For any non-negative matrix $\boldsymbol{A} \in \mathbb{R}_{\geq 0}^{\mathcal{D}\times\mathcal{D}}$, the following inequality holds,*

$$\exp(-N)\mathrm{sinkhorn}(\boldsymbol{A}) \leq \mathrm{bethe}(\boldsymbol{A})\ .$$

The above expression can be equivalently stated as,

$$\mathrm{scaledsinkhorn}(\mathbf{A}) = \exp(-N)\mathrm{sinkhorn}(\mathbf{A})\ .$$

Combining Lemma A.2 and Corollary A.4 we get the following result.

Corollary A.5 *For any matrix $\boldsymbol{A} \in \mathbb{R}_{\geq 0}^{\mathcal{D}\times\mathcal{D}}$, the following inequality holds,*

$$\mathrm{scaledsinkhorn}(\boldsymbol{A}) \leq \mathrm{bethe}(\boldsymbol{A}),$$

which further implies,

$$\mathrm{scaledsinkhorn}(\boldsymbol{A}) \leq \mathrm{perm}(\boldsymbol{A})\ .$$

A.2. The Sinkhorn permanent for distinct column case.

We start this section by defining some notation that captures the structure of repetition of columns in a matrix. For the remainder of this section we fix a matrix $\mathbf{A} \in \mathbb{R}_{\geq 0}^{\mathcal{D}\times\mathcal{D}}$. We let k denote the number of distinct columns of $\mathbf{A}$ and use $\mathbf{c}_1, \mathbf{c}_2, \dots \mathbf{c}_k$ to denote these distinct columns. Further we let $\hat{\mathbf{A}} = [\mathbf{c}_1 \mid \mathbf{c}_2 \mid \dots \mid \mathbf{c}_k]$ denote the $\mathcal{D} \times k$ matrix formed by these distinct columns. We use $\mathbf{A}_{:y}$ to denote the y'th column of matrix $\mathbf{A}$ and let $\phi_j \stackrel{\text{def}}{=} |\{y \in \mathcal{D} \mid \mathbf{A}_{:y} = \mathbf{c}_j\}|$ denote the number of columns equal to $\mathbf{c}_j$. It is immediate that,

$$\sum_{j\in[1,k]} \phi_j = N \ , \tag{12}$$

where $N = |\mathcal{D}|$ is the size of the domain. For any matrix $\mathbf{P} \in \mathbb{R}_{\geq 0}^{\mathcal{D}\times k}$ define,

$$\mathbf{f}(\mathbf{A},\mathbf{P}) \stackrel{\text{def}}{=} \sum_{x\in\mathcal{D}} \sum_{j\in[1,k]} \mathbf{P}_{x,j} \log \frac{\hat{\mathbf{A}}_{x,j}}{\mathbf{P}_{x,j}} + \sum_{j\in[1,k]} \phi_j \log \phi_j - \sum_{j\in[1,k]} \phi_j \ . \tag{13}$$

In the first half of this section, we show existence of a matrix $\mathbf{P} \in \mathbb{R}_{\geq 0}^{\mathcal{D}\times k}$ (See Lemma A.8) such that $\sum_{j\in[1,k]} \mathbf{P}_{x,j} = 1$ for all $x \in \mathcal{D}$, $\sum_{x\in\mathcal{D}} \mathbf{P}_{x,j} = \phi_j$ for all $j \in [1,k]$, and further (See Lemma A.9),

$$\log \operatorname{perm}(\mathbf{A}) \leq O\left(k \log \frac{N}{k}\right) + \mathbf{f}(\mathbf{A},\mathbf{P}) \ . \tag{14}$$

Later in the second half (See Lemma A.10), we show that for any matrix $\mathbf{P} \in \mathbb{R}_{\geq 0}^{\mathcal{D}\times k}$ that satisfies $\sum_{j\in[1,k]} \mathbf{P}_{x,j} = 1$ for all $x \in \mathcal{D}$ and $\sum_{x\in\mathcal{D}} \mathbf{P}_{x,j} = \phi_j$ for all $j \in [1,k]$, there exists a matrix $\mathbf{Q} \in \mathbf{K}_{rc}$ (recall $\mathbf{K}_{rc}$ is the set of all $\mathcal{D} \times \mathcal{D}$ doubly stochastic matrices) that satisfies,

$$\mathbf{f}(\mathbf{A},\mathbf{P}) = \mathrm{U}(\mathbf{A},\mathbf{Q}) - N \ . \tag{15}$$

However, using Corollary A.5 we already know that, $\operatorname{scaledsinkhorn}(\mathbf{A}) \leq \operatorname{perm}(\mathbf{A})$. Further using the definition of $\operatorname{scaledsinkhorn}(\mathbf{A})$ and combining with Equations (14) and (15) we get,

$$\operatorname{scaledsinkhorn}(\mathbf{A}) \leq \operatorname{perm}(\mathbf{A}) \leq \exp\left(O\left(k \log \frac{N}{k}\right)\right) \operatorname{scaledsinkhorn}(\mathbf{A}) \ .$$

In the remainder, we provide proofs for all the above mentioned inequalities and we need the following set of definitions. Let $\mathbf{K}_r \subseteq \{0,1\}^{\mathcal{D}\times k}$, be the subset of all $\mathcal{D} \times k$ matrices that are row stochastic, meaning there is exactly a single 1 in each row. Let $\mathbf{K}_{\mathbf{A}} \subseteq \mathbf{K}_r$ be the set of matrices such that any $\mathbf{X} \in \mathbf{K}_{\mathbf{A}}$ satisfies $\sum_{x\in\mathcal{D}} \mathbf{X}_{x,j} = \phi_j$ for all $j \in [1,k]$.

Definition A.6 *Let $\boldsymbol{h}_{\boldsymbol{A}} : S_{\mathcal{D}} \to \boldsymbol{K}_{\boldsymbol{A}}$ be the function that takes a permutation $\sigma \in S_{\mathcal{D}}$ as input and returns a matrix $\boldsymbol{X} \in \boldsymbol{K}_{\boldsymbol{A}}$ in the following way,*

$$\boldsymbol{X}_{x,j} = \begin{cases} 1 & \text{if } \boldsymbol{A}_{:\sigma(x)} = \boldsymbol{c}_j \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } x \in \mathcal{D}. \tag{16}$$

Remark: Note that as desired $\mathbf{h}_{\mathbf{A}}(\sigma) \in \mathbf{K}_{\mathbf{A}}$ for all $\sigma \in S_{\mathcal{D}}$ because of the following. For any $\sigma \in S_{\mathcal{D}}$, let $\mathbf{X} \stackrel{\text{def}}{=} \mathbf{h}_{\mathbf{A}}(\sigma)$. Since $\mathbf{c}_j$ for all $j \in [1,k]$ are distinct, we have $\sum_{j\in[1,k]} \mathbf{X}_{x,j} = 1$. Further for any $j \in [1,k]$, $\sum_{x\in\mathcal{D}} \mathbf{X}_{x,j} = \sum_{\{x\in\mathcal{D} \mid \mathbf{A}_{:\sigma(x)}=\mathbf{c}_j\}} 1 = \sum_{\{x\in\mathcal{D} \mid \mathbf{A}_{\cdot x}=\mathbf{c}_j\}} 1 = \phi_j$.

We next define the probability of a permutation $\sigma \in S_{\mathcal{D}}$ with respect to matrix $\mathbf{A}$ as follows,

$$\Pr(\sigma) \stackrel{\text{def}}{=} \frac{\prod_{e\in\sigma} \mathbf{A}_e}{\mathrm{perm}(\mathbf{A})} \tag{17}$$

Further we define a marginal distribution μ on $\mathbf{K}_r$ and later we will establish that this is indeed a probability distribution, that is, probabilities add up to 1.

$$\mu(\mathbf{X}) \stackrel{\text{def}}{=} \begin{cases} 0 & \text{if } \mathbf{X} \in \mathbf{K}_r \backslash \mathbf{K}_{\mathbf{A}} \\ \sum_{\{\sigma\in S_{\mathcal{D}} \mid \mathbf{h}_{\mathbf{A}}(\sigma)=\mathbf{X}\}} \Pr(\sigma) & \text{if } \mathbf{X} \in \mathbf{K}_{\mathbf{A}}\ . \end{cases} \tag{18}$$

For $\mathbf{X} \in \mathbf{K}_{\mathbf{A}}$, we next provide another equivalent expression for $\mu(\mathbf{X})$.

$$\begin{aligned} \mu(\mathbf{X}) &= \sum_{\{\sigma\in S_{\mathcal{D}} \mid \mathbf{h}_{\mathbf{A}}(\sigma)=\mathbf{X}\}} \Pr(\sigma) = \sum_{\{\sigma\in S_{\mathcal{D}} \mid \mathbf{h}_{\mathbf{A}}(\sigma)=\mathbf{X}\}} \frac{\prod_{(x,\sigma(x))} \mathbf{A}_{x,\sigma(x)}}{\mathrm{perm}(\mathbf{A})}, \\ &= \frac{1}{\mathrm{perm}(\mathbf{A})} \sum_{\{\sigma\in S_{\mathcal{D}} \mid \mathbf{h}_{\mathbf{A}}(\sigma)=\mathbf{X}\}} \prod_{x\in\mathcal{D}} \prod_{j\in[1,k]} \hat{\mathbf{A}}_{x,j}^{\mathbf{X}_{x,j}} \\ &= \left(\prod_{j\in[1,k]} \phi_j!\right) \left(\prod_{x\in\mathcal{D}} \prod_{j\in[1,k]} \hat{\mathbf{A}}_{x,j}^{\mathbf{X}_{x,j}}\right) \left(\frac{1}{\mathrm{perm}(\mathbf{A})}\right) \end{aligned} \tag{19}$$

In the first and second equality, we used definitions of $\mu(\mathbf{X})$ and $\Pr(\sigma)$ (See Equation (17)). For any $\sigma \in S_{\mathcal{D}}$, let $\mathbf{X} = \mathbf{h}_{\mathbf{A}}(\sigma)$. Further for any $x \in \mathcal{D}$, let j' be such that $\mathbf{A}_{:\sigma(x)} = \mathbf{c}_{j'}$, then $\mathbf{A}_{x,\sigma(x)} = \hat{\mathbf{A}}_{x,j'}$ that is further equal to $\prod_{j\in[1,k]} \hat{\mathbf{A}}_{x,j}^{\mathbf{X}_{x,j}}$ because $\mathbf{X}_{x,j}$ is equal to 1 if $j = j'$ and 0 otherwise. Therefore the third equality holds. For the final equality, observe that for any $\sigma \in S_{\mathcal{D}}$ if we let $\mathbf{X} = \mathbf{h}_{\mathbf{A}}(\sigma)$, then for each $j \in [1,k]$, any permutation within the subset of elements $\{x \in \mathcal{D} \mid \mathbf{A}_{:\sigma(x)} = \mathbf{c}_j\}$ results in a permutation σ' that satisfies $\mathbf{h}_{\mathbf{A}}(\sigma') = \mathbf{X}$. These permutations can be carried out independently for each $j \in [1,k]$ that corresponds to $\prod_{j\in[1,k]} \phi_j!$ number of permutations and all of them have the same $\prod_{x\in\mathcal{D}} \prod_{j\in[1,k]} \hat{\mathbf{A}}_{x,j}^{\mathbf{X}_{x,j}}$ value.

Using the derivation from above, the definition for μ can also be written as follows:

$$\mu(\mathbf{X}) = \begin{cases} 0 & \text{if } \mathbf{X} \in \mathbf{K}_r \backslash \mathbf{K}_{\mathbf{A}} \\ \left(\prod_{j\in[1,k]} \phi_j!\right) \left(\prod_{x\in\mathcal{D}} \prod_{j\in[1,k]} \hat{\mathbf{A}}_{x,j}^{\mathbf{X}_{x,j}}\right) \left(\frac{1}{\mathrm{perm}(\mathbf{A})}\right) & \text{if } \mathbf{X} \in \mathbf{K}_{\mathbf{A}}\ . \end{cases} \tag{20}$$

Note that for $\mathbf{X} \in \mathbf{K}_{\mathbf{A}}$, the expression for $\mu(\mathbf{X})$ can be equivalently written as follows:

$$\mu(\mathbf{X}) = \left(\prod_{j\in[1,k]} \phi_j!\right) \left(\prod_{\{(x,j)\in\mathcal{D}\times[1,k] \mid \mathbf{X}_{x,j}=1\}} \hat{\mathbf{A}}_{x,j}\right) \left(\frac{1}{\mathrm{perm}(\mathbf{A})}\right)\ . \tag{21}$$

We next show that the μ defined above is a valid distribution.

$$\sum_{\mathbf{X}\in\mathbf{K}_r} \mu(\mathbf{X}) = \sum_{\mathbf{X}\in\mathbf{K}_{\mathbf{A}}} \mu(\mathbf{X}) = \sum_{\mathbf{X}\in\mathbf{K}_{\mathbf{A}}} \sum_{\{\sigma\in S_{\mathcal{D}} \mid \mathbf{h}_{\mathbf{A}}(\sigma)=\mathbf{X}\}} \Pr(\sigma) = \sum_{\sigma\in S_{\mathcal{D}}} \Pr(\sigma) = 1$$

Remark: The domain of distribution μ is $\mathbf{K}_r$, but its support is subset of $\mathbf{K}_{\mathbf{A}}$.

Definition A.7 *For the distribution μ, we define a non-negative matrix $\boldsymbol{P} \in \mathbb{R}_{\geq 0}^{\mathcal{D}\times k}$ with respect to μ as follows:*

$$\boldsymbol{P}_{x,j} \stackrel{\text{def}}{=} \Pr_{\boldsymbol{X}\sim\mu}(\boldsymbol{X}_{x,j}=1) = \sum_{\{\boldsymbol{X}\in\boldsymbol{K_A} \mid \boldsymbol{X}_{x,j}=1\}} \mu(\boldsymbol{X}) \,. \tag{22}$$

Lemma A.8 *The matrix $\boldsymbol{P}$ defined in Equation (22) satisfies the following conditions:*

$$\sum_{j\in[1,k]} \boldsymbol{P}_{x,j} = 1 \textit{ for all } x \in \mathcal{D} \quad \textit{and} \quad \sum_{x\in\mathcal{D}} \boldsymbol{P}_{x,j} = \phi_j \textit{ for all } j \in [1,k] \,. \tag{23}$$

Proof We first evaluate the row sum. For each $x \in \mathcal{D}$,

$$\sum_{j\in[1,k]} \mathbf{P}_{x,j} = \sum_{j\in[1,k]} \sum_{\{\mathbf{X}\in\mathbf{K_A} \mid \mathbf{X}_{x,j}=1\}} \mu(\mathbf{X}) = \sum_{\mathbf{X}\in\mathbf{K_A}} \mu(\mathbf{X}) = 1 \,.$$

In the second inequality we used that $\mathbf{X} \in \mathbf{K_A}$, meaning for each $x \in \mathcal{D}$, $\sum_{j\in[1,k]} \mathbf{X}_{x,j} = 1$. Next we evaluate the column sum, for each $j \in [1,k]$,

$$\begin{aligned}\sum_{x\in\mathcal{D}} \mathbf{P}_{x,j} &= \sum_{x\in\mathcal{D}} \sum_{\{\mathbf{X}\in\mathbf{K_A} \mid \mathbf{X}_{x,j}=1\}} \mu(\mathbf{X}) = \sum_{\mathbf{X}\in\mathbf{K_A}} \sum_{\{x\in\mathcal{D} \mid \mathbf{X}_{x,j}=1\}} \mu(\mathbf{X}) \\ &= \sum_{\mathbf{X}\in\mathbf{K_A}} \mu(\mathbf{X}) \sum_{\{x\in\mathcal{D} \mid \mathbf{X}_{x,j}=1\}} 1 = \sum_{\mathbf{X}\in\mathbf{K_A}} \mu(\mathbf{X})\phi_j = \phi_j\end{aligned}$$

In the first equality we used the definition of $\mathbf{P}_{x,j}$. In the second inequality we interchanged the summations. In the final equality we used $\sum_{\mathbf{X}\in\mathbf{K_A}} \mu(\mathbf{X}) = 1$. ■

The matrix $\mathbf{P}$ defined in Equation (22) is important because we can upper bound the permanent of matrix $\mathbf{A}$ in terms of entries of this matrix. We formalize this result in the following lemma.

Lemma A.9 *For matrix $\boldsymbol{A} \in \mathbb{R}_{\geq 0}^{\mathcal{D}\times\mathcal{D}}$, if $\boldsymbol{P}$ is the matrix defined in Equation (22), then*

$$\log \operatorname{perm}(\boldsymbol{A}) \leq O\left(k \log \frac{N}{k}\right) + \boldsymbol{f}(\boldsymbol{A}, \boldsymbol{P})$$

Proof We first calculate the expectation of $\log(\mu(\mathbf{X}))$ and express it in terms of matrix $\mathbf{P}$.

$$\begin{aligned}\mathbb{E}_{\mathbf{X}\sim\mu}\left[\log \mu(\mathbf{X})\right] &= \sum_{\mathbf{X}\in\mathbf{K}_r} \mu(\mathbf{X}) \log \mu(\mathbf{X}) = \sum_{\mathbf{X}\in\mathbf{K_A}} \mu(\mathbf{X}) \log \mu(\mathbf{X}) \,, \\ &= \sum_{\mathbf{X}\in\mathbf{K_A}} \mu(\mathbf{X}) \log \left(\left(\prod_{j\in[1,k]} \phi_j! \right) \left(\prod_{\{(x,j)\in\mathcal{D}\times[1,k] \mid \mathbf{X}_{x,j}=1\}} \hat{\mathbf{A}}_{x,j} \right) \left(\frac{1}{\operatorname{perm}(\mathbf{A})} \right) \right) \,, \\ &= \log \left(\prod_{j\in[1,k]} \phi_j! \right) - \log \operatorname{perm}(\mathbf{A}) + \sum_{\mathbf{X}\in\mathbf{K_A}} \mu(\mathbf{X}) \log \left(\prod_{\{(x,j)\in\mathcal{D}\times[1,k] \mid \mathbf{X}_{x,j}=1\}} \hat{\mathbf{A}}_{x,j} \right) \,.\end{aligned} \tag{24}$$

The second equality holds because the support of distribution μ is subset of $\mathbf{K}_{\mathbf{A}}$. In the third equality we used Equation (21). We now simplify the last term in the final expression from the above derivation.

$$\begin{aligned}\sum_{\mathbf{X}\in\mathbf{K}_{\mathbf{A}}}\mu(\mathbf{X})\log\left(\prod_{\{(x,j)\in\mathcal{D}\times[1,k]\ \mid\ \mathbf{X}_{x,j}=1\}}\hat{\mathbf{A}}_{x,j}\right) &= \sum_{\mathbf{X}\in\mathbf{K}_{\mathbf{A}}}\mu(\mathbf{X})\sum_{\{(x,j)\in\mathcal{D}\times[1,k]\ \mid\ \mathbf{X}_{x,j}=1\}}\log\hat{\mathbf{A}}_{x,j}\ ,\\ &= \sum_{(x,j)\in\mathcal{D}\times[1,k]}\log\hat{\mathbf{A}}_{x,j}\sum_{\{\mathbf{X}\in\mathbf{K}_{\mathbf{A}}\ \mid\ \mathbf{X}_{x,j}=1\}}\mu(\mathbf{X})\ ,\\ &= \sum_{(x,j)\in\mathcal{D}\times[1,k]}\mathbf{P}_{x,j}\log\hat{\mathbf{A}}_{x,j}\ .\end{aligned} \tag{25}$$

Combining Equation (24) and Equation (25) together we get,

$$\mathbb{E}_{\mathbf{X}\sim\mu}\left[\log\mu(\mathbf{X})\right] = \log\left(\prod_{j\in[1,k]}\phi_j!\right) - \log\mathrm{perm}(\mathbf{A}) + \sum_{(x,j)\in\mathcal{D}\times[1,k]}\mathbf{P}_{x,j}\log\hat{\mathbf{A}}_{x,j}\ . \tag{26}$$

We next define a different distribution ν on $\mathbf{K}_r$ using the following sampling procedure: For each $x\in\mathcal{D}$, pick a column $j\in[1,k]$ independently with probability $\mathbf{P}_{x,j}$. Note that this is a valid sampling procedure because for each $x\in\mathcal{D}$, $\sum_{j\in[1,k]}\mathbf{P}_{x,j}=1$. The description of distribution ν is as follows: for each $\mathbf{X}\in\mathbf{K}_r$,

$$\nu(\mathbf{X}) \stackrel{\mathrm{def}}{=} \prod_{\{(x,j)\in\mathcal{D}\times[1,k]\ \mid\ \mathbf{X}_{x,j}=1\}}\mathbf{P}_{x,j} \tag{27}$$

Remark: Note that $\sum_{X\in\mathbf{K}_r}\nu(\mathbf{X})=\prod_{x\in\mathcal{D}}(\sum_{j\in[1,k]}\mathbf{P}_{x,j})=1$ and ν is a valid distribution.

We next calculate the expectation of $\log(\nu(\mathbf{X}))$ with respect to distribution μ and express it in terms of matrix $\mathbf{P}$. Note that $\sum_{\mathbf{X}\in\mathbf{K}_r}\mu(\mathbf{X})\log\nu(\mathbf{X})=\sum_{\mathbf{X}\in\mathbf{K}_{\mathbf{A}}}\mu(\mathbf{X})\log\nu(\mathbf{X})$ because $\mu(\mathbf{X})=0$ for all $\mathbf{X}\in\mathbf{K}_r\backslash\mathbf{K}_{\mathbf{A}}$ and we get,

$$\begin{aligned}\mathbb{E}_{\mathbf{X}\sim\mu}\left[\log\nu(\mathbf{X})\right] &= \sum_{\mathbf{X}\in\mathbf{K}_{\mathbf{A}}}\mu(\mathbf{X})\log\nu(\mathbf{X}) = \sum_{\mathbf{X}\in\mathbf{K}_{\mathbf{A}}}\mu(\mathbf{X})\log\left(\prod_{\{(x,j)\in\mathcal{D}\times[1,k]\ \mid\ \mathbf{X}_{x,j}=1\}}\mathbf{P}_{x,j}\right)\\ &= \sum_{\mathbf{X}\in\mathbf{K}_{\mathbf{A}}}\mu(\mathbf{X})\sum_{\{(x,j)\in\mathcal{D}\times[1,k]\mid\mathbf{X}_{x,j}=1\}}\log\mathbf{P}_{x,j} = \sum_{(x,j)\in\mathcal{D}\times[1,k]}\log\mathbf{P}_{x,j}\sum_{\{\mathbf{X}\in\mathbf{K}_{\mathbf{A}}\mid\mathbf{X}_{x,j}=1\}}\mu(\mathbf{X})\\ &= \sum_{(x,j)\in\mathcal{D}\times[1,k]}\mathbf{P}_{x,j}\log\mathbf{P}_{x,j}\end{aligned}$$

We now calculate the KL divergence $\mathrm{KL}\left(\mu\|\nu\right)$ between distributions μ and ν.

$$\begin{aligned}\mathrm{KL}\left(\mu\|\nu\right) &\stackrel{\mathrm{def}}{=} \mathbb{E}_{\mathbf{X}\sim\mu}\left[\log\mu(\mathbf{X})\right]-\mathbb{E}_{\mathbf{X}\sim\mu}\left[\log\nu(\mathbf{X})\right]\\ &= \log\left(\prod_{j\in[1,k]}\phi_j!\right)-\log\mathrm{perm}(\mathbf{A})+\sum_{(x,j)\in\mathcal{D}\times[1,k]}\mathbf{P}_{x,j}\log\hat{\mathbf{A}}_{x,j}-\sum_{(x,j)\in\mathcal{D}\times[1,k]}\mathbf{P}_{x,j}\log\mathbf{P}_{x,j}\end{aligned}$$

As KL divergence between two distributions is always non-negative, we have $\mathrm{KL}\left(\mu\|\nu\right) \geq 0$, that further implies,

$$\begin{aligned}\log \operatorname{perm}(\mathbf{A}) &\leq \log\left(\prod_{j\in[1,k]} \phi_j!\right) + \sum_{(x,j)\in\mathcal{D}\times[1,k]} \mathbf{P}_{x,j}\log\frac{\hat{\mathbf{A}}_{x,j}}{\mathbf{P}_{x,j}}\\ &\leq \sum_{j\in[1,k]} O(\log\phi_j) + \sum_{j\in[1,k]}\phi_j\log\phi_j - \sum_{j\in[1,k]}\phi_j + \sum_{(x,j)\in\mathcal{D}\times[1,k]} \mathbf{P}_{x,j}\log\frac{\hat{\mathbf{A}}_{x,j}}{\mathbf{P}_{x,j}} \quad (28)\\ &\leq O(k\log\frac{N}{k}) + \sum_{j\in[1,k]}\phi_j\log\phi_j - \sum_{j\in[1,k]}\phi_j + \sum_{(x,j)\in\mathcal{D}\times[1,k]} \mathbf{P}_{x,j}\log\frac{\hat{\mathbf{A}}_{x,j}}{\mathbf{P}_{x,j}}\end{aligned}$$

In the second inequality we used Lemma 3.1 on each ϕ_j and further in the third inequality we used $\sum_{j\in[1,k]}\phi_j = N$ and the fact that the function $\sum_{j\in[1,k]}\log\phi_j$ is always upper bounded by $O(k\log\frac{N}{k})$. Further using the definition of $\mathbf{f}(\mathbf{A},\mathbf{P})$ (See Equation (13)), we conclude the proof. ■

We provided an upper bound to the permanent of matrix $\mathbf{A}$ and all that remains is to relate this upper bound to the scaled Sinkhorn permanent of matrix $\mathbf{A}$. Our next lemma will serve this purpose.

Lemma A.10 *For any matrix $\boldsymbol{P} \in \mathbb{R}_{\geq 0}^{\mathcal{D}\times[1,k]}$ that satisfies,*

$$\sum_{j\in[1,k]} \boldsymbol{P}_{x,j} = 1 \textit{ for all } x\in\mathcal{D} \quad \textit{and} \quad \sum_{x\in\mathcal{D}} \boldsymbol{P}_{x,j} = \phi_j \textit{ for all } j\in[1,k]\ . \quad (29)$$

there exists a doubly stochastic matrix $\boldsymbol{Q} \in \mathbb{R}_{\geq 0}^{\mathcal{D}\times\mathcal{D}}$ such that,

$$\boldsymbol{f}(\boldsymbol{A},\boldsymbol{P}) = \mathrm{U}(\boldsymbol{A},\boldsymbol{Q}) - N\ . \quad (30)$$

Proof Define matrix $\mathbf{Q} \in \mathbb{R}^{\mathcal{D}\times\mathcal{D}}$ as follows,

$$\mathbf{Q}_{x,y} \stackrel{\mathrm{def}}{=} \frac{\mathbf{P}_{x,j}}{\phi_j}$$

where in the definition above j is such that $\mathbf{A}_{:y} = \mathbf{c}_j$. Now we verify the row and column sums of matrix $\mathbf{Q}$. For each $x\in\mathcal{D}$,

$$\begin{aligned}\sum_{y\in\mathcal{D}}\mathbf{Q}_{x,y} &= \sum_{j\in[1,k]}\sum_{\{y\in\mathcal{D}\ \mid\ \mathbf{A}_{:y}=\mathbf{c}_j\}}\frac{\mathbf{P}_{x,j}}{\phi_j} = \sum_{j\in[1,k]}\frac{\mathbf{P}_{x,j}}{\phi_j}\sum_{\{y\in\mathcal{D}\ \mid\ \mathbf{A}_{:y}=\mathbf{c}_j\}} 1\\ &= \sum_{j\in[1,k]}\frac{\mathbf{P}_{x,j}}{\phi_j}\cdot\phi_j = \sum_{j\in[1,k]}\mathbf{P}_{x,j} = 1\end{aligned} \quad (31)$$

We next evaluate the column sums. For each $y\in\mathcal{D}$, let j [7] be such that $\mathbf{A}_{:y} = \mathbf{c}_j$, then

$$\sum_{x\in\mathcal{D}}\mathbf{Q}_{x,y} = \sum_{x\in\mathcal{D}}\frac{\mathbf{P}_{x,j}}{\phi_j} = \frac{1}{\phi_j}\sum_{x\in\mathcal{D}}\mathbf{P}_{x,j} = \frac{1}{\phi_j}\phi_j = 1\ . \quad (32)$$

[7] Note that j is a function of y. For convenience, in our notation we don't capture its dependence on y.

Therefore the matrix $\mathbf{Q}$ is doubly stochastic and we next relate $\mathrm{U}(\mathbf{A},\mathbf{Q})$ with $\mathbf{f}(\mathbf{A},\mathbf{P})$. Recall the definition of $\mathrm{U}(\mathbf{A},\mathbf{Q})$ (Equation (1)),

$$\mathrm{U}(\mathbf{A},\mathbf{Q}) = \sum_{(x,y)\in\mathcal{D}\times\mathcal{D}} \mathbf{Q}_{x,y}\log(\frac{\mathbf{A}_{x,y}}{\mathbf{Q}_{x,y}})\ . \tag{33}$$

We analyze the above term and express it in terms of entries of matrices $\mathbf{P}$ and $\hat{\mathbf{A}}$.

$$\begin{aligned}\sum_{(x,y)\in\mathcal{D}\times\mathcal{D}} \mathbf{Q}_{x,y}\log(\frac{\mathbf{A}_{x,y}}{\mathbf{Q}_{x,y}}) &= \sum_{x\in\mathcal{D}}\sum_{j\in[1,k]}\left[\sum_{\{y\in\mathcal{D}\ \mid\ \mathbf{A}_{:y}=\mathbf{c}_j\}} \mathbf{Q}_{x,y}\log(\frac{\mathbf{A}_{x,y}}{\mathbf{Q}_{x,y}})\right] \\ &= \sum_{x\in\mathcal{D}}\sum_{j\in[1,k]}\left[\sum_{\{y\in\mathcal{D}\ \mid\ \mathbf{A}_{:y}=\mathbf{c}_j\}} \frac{\mathbf{P}_{x,j}}{\phi_j}\log(\frac{\hat{\mathbf{A}}_{x,j}\phi_j}{\mathbf{P}_{x,j}})\right] \\ &= \sum_{x\in\mathcal{D}}\sum_{j\in[1,k]}\left[\phi_j\cdot\frac{\mathbf{P}_{x,j}}{\phi_j}\log(\frac{\hat{\mathbf{A}}_{x,j}\phi_j}{\mathbf{P}_{x,j}})\right] = \sum_{x\in\mathcal{D}}\sum_{j\in[1,k]}\left[\mathbf{P}_{x,j}\log(\frac{\hat{\mathbf{A}}_{x,j}\phi_j}{\mathbf{P}_{x,j}})\right]\end{aligned} \tag{34}$$

The first equality follows because $\mathbf{c}_j$ for all $j\in[1,k]$ are distinct. The second equality follows because for each $x\in\mathcal{D}$, consider any $y\in\mathcal{D}$ such that $\mathbf{A}_{:y}=\mathbf{c}_j$ and note that for all such y's, $\mathbf{A}_{x,y}=\hat{\mathbf{A}}_{x,j}$ and $\mathbf{Q}_{x,y}=\frac{\mathbf{P}_{x,j}}{\phi_j}$. The third equality follows because $\sum_{\{y\in\mathcal{D}\ \mid\ \mathbf{A}_{:y}=\mathbf{c}_j\}}1 = |\{y\in\mathcal{D}\mid\mathbf{A}_{:y}=\mathbf{c}_j\}|=\phi_j$.

We further simplify the final term in the above derivation.

$$\begin{aligned}\sum_{x\in\mathcal{D}}\sum_{j\in[1,k]}\left[\mathbf{P}_{x,j}\log(\frac{\hat{\mathbf{A}}_{x,j}\phi_j}{\mathbf{P}_{x,j}})\right] &= \sum_{x\in\mathcal{D}}\sum_{j\in[1,k]}\left[\mathbf{P}_{x,j}\log(\frac{\hat{\mathbf{A}}_{x,j}}{\mathbf{P}_{x,j}})\right] + \sum_{x\in\mathcal{D}}\sum_{j\in[1,k]}\mathbf{P}_{x,j}\log\phi_j \\ &= \sum_{x\in\mathcal{D}}\sum_{j\in[1,k]}\left[\mathbf{P}_{x,j}\log(\frac{\hat{\mathbf{A}}_{x,j}}{\mathbf{P}_{x,j}})\right] + \sum_{j\in[1,k]}\log\phi_j\sum_{x\in\mathcal{D}}\mathbf{P}_{x,j} \\ &= \sum_{x\in\mathcal{D}}\sum_{j\in[1,k]}\left[\mathbf{P}_{x,j}\log(\frac{\hat{\mathbf{A}}_{x,j}}{\mathbf{P}_{x,j}})\right] + \sum_{j\in[1,k]}\phi_j\log\phi_j\ .\end{aligned} \tag{35}$$

Combining Equation (34), Equation (35) and further substituting back in Equation (33) we get,

$$\begin{aligned}\mathrm{U}(\mathbf{A},\mathbf{Q}) &= \sum_{x\in\mathcal{D}}\sum_{j\in[1,k]}\left[\mathbf{P}_{x,j}\log(\frac{\hat{\mathbf{A}}_{x,j}}{\mathbf{P}_{x,j}})\right] + \sum_{j\in[1,k]}\phi_j\log\phi_j \\ &= \mathbf{f}(\mathbf{A},\mathbf{Q}) + N\ .\end{aligned} \tag{36}$$

In the final expression, we used the definition of $\mathbf{f}(\mathbf{A},\mathbf{Q})$ and combined it with $N=\sum_{j\in[1,k]}\phi_j$. ■

We are now ready to prove our main lemma of this section and is restated for convenience.

Lemma A.1 (Scaled Sinkhorn permanent approximation) *For any matrix $\boldsymbol{A}\in\mathbb{R}_{\geq 0}^{\mathcal{D}\times\mathcal{D}}$ with at most k distinct columns, the following holds,*

$$\mathrm{scaledsinkhorn}(\boldsymbol{A}) \leq \mathrm{perm}(\boldsymbol{A}) \leq \exp\left(O\left(k\log\frac{N}{k}\right)\right)\mathrm{scaledsinkhorn}(\boldsymbol{A})\ . \tag{11}$$

Proof Consider the matrix $\mathbf{P}$ defined in Equation (22). By Lemma A.8, matrix $\mathbf{P}$ satisfies the conditions of Lemma A.10; therefore, there exists a doubly stochastic matrix $\mathbf{Q} \in \mathbf{K}_{rc}$ such that $\mathbf{f}(\mathbf{A}, \mathbf{P}) = \mathrm{U}(\mathbf{A}, \mathbf{Q}) - N$. Combining it with Lemma A.9 we get $\log \mathrm{perm}(\mathbf{A}) \leq O(k \log \frac{N}{k}) + \mathrm{U}(\mathbf{A}, \mathbf{Q}) - N$, which further implies $\mathrm{perm}(\mathbf{A}) \leq \exp(O(k \log \frac{N}{k}))\mathrm{scaledsinkhorn}(\mathbf{A})$. The lower bound for the $\mathrm{perm}(\mathbf{A})$ follows from Corollary A.5 and we conclude the proof. ■

We next state another interesting property of the matrix $\mathbf{P}$ defined in Equation (22). This property will be useful for the purposes of PML (Appendix C).

Theorem A.11 *For matrix $\boldsymbol{A} \in \mathbb{R}_{\geq 0}^{\mathcal{D} \times \mathcal{D}}$, the matrix $\boldsymbol{P}$ defined in Equation (22) satisfies the following: If $x, y \in \mathcal{D}$ are such that $\boldsymbol{A}_{x.} = \boldsymbol{A}_{y.}$ then, for all $j \in [1, k]$ we have $\boldsymbol{P}_{x,j} = \boldsymbol{P}_{y,j}$.*

Proof For any $j \in [1, k]$, recall by the definitions of terms $\mathbf{P}_{x,j}$ and $\mathbf{P}_{y,j}$,

$$\begin{aligned}\mathbf{P}_{x,j} &= \sum_{\{\mathbf{X} \in \mathbf{K}_{\mathbf{A}} \mid \mathbf{X}_{x,j}=1\}} \left(\prod_{j' \in [1,k]} \phi_{j'}!\right) \left(\prod_{(z,j') \in \mathcal{D} \times [1,k]} \hat{\mathbf{A}}_{z,j'}^{\mathbf{X}_{z,j'}}\right) \left(\frac{1}{\mathrm{perm}(\mathbf{A})}\right), \\ &= \left(\prod_{j' \in [1,k]} \phi_{j'}!\right) \left(\frac{1}{\mathrm{perm}(\mathbf{A})}\right) \sum_{\{\mathbf{X} \in \mathbf{K}_{\mathbf{A}} \mid \mathbf{X}_{x,j}=1\}} \prod_{(z,j') \in \mathcal{D} \times [1,k]} \hat{\mathbf{A}}_{z,j'}^{\mathbf{X}_{z,j'}} .\end{aligned} \tag{37}$$

$$\mathbf{P}_{y,j} = \left(\prod_{j' \in [1,k]} \phi_{j'}!\right) \left(\frac{1}{\mathrm{perm}(\mathbf{A})}\right) \sum_{\{\mathbf{X}' \in \mathbf{K}_{\mathbf{A}} \mid \mathbf{X}'_{y,j}=1\}} \prod_{(z,j') \in \mathcal{D} \times [1,k]} \hat{\mathbf{A}}_{z,j'}^{\mathbf{X}'_{z,j'}} . \tag{38}$$

For any $\mathbf{Y} \in \{\mathbf{X} \in \mathbf{K}_{\mathbf{A}} \mid \mathbf{X}_{x,j} = 1\}$ we next construct a unique $\mathbf{Y}' \in \{\mathbf{X}' \in \mathbf{K}_{\mathbf{A}} \mid \mathbf{X}'_{y,j} = 1\}$ (and vice versa) such that,

$$\prod_{(z,j') \in \mathcal{D} \times [1,k]} \hat{\mathbf{A}}_{z,j'}^{\mathbf{Y}_{z,j'}} = \prod_{(z,j') \in \mathcal{D} \times [1,k]} \hat{\mathbf{A}}_{z,j'}^{\mathbf{Y}'_{z,j'}}$$

Each $\mathbf{Y} \in \mathbf{K}_{\mathbf{A}}$ corresponds to a bipartite graph where vertices correspond to set $\mathcal{D}$ on left side and $[1, k]$ on the other, such that, degree of every left vertex $x \in \mathcal{D}$ is 1 and degree of every right vertex $j \in [1, k]$ is ϕ_j.

Consider $\mathbf{Y} \in \{\mathbf{X} \in \mathbf{K}_{\mathbf{A}} \mid \mathbf{X}_{x,j} = 1\}$, we divide the analysis into the following two cases,

1. If $\mathbf{Y}_{y,j} = 1$, meaning both vertices $x, y \in \mathcal{D}$ are connected to $j \in [1, k]$ in our bipartite graph representation. Then, $\mathbf{Y}' \overset{\mathrm{def}}{=} \mathbf{Y}$.

2. If $\mathbf{Y}_{y,j} = 0$, meaning vertex x is connected to j and y to some other vertex $j' \neq j$. In this case we swap the edges, meaning we remove edges $(x, j), (y, j')$ and add $(x, j'), (y, j)$ to construct $\mathbf{Y}'$. We formally define $\mathbf{Y}'$ next,

$$\mathbf{Y}'_{z,j''} \overset{\mathrm{def}}{=} \begin{cases} 1 \text{ if } z = y \text{ and } j'' = j, \\ 0 \text{ if } z = y \text{ and } j'' = j', \\ 1 \text{ if } z = x \text{ and } j'' = j', \\ 0 \text{ if } z = x \text{ and } j'' = j, \\ \mathbf{Y}_{z,j'} \text{ otherwise} . \end{cases} \tag{39}$$

In both cases, clearly $\mathbf{Y}' \in \{\mathbf{X}' \in \mathbf{K_A} \mid \mathbf{X}'_{y,j} = 1\}$. Further, $\mathbf{A}_{x.} = \mathbf{A}_{y.}$ implies $\hat{\mathbf{A}}_{x,j'} = \hat{\mathbf{A}}_{y,j}$ for all $j' \in [1,k]$ and the following equality holds,

$$\prod_{(z,j')\in\mathcal{D}\times[1,k]} \hat{\mathbf{A}}_{z,j'}^{\mathbf{Y}_{z,j'}} = \prod_{(z,j')\in\mathcal{D}\times[1,k]} \hat{\mathbf{A}}_{z,j'}^{\mathbf{Y}'_{z,j'}}$$

The same analysis also holds when we start with a $\mathbf{Y}' \in \{\mathbf{X}' \in \mathbf{K_A} \mid \mathbf{X}'_{y,j} = 1\}$ and construct $\mathbf{Y} \in \{\mathbf{X} \in \mathbf{K_A} \mid \mathbf{X}_{x,j} = 1\}$. We have a one to one correspondence between elements $\mathbf{Y}$ and $\mathbf{Y}'$ in the sets $\{\mathbf{X} \in \mathbf{K_A} \mid \mathbf{X}_{x,j} = 1\}$ and $\{\mathbf{X}' \in \mathbf{K_A} \mid \mathbf{X}'_{y,j} = 1\}$ respectively, satisfying,

$$\prod_{(z,j')\in\mathcal{D}\times[1,k]} \hat{\mathbf{A}}_{z,j'}^{\mathbf{Y}_{z,j'}} = \prod_{(z,j')\in\mathcal{D}\times[1,k]} \hat{\mathbf{A}}_{z,j'}^{\mathbf{Y}'_{z,j'}} .$$

Therefore, $\mathbf{P}_{x,j} = \mathbf{P}_{y,j}$ and we conclude the proof. ■

A.3. Generalization to low non-negative rank matrices

Here we prove our main result for the scaled Sinkhorn permanent of low non-negative rank matrices (Theorem 4.3). To prove this result, we use the performance result of the scaled Sinkhorn permanent for non-negative matrices with k distinct columns. The following lemma relates the permanent of a matrix $\mathbf{A}$ of non-negative rank k to matrices with at most k distinct columns and will be crucial for our analysis.

Lemma A.12 (Barvinok (1996)) *Let $\boldsymbol{A} \in \mathbb{R}_{\geq 0}^{\mathcal{D}\times\mathcal{D}}$ be a matrix of non-negative rank k. If $\boldsymbol{A} \stackrel{\text{def}}{=} \sum_{j\in[k]} \boldsymbol{v}_j \boldsymbol{u}_j^\top$ for $\boldsymbol{v}_j, \boldsymbol{u}_j \in \mathbb{R}_{\geq 0}^{\mathcal{D}}$, then*

$$\mathrm{perm}(\boldsymbol{A}) = \sum_{\{\alpha\subseteq\mathbb{Z}_+^k \mid \sum_{j\in[k]}\alpha_j = N\}} \frac{1}{\prod_{j\in[k]}\alpha_j!}\mathrm{perm}(\boldsymbol{V}^\alpha)\mathrm{perm}(\boldsymbol{U}^\alpha),$$

where $\boldsymbol{V}^\alpha \stackrel{\text{def}}{=} [\underbrace{\boldsymbol{v}_1\ldots\boldsymbol{v}_1}_{\alpha_1} \mid \underbrace{\boldsymbol{v}_2\ldots\boldsymbol{v}_2}_{\alpha_2} \mid \ldots \mid \underbrace{\boldsymbol{v}_k\ldots\boldsymbol{v}_k}_{\alpha_k}]$ *and* $\boldsymbol{U}^\alpha \stackrel{\text{def}}{=} [\underbrace{\boldsymbol{u}_1\ldots\boldsymbol{u}_1}_{\alpha_1} \mid \underbrace{\boldsymbol{u}_2\ldots\boldsymbol{u}_2}_{\alpha_2} \mid \ldots \mid \underbrace{\boldsymbol{u}_k\ldots\boldsymbol{u}_k}_{\alpha_k}]$.

As the number of terms in the above summation is low, the maximizing term is a good approximation to the permanent of $\mathbf{A}$.

Corollary A.13 *Given a non-negative matrix $\boldsymbol{A} \in \mathbb{R}_{\geq 0}^{\mathcal{D}\times\mathcal{D}}$, let k denote the non-negative rank of the matrix. If $\boldsymbol{A} = \sum_{j\in[k]} \boldsymbol{v}_j \boldsymbol{u}_j^\top$ for $\boldsymbol{v}_j, \boldsymbol{u}_j \in \mathbb{R}_{\geq 0}^{\mathcal{D}}$ is any non-negative matrix factorization of $\boldsymbol{A}$, then*

$$\mathrm{perm}(\boldsymbol{A}) \leq \exp\left(O(k\log\frac{N}{k})\right) \max_{\{\alpha\subseteq\mathbb{Z}_+^k \mid \sum_{j\in[k]}\alpha_j = N\}} \frac{1}{\prod_{j\in[k]}\alpha_j!}\mathrm{perm}(\boldsymbol{V}^\alpha)\mathrm{perm}(\boldsymbol{U}^\alpha)\,. \tag{40}$$

Proof The number of feasible α's in the set $\{\alpha \subseteq \mathbb{Z}_+^k \mid \sum_{j\in[k]}\alpha_j = N\}$ is at most $\binom{N+k-1}{k-1} \in \exp\left(O(k\log\frac{N}{k})\right)$ and we conclude the proof. ■

Lemma A.14 *Let $\boldsymbol{Q}', \boldsymbol{Q}'' \in \mathbb{R}_{\geq 0}^{\mathcal{D}\times\mathcal{D}}$ be any doubly stochastic matrices. Then $\boldsymbol{Q} \stackrel{\mathrm{def}}{=} \boldsymbol{Q}'\boldsymbol{Q}''$ is a doubly stochastic matrix.*

Proof We first consider the row sums,

$$\mathbf{Q}\mathbf{1} = \mathbf{Q}'\mathbf{Q}''\mathbf{1} = \mathbf{Q}'\mathbf{1} = \mathbf{1}\ .$$

Therefore the matrix $\mathbf{Q}$ is row stochastic. In the above derivation, the second and third equalities follow because $\mathbf{Q}''$ and $\mathbf{Q}'$ are row stochastic matrices respectively. We now consider the column sums,

$$\mathbf{Q}^\top\mathbf{1} = \mathbf{Q}''^\top\mathbf{Q}'^\top\mathbf{1} = \mathbf{Q}''^\top\mathbf{1} = \mathbf{1}\ .$$

The above derivation follows because $\mathbf{Q}'$ and $\mathbf{Q}''$ are column stochastic and therefore the matrix $\mathbf{Q}$ is column stochastic. As the matrix $\mathbf{Q}$ is both row and column stochastic we conclude the proof. ■

We are now ready to prove our main result of this section and we restate it for convenience.

Theorem 4.3 (Scaled Sinkhorn permanent approximation for low non-negative rank matrices) *For any matrix $\boldsymbol{A} \in \mathbb{R}_{\geq 0}^{\mathcal{D}\times\mathcal{D}}$ with non-negative rank at most k, the following inequality holds,*

$$\mathrm{scaledsinkhorn}(\boldsymbol{A}) \leq \mathrm{perm}(\boldsymbol{A}) \leq \exp\left(O\left(k\log\frac{N}{k}\right)\right)\mathrm{scaledsinkhorn}(\boldsymbol{A})\ . \tag{3}$$

Proof Let α be the maximizer of the optimization problem 40, then

$$\mathrm{perm}(\mathbf{A}) \leq \exp\left(O(k\log\frac{N}{k})\right)\frac{1}{\prod_{j\in[k]}\alpha_j!}\mathrm{perm}(\mathbf{V}^\alpha)\mathrm{perm}(\mathbf{U}^\alpha)\ . \tag{41}$$

Recall to prove the theorem, we need to construct a doubly stochastic witness $\mathbf{Q}$ that satisfies:

$$\log\mathrm{perm}(\mathbf{A}) \leq O(k\log\frac{N}{k}) + \mathrm{U}(\mathbf{A},\mathbf{Q}) - N\ .$$

We construct such a witness $\mathbf{Q}$ from the doubly stochastic witnesses for matrices $\mathbf{V}^\alpha$ and $\mathbf{U}^\alpha$. For all $j \in [k]$ define $S_j \stackrel{\mathrm{def}}{=} \{y \in \mathcal{D} \mid \mathbf{V}^\alpha_{:y} = \mathbf{v}_j\}$, equivalently $S_j = \{y \in \mathcal{D} \mid \mathbf{U}^\alpha_{:y} = \mathbf{u}_j\}$ and note that $\alpha_j = |S_j|$. Let $\mathbf{Q}'$ and $\mathbf{Q}''$ be the doubly stochastic matrices that maximize the scaled Sinkhorn permanent for matrices $\mathbf{V}^\alpha$ and $\mathbf{U}^\alpha$ respectively. Therefore by Lemma A.1 the following inequalities hold,

$$\log\mathrm{perm}(\mathbf{V}^\alpha) \leq O(k\log\frac{N}{k}) + \mathrm{U}(\mathbf{V}^\alpha,\mathbf{Q}') - N\ , \tag{42}$$

$$\log\mathrm{perm}(\mathbf{U}^\alpha) \leq O(k\log\frac{N}{k}) + \mathrm{U}(\mathbf{U}^\alpha,\mathbf{Q}'') - N\ , \tag{43}$$

where recall $\mathrm{U}(\mathbf{V}^\alpha,\mathbf{Q}') = \sum_{x,y\in\mathcal{D}\times\mathcal{D}}\mathbf{Q}'_{x,y}\log\frac{\mathbf{V}^\alpha_{x,y}}{\mathbf{Q}'_{x,y}}$ and $\mathrm{U}(\mathbf{U}^\alpha,\mathbf{Q}'') = \sum_{x,y\in\mathcal{D}\times\mathcal{D}}\mathbf{Q}''_{x,y}\log\frac{\mathbf{U}^\alpha_{x,y}}{\mathbf{Q}''_{x,y}}$. Without loss of generality by the symmetry (with respect to columns within S_j) and concavity of the scaled Sinkhorn objective, we can assume that the maximizing matrices $\mathbf{Q}'$ and $\mathbf{Q}''$ satisfy the following: for all $x \in \mathcal{D}$ and $j \in [k]$,

$$\mathbf{Q}'_{x,y} = \mathbf{Q}'_{x,y'} \text{ and } \mathbf{Q}''_{x,y} = \mathbf{Q}''_{x,y'} \text{ for all } y, y' \in S_j \text{ and } x \in \mathcal{D}\ . \tag{44}$$

Note that the doubly stochastic matrix that we constructed for the proof of Lemma A.1 also satisfies the above collection of equalities. Now combining Equations (41) to (43) we get,

$$\begin{aligned}\log \mathrm{perm}(\mathbf{A}) &\leq O(k\log\frac{N}{k}) - \log\prod_{j\in[k]}\alpha_j! + \mathrm{U}(\mathbf{V}^{\alpha},\mathbf{Q}') - N + \mathrm{U}(\mathbf{U}^{\alpha},\mathbf{Q}'') - N\ ,\\ &\leq O(k\log\frac{N}{k}) - \sum_{j\in[k]}(\alpha_j\log\alpha_j - \alpha_j) + \mathrm{U}(\mathbf{V}^{\alpha},\mathbf{Q}') - N + \mathrm{U}(\mathbf{U}^{\alpha},\mathbf{Q}'') - N\ ,\\ &\leq O(k\log\frac{N}{k}) - \sum_{j\in[k]}\alpha_j\log\alpha_j + \mathrm{U}(\mathbf{V}^{\alpha},\mathbf{Q}') + \mathrm{U}(\mathbf{U}^{\alpha},\mathbf{Q}'') - N\ .\end{aligned} \tag{45}$$

In the second inequality we use the Stirling's approximation (Lemma 3.1) and the error term due to this approximation is upper bounded by $O(k\log\frac{N}{k})$. In the third inequality we used $\sum_{j\in[k]}\alpha_j = N$.

Let $\mathbf{Q} = \mathbf{Q}'\mathbf{Q}''^{\top}$, then by Lemma A.14 the matrix $\mathbf{Q}$ is doubly stochastic. In the remainder of the proof we show that,

$$-\sum_{j\in[k]}\alpha_j\log\alpha_j + \mathrm{U}(\mathbf{V}^{\alpha},\mathbf{Q}') + \mathrm{U}(\mathbf{U}^{\alpha},\mathbf{Q}'') \leq \mathrm{U}(\mathbf{A},\mathbf{Q})\ , \tag{46}$$

where recall $\mathrm{U}(\mathbf{A},\mathbf{Q}) = \sum_{x,y\in\mathcal{D}\times\mathcal{D}}\mathbf{Q}_{x,y}\log\frac{\mathbf{A}_{x,y}}{\mathbf{Q}_{x,y}}$. As matrix $\mathbf{Q}$ is doubly stochastic, the above inequality combined with Equation (45) concludes the proof. Therefore in the remainder we focus our attention to prove Equation (46) and we start by simplifying the above expression. Define,

$$\beta^{j}_{x,y} \stackrel{\mathrm{def}}{=} \frac{1}{\mathbf{Q}_{x,y}}\sum_{z\in S_j}\mathbf{Q}'_{x,z}\mathbf{Q}''_{y,z} \quad \text{for all } x\in\mathcal{D}, y\in\mathcal{D} \text{ and for all } j\in[k]\ . \tag{47}$$

For all $x\in\mathcal{D}$ and $y\in\mathcal{D}$ the variables defined above satisfy the following,

$$\sum_{j\in[k]}\beta^{j}_{x,y} = \frac{1}{\mathbf{Q}_{x,y}}\sum_{j\in[k]}\sum_{z\in S_j}\mathbf{Q}'_{x,z}\mathbf{Q}''_{y,z} = \frac{1}{\mathbf{Q}_{x,y}}\sum_{z\in\mathcal{D}}\mathbf{Q}'_{x,z}\mathbf{Q}''_{y,z} = \frac{1}{\mathbf{Q}_{x,y}}\mathbf{Q}_{x,y} = 1\ , \tag{48}$$

where in the third inequality we used the definition of $\mathbf{Q} = \mathbf{Q}'\mathbf{Q}''^{\top}$. We next simplify and lower bound the term $\mathrm{U}(\mathbf{A},\mathbf{Q})$ in terms of these newly defined variables.

$$\log\mathbf{A}_{x,y} = \log\sum_{j\in[k]}\mathbf{v}_j(x)\mathbf{u}_j(y) \geq \log\prod_{j\in[k]}\left(\frac{\mathbf{v}_j(x)\mathbf{u}_j(y)}{\beta^{j}_{x,y}}\right)^{\beta^{j}_{x,y}} = \sum_{j\in[k]}\beta^{j}_{x,y}\log\left(\frac{\mathbf{v}_j(x)\mathbf{u}_j(y)}{\beta^{j}_{x,y}}\right)\ , \tag{49}$$

where in the first equality we used $\mathbf{A} = \sum_{j\in[k]}\mathbf{v}_j\mathbf{u}_j^{\top}$. In the second inequality we used weighted AM-GM inequality. Now consider the term $\mathbf{Q}_{x,y}\log\mathbf{A}_{x,y}$ and substitute the above lower bound,

$$\mathbf{Q}_{x,y}\log\mathbf{A}_{x,y} \geq \mathbf{Q}_{x,y}\sum_{j\in[k]}\beta^{j}_{x,y}(\log\mathbf{v}_j(x) + \log\mathbf{u}_j(y)) - \mathbf{Q}_{x,y}\sum_{j\in[k]}\beta^{j}_{x,y}\log\beta^{j}_{x,y}\ . \tag{50}$$

Summing over all the (x, y) pairs we get,

$$\begin{aligned}\sum_{x,y\in\mathcal{D}\times\mathcal{D}} \mathbf{Q}_{x,y}\log \mathbf{A}_{x,y} \geq & \sum_{x\in\mathcal{D}}\sum_{j\in[k]} \log \mathbf{v}_j(x)\bigl(\sum_{y\in\mathcal{D}} \mathbf{Q}_{x,y}\beta^j_{x,y}\bigr) + \sum_{y\in\mathcal{D}}\sum_{j\in[k]} \log \mathbf{v}_j(y)\bigl(\sum_{x\in\mathcal{D}} \mathbf{Q}_{x,y}\beta^j_{x,y}\bigr)~, \\ & - \sum_{x,y\in\mathcal{D}\times\mathcal{D}} \mathbf{Q}_{x,y}\sum_{j\in[k]}\beta^j_{x,y}\log\beta^j_{x,y}~. \end{aligned} \tag{51}$$

In the above expression the following terms simplify,

$$\sum_{y\in\mathcal{D}} \mathbf{Q}_{x,y}\beta^j_{x,y} = \sum_{y\in\mathcal{D}} \mathbf{Q}_{x,y}\frac{1}{\mathbf{Q}_{x,y}}\sum_{z\in S_j}\mathbf{Q}'_{x,z}\mathbf{Q}''_{y,z} = \sum_{z\in S_j}\mathbf{Q}'_{x,z}\sum_{y\in\mathcal{D}}\mathbf{Q}''_{y,z} = \sum_{z\in S_j}\mathbf{Q}'_{x,z}~. \tag{52}$$

Similarly,

$$\sum_{x\in\mathcal{D}} \mathbf{Q}_{x,y}\beta^j_{x,y} = \sum_{z\in S_j}\mathbf{Q}''_{y,z}~. \tag{53}$$

Also note that,

$$\begin{aligned}\sum_{x,y\in\mathcal{D}\times\mathcal{D}} \mathbf{Q}_{x,y}\sum_{j\in[k]}\beta^j_{x,y}\log\beta^j_{x,y} &= \sum_{x,y\in\mathcal{D}\times\mathcal{D}} \mathbf{Q}_{x,y}\sum_{j\in[k]}\beta^j_{x,y}\log\frac{\beta^j_{x,y}\mathbf{Q}_{x,y}}{\mathbf{Q}_{x,y}}, \\ &= \sum_{x,y\in\mathcal{D}\times\mathcal{D}} \mathbf{Q}_{x,y}\sum_{j\in[k]}\beta^j_{x,y}\log(\beta^j_{x,y}\mathbf{Q}_{x,y}) - \sum_{x,y\in\mathcal{D}\times\mathcal{D}} \mathbf{Q}_{x,y}\sum_{j\in[k]}\beta^j_{x,y}\log\mathbf{Q}_{x,y}, \\ &= \sum_{x,y\in\mathcal{D}\times\mathcal{D}}\sum_{j\in[k]}\beta^j_{x,y}\mathbf{Q}_{x,y}\log(\beta^j_{x,y}\mathbf{Q}_{x,y}) - \sum_{x,y\in\mathcal{D}\times\mathcal{D}}\mathbf{Q}_{x,y}\log\mathbf{Q}_{x,y}~, \\ &= \sum_{x,y\in\mathcal{D}\times\mathcal{D}}\sum_{j\in[k]}\bigl(\sum_{z\in S_j}\mathbf{Q}'_{x,z}\mathbf{Q}''_{y,z}\bigr)\log\bigl(\sum_{z\in S_j}\mathbf{Q}'_{x,z}\mathbf{Q}''_{y,z}\bigr) - \sum_{x,y\in\mathcal{D}\times\mathcal{D}}\mathbf{Q}_{x,y}\log\mathbf{Q}_{x,y}~. \end{aligned} \tag{54}$$

In the third and fourth inequality we used Equation (48) and Equation (47) respectively. Substituting Equations (52) to (54) in Equation (51) we get,

$$\begin{aligned}\sum_{x,y\in\mathcal{D}\times\mathcal{D}} \mathbf{Q}_{x,y}\log \mathbf{A}_{x,y} \geq & \sum_{x\in\mathcal{D}}\sum_{j\in[k]} \log \mathbf{v}_j(x)\bigl(\sum_{z\in S_j}\mathbf{Q}'_{x,z}\bigr) + \sum_{y\in\mathcal{D}}\sum_{j\in[k]} \log \mathbf{v}_j(y)\bigl(\sum_{z\in S_j}\mathbf{Q}''_{y,z}\bigr) \\ & - \sum_{x,y\in\mathcal{D}\times\mathcal{D}}\sum_{j\in[k]}\bigl(\sum_{z\in S_j}\mathbf{Q}'_{x,z}\mathbf{Q}''_{y,z}\bigr)\log\bigl(\sum_{z\in S_j}\mathbf{Q}'_{x,z}\mathbf{Q}''_{y,z}\bigr) + \sum_{x,y\in\mathcal{D}\times\mathcal{D}}\mathbf{Q}_{x,y}\log\mathbf{Q}_{x,y}~. \end{aligned} \tag{55}$$

By rearranging terms the above expression can be equivalently written as,

$$\begin{aligned}\mathrm{U}(\mathbf{A},\mathbf{Q}) = \sum_{x,y\in\mathcal{D}\times\mathcal{D}} \mathbf{Q}_{x,y}\log\frac{\mathbf{A}_{x,y}}{\mathbf{Q}_{x,y}} \geq & \sum_{x\in\mathcal{D}}\sum_{j\in[k]} \log \mathbf{v}_j(x)\bigl(\sum_{z\in S_j}\mathbf{Q}'_{x,z}\bigr) + \sum_{y\in\mathcal{D}}\sum_{j\in[k]} \log \mathbf{u}_j(y)\bigl(\sum_{z\in S_j}\mathbf{Q}''_{y,z}\bigr) \\ & - \sum_{x,y\in\mathcal{D}\times\mathcal{D}}\sum_{j\in[k]}\bigl(\sum_{z\in S_j}\mathbf{Q}'_{x,z}\mathbf{Q}''_{y,z}\bigr)\log\bigl(\sum_{z\in S_j}\mathbf{Q}'_{x,z}\mathbf{Q}''_{y,z}\bigr)~. \end{aligned} \tag{56}$$

In the above expression we have a lower bound for the term $\mathrm{U}(\mathbf{A}, \mathbf{Q})$ and we relate it to terms $\mathrm{U}(\mathbf{V}^{\alpha}, \mathbf{Q}')$ and $\mathrm{U}(\mathbf{U}^{\alpha}, \mathbf{Q}'')$. Consider the following term,

$$\begin{aligned}\sum_{x,y \in \mathcal{D}\times\mathcal{D}} \mathbf{Q}'_{x,y} \log \mathbf{V}^{\alpha}_{x,y} &= \sum_{x \in \mathcal{D}} \sum_{j \in [k]} \sum_{y \in S_j} \mathbf{Q}'_{x,y} \log \mathbf{V}^{\alpha}_{x,y} = \sum_{x \in \mathcal{D}} \sum_{j \in [k]} \sum_{y \in S_j} \mathbf{Q}'_{x,y} \log \mathbf{v}_j(x) \ , \\ &= \sum_{x \in \mathcal{D}} \sum_{j \in [k]} \log \mathbf{v}_j(x) \big(\sum_{y \in S_j} \mathbf{Q}'_{x,y} \big) = \sum_{x \in \mathcal{D}} \sum_{j \in [k]} \log \mathbf{v}_j(x) \big(\sum_{z \in S_j} \mathbf{Q}'_{x,z} \big) \ ,\end{aligned} \tag{57}$$

In the final equality we renamed the variables and the rest of equalities are straightforward. Carrying out similar derivation we also get,

$$\sum_{x,y \in \mathcal{D}\times\mathcal{D}} \mathbf{Q}''_{x,y} \log \mathbf{U}^{\alpha}_{x,y} = \sum_{x \in \mathcal{D}} \sum_{j \in [k]} \log \mathbf{u}_j(x) \big(\sum_{y \in S_j} \mathbf{Q}''_{x,y} \big) = \sum_{y \in \mathcal{D}} \sum_{j \in [k]} \log \mathbf{u}_j(y) \big(\sum_{z \in S_j} \mathbf{Q}''_{y,z} \big) \ . \tag{58}$$

As before in the final equality we renamed variables. Substituting Equations (57) and (58) in Equation (56) we get,

$$\begin{aligned}\mathrm{U}(\mathbf{A}, \mathbf{Q}) \geq &\sum_{x,y \in \mathcal{D}\times\mathcal{D}} \mathbf{Q}'_{x,y} \log \mathbf{V}^{\alpha}_{x,y} + \sum_{x,y \in \mathcal{D}\times\mathcal{D}} \mathbf{Q}''_{x,y} \log \mathbf{U}^{\alpha}_{x,y} - \sum_{x,y \in \mathcal{D}\times\mathcal{D}} \big(\sum_{z \in S_j} \mathbf{Q}'_{x,z} \mathbf{Q}''_{y,z} \big) \log \big(\sum_{z \in S_j} \mathbf{Q}'_{x,z} \mathbf{Q}''_{y,z} \big) \\ &= \mathrm{U}(\mathbf{V}^{\alpha}, \mathbf{Q}') + \mathrm{U}(\mathbf{U}^{\alpha}, \mathbf{Q}'') + \sum_{x,y \in \mathcal{D}\times\mathcal{D}} \mathbf{Q}'_{x,y} \log \mathbf{Q}'_{x,y} + \sum_{x,y \in \mathcal{D}\times\mathcal{D}} \mathbf{Q}''_{x,y} \log \mathbf{Q}''_{x,y} \\ &\quad - \sum_{x,y \in \mathcal{D}\times\mathcal{D}} \sum_{j \in [k]} \big(\sum_{z \in S_j} \mathbf{Q}'_{x,z} \mathbf{Q}''_{y,z} \big) \log \big(\sum_{z \in S_j} \mathbf{Q}'_{x,z} \mathbf{Q}''_{y,z} \big).\end{aligned} \tag{59}$$

Therefore to prove Equation (46), all that remains is to show that,

$$\sum_{x,y \in \mathcal{D}\times\mathcal{D}} \big(\mathbf{Q}'_{x,y} \log \mathbf{Q}'_{x,y} + \mathbf{Q}''_{x,y} \log \mathbf{Q}''_{x,y} \big) - \sum_{x,y \in \mathcal{D}\times\mathcal{D}} \big(\sum_{z \in S_j} \mathbf{Q}'_{x,z} \mathbf{Q}''_{y,z} \big) \log \big(\sum_{z \in S_j} \mathbf{Q}'_{x,z} \mathbf{Q}''_{y,z} \big) \geq - \sum_{j \in [k]} \alpha_j \log \alpha_j \ . \tag{60}$$

To prove the above inequality we use the symmetry in the solutions $\mathbf{Q}'$ and $\mathbf{Q}''$. Recall from Equation (44), for all $x \in \mathcal{D}$ and $j \in [k]$ we have $\mathbf{Q}'_{x,y} = \mathbf{Q}'_{x,y'}$ and $\mathbf{Q}''_{x,y} = \mathbf{Q}''_{x,y'}$ for all $y, y' \in S_j$ and $x \in \mathcal{D}$. Define $\mathbf{R}'_{x,j} = \mathbf{Q}'_{x,y}$ and $\mathbf{R}''_{x,j} = \mathbf{Q}''_{x,y}$ for any $y \in S_j$. We next substitute these definitions and simplify terms on the left hand side of Equation (60),

$$\begin{aligned}\sum_{x,y \in \mathcal{D}\times\mathcal{D}} \mathbf{Q}'_{x,y} \log \mathbf{Q}'_{x,y} &= \sum_{x \in \mathcal{D}} \sum_{j \in [k]} \sum_{y \in S_j} \mathbf{Q}'_{x,y} \log \mathbf{Q}'_{x,y} = \sum_{x \in \mathcal{D}} \sum_{j \in [k]} \sum_{y \in S_j} \mathbf{R}'_{x,j} \log \mathbf{R}'_{x,j}, \\ &= \sum_{x \in \mathcal{D}} \sum_{j \in [k]} \alpha_j \mathbf{R}'_{x,j} \log \mathbf{R}'_{x,j} \ .\end{aligned} \tag{61}$$

In the final equality we used $|S_j| = \alpha_j$ and the rest of the equalities are straightforward. Similar argument as above also gets us,

$$\sum_{x,y \in \mathcal{D}\times\mathcal{D}} \mathbf{Q}''_{x,y} \log \mathbf{Q}''_{x,y} = \sum_{x \in \mathcal{D}} \sum_{j \in [k]} \alpha_j \mathbf{R}''_{x,j} \log \mathbf{R}''_{x,j} = \sum_{y \in \mathcal{D}} \sum_{j \in [k]} \alpha_j \mathbf{R}''_{y,j} \log \mathbf{R}''_{y,j} \ . \tag{62}$$

Note that in the final equality we renamed variables. Finally,

$$\begin{aligned}\sum_{x,y\in\mathcal{D}\times\mathcal{D}}\sum_{j\in[k]}\Big(\sum_{z\in S_j}\mathbf{Q}'_{x,z}\mathbf{Q}''_{y,z}\Big)\log\Big(\sum_{z\in S_j}\mathbf{Q}'_{x,z}\mathbf{Q}''_{y,z}\Big) &= \sum_{x,y\in\mathcal{D}\times\mathcal{D}}\sum_{j\in[k]}\alpha_j\mathbf{R}'_{x,j}\mathbf{R}''_{y,j}\log\alpha_j\mathbf{R}'_{x,j}\mathbf{R}''_{y,j}\ ,\\ &= \sum_{x,y\in\mathcal{D}\times\mathcal{D}}\sum_{j\in[k]}\alpha_j\mathbf{R}'_{x,j}\mathbf{R}''_{y,j}\big(\log\alpha_j+\log\mathbf{R}'_{x,j}+\log\mathbf{R}''_{y,j}\big)\ ,\end{aligned} \tag{63}$$

Again each of the terms in the parenthesis further simplify as follows,

$$\begin{aligned}\sum_{x,y\in\mathcal{D}\times\mathcal{D}}\sum_{j\in[k]}\alpha_j\mathbf{R}'_{x,j}\mathbf{R}''_{y,j}\log\alpha_j &= \sum_{j\in[k]}\alpha_j\log\alpha_j\sum_{x,y\in\mathcal{D}\times\mathcal{D}}\mathbf{R}'_{x,j}\mathbf{R}''_{y,j} = \sum_{j\in[k]}\alpha_j\log\alpha_j\sum_{x\in\mathcal{D}}\mathbf{R}'_{x,j}\sum_{y\in\mathcal{D}}\mathbf{R}''_{y,j},\\ &= \sum_{j\in[k]}\alpha_j\log\alpha_j\ .\end{aligned}$$

$$\sum_{x,y\in\mathcal{D}\times\mathcal{D}}\sum_{j\in[k]}\alpha_j\mathbf{R}'_{x,j}\mathbf{R}''_{y,j}\log\mathbf{R}'_{x,j} = \sum_{x\in\mathcal{D}}\sum_{j\in[k]}\alpha_j\mathbf{R}'_{x,j}\log\mathbf{R}'_{x,j}\sum_{y\in\mathcal{D}}\mathbf{R}''_{y,j} = \sum_{x\in\mathcal{D}}\sum_{j\in[k]}\alpha_j\mathbf{R}'_{x,j}\log\mathbf{R}'_{x,j}\ .$$

Similarly,

$$\sum_{x,y\in\mathcal{D}\times\mathcal{D}}\sum_{j\in[k]}\alpha_j\mathbf{R}'_{x,j}\mathbf{R}''_{y,j}\log\mathbf{R}''_{y,j} = \sum_{y\in\mathcal{D}}\sum_{j\in[k]}\alpha_j\mathbf{R}''_{y,j}\log\mathbf{R}''_{y,j}\ .$$

Substituting back all the above three expressions in Equation (63) we get,

$$\begin{aligned}\sum_{x,y\in\mathcal{D}\times\mathcal{D}}\sum_{j\in[k]}\Big(\sum_{z\in S_j}\mathbf{Q}'_{x,z}\mathbf{Q}''_{y,z}\Big)\log\Big(\sum_{z\in S_j}\mathbf{Q}'_{x,z}\mathbf{Q}''_{y,z}\Big) &= \sum_{x\in\mathcal{D}}\sum_{j\in[k]}\alpha_j\mathbf{R}'_{x,j}\log\mathbf{R}'_{x,j}+\sum_{y\in\mathcal{D}}\sum_{j\in[k]}\alpha_j\mathbf{R}''_{y,j}\log\mathbf{R}''_{y,j}\\ &\quad+\sum_{j\in[k]}\alpha_j\log\alpha_j\ .\end{aligned} \tag{64}$$

Further substituting Equations (61), (62) and (64) in the derivation below we get,

$$\sum_{x,y\in\mathcal{D}\times\mathcal{D}}(\mathbf{Q}'_{x,y}\log\mathbf{Q}'_{x,y}+\mathbf{Q}''_{x,y}\log\mathbf{Q}''_{x,y})-\sum_{x,y\in\mathcal{D}\times\mathcal{D}}\Big(\sum_{z\in S_j}\mathbf{Q}'_{x,z}\mathbf{Q}''_{y,z}\Big)\log\Big(\sum_{z\in S_j}\mathbf{Q}'_{x,z}\mathbf{Q}''_{y,z}\Big) = -\sum_{j\in[k]}\alpha_j\log\alpha_j\ .$$

Therefore the above derivation proves Equation (60) and we further substitute it in Equation (59) to get,

$$\mathrm{U}(\mathbf{A},\mathbf{Q}) \geq \mathrm{U}(\mathbf{V}^{\alpha},\mathbf{Q}')+\mathrm{U}(\mathbf{U}^{\alpha},\mathbf{Q}'')-\sum_{j\in[k]}\alpha_j\log\alpha_j\ . \tag{65}$$

The above expression combined with Equation (45) gives the following upper bound on the log of permanent,

$$\log\mathrm{perm}(\mathbf{A}) \leq O(k\log\frac{N}{k})+\mathrm{U}(\mathbf{A},\mathbf{Q})-N\ . \tag{66}$$

The above expression combined with definition of the scaled Sinkhorn permanent concludes the proof. ■

Appendix B. Lower bound for Bethe and scaled Sinkhorn permanent approximations

Here we provide the proof for Theorem 4.5 that intuitively works as follows. The performance of the Bethe permanent for the all 1's matrix is not hard to analyze and the approximation ratio is lower bounded by $\Omega(d)$ for the $d \times d$ all 1's matrix. The all 1's matrix has non-negative rank 1. We construct a matrix of non-negative rank k by creating a block diagonal matrix consisting of k blocks, where each block is a $\frac{N}{k} \times \frac{N}{k}$ dimensional all 1's matrix. The key property used in the analysis is that the value of the permanent of such a block diagonal matrix is equal to the permanent of each block raised to the power of k. Such a property also holds for the Bethe permanent and yields a gap of $\exp\left(\Omega(k \log \frac{N}{k})\right)$.

Theorem 4.5 (Lower bound for the Bethe and the scaled Sinkhorn permanents approximation) *There exists a matrix $\boldsymbol{A} \in \mathbb{R}_{\geq 0}^{\mathcal{D}\times\mathcal{D}}$ with non-negative rank k, that satisfies*

$$\mathrm{perm}(\boldsymbol{A}) \geq \exp\left(\Omega\left(k \log \frac{N}{k}\right)\right) \mathrm{bethe}(\boldsymbol{A}) \, , \tag{5}$$

which further implies,

$$\mathrm{perm}(\boldsymbol{A}) \geq \exp\left(\Omega\left(k \log \frac{N}{k}\right)\right) \mathrm{scaledsinkhorn}(\boldsymbol{A}) \, . \tag{6}$$

Proof Assume N is divisible by k. Let $\mathbf{1}$ and $\mathbf{0}$ be $\frac{N}{k} \times \frac{N}{k}$ all ones and all zeros matrices respectively. Note that $\log \mathrm{bethe}(\mathbf{1}) \leq \frac{N}{k} \log \frac{N}{k} - \frac{N}{k} + 1$ and the proof for this statement follows because $\frac{k}{N}\mathbf{1}$ is the maximizer of the optimization problem $\max_{\mathbf{Q}} \mathrm{F}(\mathbf{1}, \mathbf{Q})$ over all doubly stochastic matrices $\mathbf{Q}$. On the other hand $\log \mathrm{perm}(\mathbf{1}) = \log \frac{N}{k}! \geq \frac{N}{k} \log \frac{N}{k} - \frac{N}{k} + \Omega(\log \frac{N}{k})$, where in the last inequality we used the Stirling's approximation. Now consider the following matrix,

$$\mathbf{A} \stackrel{\mathrm{def}}{=} \begin{bmatrix} \mathbf{1} & \mathbf{0} & \ldots \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \ldots \mathbf{0} \\ \vdots & \ldots & \ddots \\ \mathbf{0} & \mathbf{0} & \ldots \mathbf{1} \end{bmatrix}$$

In the above definition $\mathbf{A}$ is a $N \times N$ matrix with k blocks, where each block is a $\frac{N}{k} \times \frac{N}{k}$ dimensional all ones matrix. For the matrix $\mathbf{A}$ we have, $\log \mathrm{perm}(\mathbf{A}) = k \cdot \log \mathrm{perm}(\mathbf{1}) \geq k\left(\frac{N}{k} \log \frac{N}{k} - \frac{N}{k} + \Omega(\log \frac{N}{k})\right)$ and $\log \mathrm{bethe}(\mathbf{A}) = k \cdot \log \mathrm{bethe}(\mathbf{1}) \leq k\left(\frac{N}{k} \log \frac{N}{k} - \frac{N}{k} + 1\right)$. Therefore $\log \mathrm{perm}(\mathbf{A}) - \log \mathrm{bethe}(\mathbf{A}) \geq \Omega(k \log \frac{N}{k})$.

The proof for the case when N is not divisible by k is similar. Here matrix $\mathbf{A}$ is the $N \times N$ block diagonal matrix where the first k blocks correspond to $\lfloor \frac{N}{k} \rfloor \times \lfloor \frac{N}{k} \rfloor$ all ones matrix and the final block corresponds to $r \times r$ all ones matrix, where $r \stackrel{\mathrm{def}}{=} N - k\lfloor \frac{N}{k} \rfloor$. For this definition of matrix $\mathbf{A}$ we have, $\log \mathrm{perm}(\mathbf{A}) = k \cdot \log \lfloor \frac{N}{k} \rfloor ! + \log r! \geq k\left(\lfloor \frac{N}{k} \rfloor \log \lfloor \frac{N}{k} \rfloor - \lfloor \frac{N}{k} \rfloor + \Omega(\log \frac{N}{k})\right) + r \log r - r + \Omega(\log r)$ and $\log \mathrm{bethe}(\mathbf{A}) = k \cdot \log \mathrm{bethe}(\mathbf{1}) \leq k\left(\lfloor \frac{N}{k} \rfloor \log \lfloor \frac{N}{k} \rfloor - \lfloor \frac{N}{k} \rfloor + 1\right) + r \log r - r + 1$. Therefore $\log \mathrm{perm}(\mathbf{A}) - \log \mathrm{bethe}(\mathbf{A}) \geq \Omega(k \log \frac{N}{k})$. The first condition of the theorem follows by taking exponential on both sides of the previous inequality.

The second inequality in the theorem follows by using $\mathrm{bethe}(\mathbf{A}) \geq \mathrm{scaledsinkhorn}(\mathbf{A})$ (See Corollary A.5). As the matrix $\mathbf{A}$ constructed here is of non-negative rank k, we conclude the proof. ■

Appendix C. Improved approximation to profile maximum likelihood

In this section, we provide an efficient algorithm to compute an $\exp\left(-O(\sqrt{n}\log n)\right)$-approximate PML distribution. We first introduce the setup and some new notation. For convenience, we also recall some definitions from Section 3.

We are given access to n independent samples from a hidden distribution $\mathbf{p} \in \Delta^{\mathcal{D}}$ supported on domain $\mathcal{D}$. Let x^n be this length n sequence and $\phi = \Phi(x^n)$ be its corresponding profile. Let $\mathbf{f}(x^n, y)$ be the frequency for domain element $y \in \mathcal{D}$ in sequence x^n. Let k be the number of non-zero distinct frequencies and we use $\{\mathbf{m}_1, \dots \mathbf{m}_j, \dots \mathbf{m}_k\}$ to denote these distinct frequencies. Note that the number of non-zero distinct frequencies k is always upper bounded by $\sqrt{n}$. For $j \in [1,k]$, we define $\phi_j \overset{\text{def}}{=} |\{y \in \mathcal{D} \mid \mathbf{f}(x^n, y) = \mathbf{m}_j\}|$. Let $\mathbf{p}_{\text{pml}}$ be the PML distribution with respect to profile ϕ and is formally defined as follows,

$$\mathbf{p}_{\text{pml}} \in \arg\max_{\mathbf{p} \in \Delta^{\mathcal{D}}} \mathbb{P}(\mathbf{p}, \phi)\ .$$

Let x^n be a sequence such that $\Phi(x^n) = \phi$. We define a *profile probability matrix* $\mathbf{A}^{\mathbf{p},\phi}$ with respect to sequence x^n (therefore profile ϕ) and distribution $\mathbf{p}$ as follows,

$$\mathbf{A}^{\mathbf{p},\phi}_{z,y} \overset{\text{def}}{=} \mathbf{p}_z^{\mathbf{f}_y} \text{ for all } z, y \in \mathcal{D}, \tag{67}$$

where $\mathbf{f}_y \overset{\text{def}}{=} \mathbf{f}(x^n, y)$ is the frequency of domain element $y \in \mathcal{D}$ in sequence x^n and recall $\Phi(x^n) = \phi$. We are interested in the permanent of the matrix $\mathbf{A}^{\mathbf{p},\phi}$, and note that the $\text{perm}(\mathbf{A}^{\mathbf{p},\phi})$ is invariant under the choice of sequences x^n that satisfy $\Phi(x^n) = \phi$. Therefore we index the matrix $\mathbf{A}^{\mathbf{p},\phi}$ with profile ϕ rather than sequence x^n itself. The number of distinct columns in $\mathbf{A}^{\mathbf{p},\phi}$ is equal to number of distinct observed frequencies plus one (for the unseen), i.e. $k+1$.

The probability of a profile $\phi \in \Phi^n$ with respect to distribution $\mathbf{p}$ (from Equation 20 in Orlitsky et al. (2003), Equation 15 in Pavlichin et al. (2017)) in terms of permanent of matrix $\mathbf{A}^{\mathbf{p},\phi}$ is given below:

$$\mathbb{P}(\mathbf{p}, \phi) = C_\phi \cdot \left(\prod_{j \in [0,k]} \frac{1}{\phi_j !}\right) \cdot \text{perm}(\mathbf{A}^{\mathbf{p},\phi}) \tag{68}$$

where $C_\phi \overset{\text{def}}{=} \frac{n!}{\prod_{j \in [1,k]} (\mathbf{m}_j!)^{\phi_j}}$ and ϕ_0 is the number of unseen domain elements and note that it is not part of the profile. Given a distribution $\mathbf{p}$ we know its domain $\mathcal{D}$ therefore the unseen domain elements. Also, note that C_ϕ is independent of the term ϕ_0, therefore it depends just on the profile ϕ and not on the underlying distribution $\mathbf{p}$.

We now provide the motivation behind the techniques used in this section. Recall that the goal of this section is to compute an approximate PML distribution and we wish to do this using the results from the previous section. A first attempt would be to use the scaled Sinkhorn (or the Bethe) permanent as a proxy for the term $\text{perm}(\mathbf{A}^{\mathbf{p},\phi})$ in Equation (68) and solve the following optimization problem:

$$\max_{\mathbf{p} \in \Delta^{\mathcal{D}}} C_\phi \cdot \left(\prod_{j \in [0,k]} \frac{1}{\phi_j !}\right) \cdot \text{scaledsinkhorn}(\mathbf{A}^{\mathbf{p},\phi})\ .$$

The above optimization problem is indeed a good proxy for the PML objective and recall the above optimization problem is equivalent to the following:

$$\max_{\mathbf{p}\in\Delta^{\mathcal{D}}} C_{\phi}\cdot\left(\prod_{j\in[0,k]}\frac{1}{\phi_j!}\right)\cdot\max_{\mathbf{Q}\in\mathbf{Z}_{rc}}\exp\left(\mathrm{U}(\mathbf{A}^{\mathbf{p},\phi},\mathbf{Q})\right)~.$$

Taking log and ignoring the constants we get the following equivalent optimization problem,

$$\max_{\mathbf{p}\in\Delta^{\mathcal{D}}}\max_{\mathbf{Q}\in\mathbf{Z}_{rc}}\left(\log\frac{1}{\phi_0!}+\mathrm{U}(\mathbf{A}^{\mathbf{p},\phi},\mathbf{Q})\right)$$

Interestingly, the function $\mathrm{U}(\mathbf{A}^{\mathbf{p},\phi},\mathbf{Q})$, is concave with respect to $\mathbf{p}$ for a fixed $\mathbf{Q}$ and concave with respect to $\mathbf{Q}$ for a fixed $\mathbf{p}$ (See Vontobel (2014)). However, unfortunately the function $\mathrm{U}(\mathbf{A}^{\mathbf{p},\phi},\mathbf{Q})$ in general is not a concave function with respect to $\mathbf{p}$ and $\mathbf{Q}$ simultaneously Vontobel (2014) and we do not know how to solve the above optimization problem. Vontobel Vontobel (2014) proposed an alternating maximization algorithm to solve the above optimization problem, and studied its implementation and convergence to a stationary point; see Vontobel (2014) for empirical performance of this approach. Using the Bethe permanent as a proxy in the above optimization problem has similar issues; see Vontobel (2012, 2014) for further details.

To address the above issue we use the idea of probability discretization from Charikar et al. (2019a), meaning we assume distribution takes all its probability values from some fixed predefined set. We use this idea in a different way than Charikar et al. (2019a) and further exploit the structure of optimal solution $\mathbf{Q}$ to write a convex optimization problem that approximates the PML objective. The solution of this convex optimization problem returns a fractional representation of the distribution that we later round to return the approximate PML distribution with desired guarantees. Surprisingly, the final convex optimization problem we write is exactly same as the one in Charikar et al. (2019a) and our work gives a better analysis of the same convex program by a completely different approach.

The rest of this section is organized as follows. In Appendix C.1, we study the probability discretization. In the same section, we also study the application of results from Appendix A for approximating the permanent of profile probability matrix ($\mathbf{A}^{\mathbf{p},\phi}$). We further provide the convex optimization problem at the end of this section that can be solved efficiently and returns a fractional representation of the approximate PML distribution. In Appendix C.2, we provide the rounding algorithm that returns our final approximate PML distribution. Till this point, all our results are independent of the choice of the probability discretization set. Later in Appendix C.3, we choose an appropriate probability discretization set and further combine analysis from all the previous sections. In this section, we state and analyze our final algorithm that returns a $\exp\left(-O(\sqrt{n}\log n)\right)$-approximate PML distribution. Note that our rounding algorithm is technical and for the continuity of reading we defer all the proofs for results in Appendix C.2 to Appendix C.4.

C.1. Probability discretization

Here we study the idea of probability discretization that is also used in Charikar et al. (2019a). We combine this with other ideas from Appendix A to provide a convex program that approximates the PML objective.

Let $\mathbf{R}\subseteq[0,1]_{\mathbb{R}}$ be some discretization of the probability space and in this section we consider distributions that take all its probability values in set $\mathbf{R}$. All results in this section hold for finite set $\mathbf{R}$ and we specify the exact definition of $\mathbf{R}$ in Appendix C.3.

The discretization introduces a technicality of probability values not summing up to one and we redefine pseudo-distribution and discrete pseudo-distribution from Charikar et al. (2019a) to deal with these.

Definition C.1 (Pseudo-distribution) *$\boldsymbol{q} \in [0,1]^{\mathcal{D}}_{\mathbb{R}}$ is a* pseudo-distribution *if $\|\boldsymbol{q}\|_1 \leq 1$ and a* discrete pseudo-distribution *with respect to $\boldsymbol{R}$ if all its entries are in $\boldsymbol{R}$ as well. We use $\Delta^{\mathcal{D}}_{pseudo}$ and $\Delta^{\mathcal{D}}_{\boldsymbol{R}}$ to denote the set of all such pseudo-distributions and discrete pseudo-distributions with respect to $\boldsymbol{R}$ respectively.*

We extend and use the following definition for $\mathbb{P}(\mathbf{v}, y^n)$ for any vector $\mathbf{v} \in \mathbb{R}^{\mathcal{D}}_{\geq 0}$ and therefore for pseudo-distributions as well,

$$\mathbb{P}(\mathbf{v}, y^n) \stackrel{\text{def}}{=} \prod_{x \in \mathcal{D}} \mathbf{v}_x^{\mathbf{f}(y^n,x)} .$$

Further, for any probability terms defined involving $\mathbf{p}$, we define those terms for any vector $\mathbf{v} \in \mathbb{R}^{\mathcal{D}}_{\geq 0}$ just by replacing $\mathbf{p}_x$ by $\mathbf{v}_x$ everywhere. For convenience we refer to $\mathbb{P}(\mathbf{q}, \phi)$ for any pseudo-distribution $\mathbf{q}$ as the "probability" of profile ϕ with respect to $\mathbf{q}$.

For a scalar c and set $\mathbf{S}$, define $\lfloor c \rfloor_{\mathbf{S}}$ and $\lceil c \rceil_{\mathbf{S}}$ as follows:

$$\lfloor c \rfloor_{\mathbf{S}} \stackrel{\text{def}}{=} \sup_{s \in \mathbf{S} : s \leq c} s \quad \text{and} \quad \lceil c \rceil_{\mathbf{S}} \stackrel{\text{def}}{=} \inf_{s \in \mathbf{S} : s \geq c} s$$

Definition C.2 (Discrete pseudo-distribution) *For any distribution $\boldsymbol{p} \in \Delta^{\mathcal{D}}$, its* discrete pseudo-distribution *$\boldsymbol{q} = \mathrm{disc}(\boldsymbol{p}) \in \Delta^{\mathcal{D}}_{\boldsymbol{R}}$ with respect to $\boldsymbol{R}$ is defined as:*

$$\boldsymbol{q}_x \stackrel{\text{def}}{=} \lfloor \boldsymbol{p}_x \rfloor_{\boldsymbol{R}} \quad \forall x \in \mathcal{D}$$

We now define some additional definitions and notation that will help us lower and upper bound the permanent of profile probability matrix by a convex optimization problem.

- Let $\ell \stackrel{\text{def}}{=} |\mathbf{R}|$ be the cardinality of set $\mathbf{R}$ and $\mathbf{r}_i$ be the i'th element of set $\mathbf{R}$.

- For any discrete pseudo-distribution $\mathbf{q}$ with respect to $\mathbf{R}$, that is $\mathbf{q} \in \Delta^{\mathcal{D}}_{\mathbf{R}}$, we let $\ell^{\mathbf{q}}_i \stackrel{\text{def}}{=} |\{y \in \mathcal{D} \mid \mathbf{q}_y = \mathbf{r}_i\}|$, be the number of domain elements with probability $\mathbf{r}_i$.

- Let $\mathbf{Z}^{\mathbf{q},\phi}_{\mathbf{R}} \subseteq \mathbb{R}^{\ell \times (k+1)}_{\geq 0}$ be the set of non-negative matrices such that, for any $\mathbf{S} \in \mathbf{Z}^{\mathbf{q},\phi}_{\mathbf{R}}$ the following holds:

$$\sum_{j \in [0,k]} \mathbf{S}_{i,j} = \ell^{\mathbf{q}}_i \text{ for all } i \in [1,\ell] \quad \text{and} \quad \sum_{i \in [1,\ell]} \mathbf{S}_{i,j} = \phi_j \text{ for all } j \in [0,k] , \qquad (69)$$

  where ϕ_0[8] is the number of unseen domain elements and we use $\mathbf{m}_0 \stackrel{\text{def}}{=} 0$ to denote the corresponding frequency element.

[8] ϕ_0 is not part of the profile and is not given to us. Later in this section, we get rid of this dependency on ϕ_0.

- For any $\mathbf{S} \in \mathbb{R}_{\geq 0}^{\ell \times (k+1)}$ define,

$$\mathbf{h}(\mathbf{S}) = \sum_{i\in[1,\ell]} \sum_{j\in[0,k]} \left[\mathbf{S}_{i,j} \log(\frac{\mathbf{r}_i^{\mathbf{m}_j}}{\mathbf{S}_{i,j}})\right] + \sum_{i\in[1,\ell]} \left(\sum_{j\in[0,k]} \mathbf{S}_{i,j}\right) \log\left(\sum_{j\in[0,k]} \mathbf{S}_{i,j}\right) + \sum_{j\in[0,k]} \phi_j \log \phi_j - \sum_{j\in[0,k]} \phi_j \ . \tag{70}$$

- Throughout this section, for convenience unless stated otherwise we abuse notation and use $\mathbf{A}$ to denote the matrix $\mathbf{A}^{\mathbf{q},\phi}$. The underlying pseudo-distribution $\mathbf{q}$ and profile ϕ with respect to matrix $\mathbf{A}$ will be clear from the context.

The first half of this section is dedicated to bound the $\mathrm{perm}(\mathbf{A})$ in terms of function $\mathbf{h}(\mathbf{S})$. For any fixed discrete pseudo-distribution $\mathbf{q}$ and profile ϕ, we will show that,

$$\max_{\mathbf{S}\in\mathbf{Z}_{\mathbf{R}}^{\mathbf{q},\phi}} \mathbf{h}(\mathbf{S}) \leq \log \mathrm{perm}(\mathbf{A}^{\mathbf{q},\phi}) \leq O(k \log \frac{N}{k}) + \max_{\mathbf{S}\in\mathbf{Z}_{\mathbf{R}}^{\mathbf{q},\phi}} \mathbf{h}(\mathbf{S}) \ .$$

Later in the second half, we use the above inequality to maximize over all the discrete pseudo-distributions to find the approximate PML distribution and the summary of which is stated later. We start by showing the lower bound first and later in Theorem C.4 we prove the upper bound.

Theorem C.3 *For any discrete pseudo-distribution* $\boldsymbol{q}$ *with respect to* $\boldsymbol{R}$ *and profile* ϕ*, let* $\boldsymbol{A}$ *be the matrix defined (with respect to* $\boldsymbol{q}$ *and* ϕ*) in Equation* (67)*, then the following holds,*

$$\log \mathrm{perm}(\boldsymbol{A}) \geq \max_{\boldsymbol{S}\in\boldsymbol{Z}_{\boldsymbol{R}}^{\boldsymbol{q},\phi}} \boldsymbol{h}(\boldsymbol{S}) \ . \tag{71}$$

Proof For any matrix $\mathbf{S} \in \mathbf{Z}_{\mathbf{R}}^{\mathbf{q},\phi}$, define matrix $\mathbf{Q} \in \mathbb{R}^{\mathcal{D}\times\mathcal{D}}$ as follows,

$$\mathbf{Q}_{x,y} \stackrel{\mathrm{def}}{=} \frac{\mathbf{S}_{i,j}}{\ell_i^{\mathbf{q}} \phi_j}$$

where in the definition above i and j are such that $\mathbf{q}_x = \mathbf{r}_i$ and $\mathbf{f}_y = \mathbf{m}_j$. We now establish that matrix $\mathbf{Q}$ is doubly stochastic. For each $x \in \mathcal{D}$, let i be such that $\mathbf{q}_x = \mathbf{r}_i$, then

$$\begin{aligned}\sum_{y\in\mathcal{D}} \mathbf{Q}_{x,y} &= \sum_{j\in[0,k]} \sum_{\{y\in\mathcal{D} \mid \mathbf{f}_y=\mathbf{m}_j\}} \frac{\mathbf{S}_{i,j}}{\ell_i^{\mathbf{q}} \phi_j} = \sum_{j\in[0,k]} \frac{\mathbf{S}_{i,j}}{\ell_i^{\mathbf{q}} \phi_j} \sum_{\{y\in\mathcal{D} \mid \mathbf{f}_y=\mathbf{m}_j\}} 1 \\ &= \sum_{j\in[0,k]} \frac{\mathbf{S}_{x,\mathbf{m}_j}}{\ell_i^{\mathbf{q}} \phi_j} \cdot \phi_j = \frac{1}{\ell_i^{\mathbf{q}}} \sum_{j\in[0,k]} \mathbf{S}_{x,\mathbf{m}_j} = 1 \ .\end{aligned} \tag{72}$$

For each $y \in \mathcal{D}$, let j be such that $\mathbf{f}_y = \mathbf{m}_j$, then

$$\begin{aligned}\sum_{x\in\mathcal{D}} \mathbf{Q}_{x,y} &= \sum_{i\in[1,\ell]} \sum_{\{x\in\mathcal{D} \mid \mathbf{q}_x=\mathbf{r}_i\}} \frac{\mathbf{S}_{i,j}}{\ell_i^{\mathbf{q}} \phi_j} = \sum_{i\in[1,\ell]} \frac{\mathbf{S}_{i,j}}{\ell_i^{\mathbf{q}} \phi_j} \sum_{\{x\in\mathcal{D} \mid \mathbf{q}_x=\mathbf{r}_i\}} 1 \\ &= \sum_{i\in[1,\ell]} \frac{\mathbf{S}_{x,\mathbf{m}_j}}{\ell_i^{\mathbf{q}} \phi_j} \cdot \ell_i^{\mathbf{q}} = \frac{1}{\phi_j} \sum_{i\in[1,\ell]} \mathbf{S}_{x,\mathbf{m}_j} = 1 \ .\end{aligned} \tag{73}$$

Since matrix $\mathbf{Q}$ is doubly stochastic, by the definition of the scaled Sinkhorn permanent and Corollary A.5 we have $\log \mathrm{perm}(\mathbf{A}) \geq \mathrm{U}(\mathbf{A},\mathbf{Q}) - N$. To conclude the proof we show that $\mathrm{U}(\mathbf{A},\mathbf{Q}) - N = \mathbf{h}(\mathbf{S})$.

$$\begin{aligned}\mathrm{U}(\mathbf{A},\mathbf{Q}) &= \sum_{(x,y)\in\mathcal{D}\times\mathcal{D}} \mathbf{Q}_{x,y}\log(\frac{\mathbf{A}_{x,y}}{\mathbf{Q}_{x,y}}) = \sum_{i\in[1,\ell]}\sum_{j\in[0,k]} \ell_i^{\mathbf{q}}\phi_j \cdot \frac{\mathbf{S}_{i,j}}{\ell_i^{\mathbf{q}}\phi_j}\log(\frac{\mathbf{r}_i^{\mathbf{m}_j}\ell_i^{\mathbf{q}}\phi_j}{\mathbf{S}_{i,j}}) \\ &= \sum_{i\in[1,\ell]}\sum_{j\in[0,k]} \mathbf{S}_{i,j}\log(\frac{\mathbf{r}_i^{\mathbf{m}_j}\ell_i^{\mathbf{q}}\phi_j}{\mathbf{S}_{i,j}}) \; .\end{aligned} \tag{74}$$

We consider the final expression above and simplify it. First note that,

$$\sum_{i\in[1,\ell]}\sum_{j\in[0,k]} \mathbf{S}_{i,j}\log \ell_i^{\mathbf{q}} = \sum_{i\in[1,\ell]} \log \ell_i^{\mathbf{q}} \sum_{j\in[0,k]} \mathbf{S}_{i,j} = \sum_{i\in[1,\ell]} \ell_i^{\mathbf{q}} \log \ell_i^{\mathbf{q}} \; .$$

Similarly,

$$\sum_{i\in[1,\ell]}\sum_{j\in[0,k]} \mathbf{S}_{i,j}\log \phi_j = \sum_{j\in[0,k]} \log \phi_j \sum_{i\in[1,\ell]} \mathbf{S}_{i,j} = \sum_{j\in[0,k]} \phi_j \log \phi_j \; .$$

Using the above two expressions, the final expression of Equation (74) can be equivalently written as,

$$\sum_{i\in[1,\ell]}\sum_{j\in[0,k]} \mathbf{S}_{i,j}\log(\frac{\mathbf{r}_i^{\mathbf{m}_j}\ell_i^{\mathbf{q}}\phi_j}{\mathbf{S}_{i,j}}) = \sum_{i\in[1,\ell]}\sum_{j\in[0,k]} \left[\mathbf{S}_{i,j}\log(\frac{\mathbf{r}_i^{\mathbf{m}_j}}{\mathbf{S}_{i,j}})\right] + \sum_{i\in[1,\ell]} \ell_i^{\mathbf{q}}\log \ell_i^{\mathbf{q}} + \sum_{j\in[0,k]} \phi_j \log \phi_j \; . \tag{75}$$

Combining Equation (74), Equation (75) and substituting $N = \sum_{j\in[0,k]}\phi_j$, we get:

$$\mathrm{U}(\mathbf{A},\mathbf{Q}) - N = \sum_{i\in[1,\ell]}\sum_{j\in[0,k]} \mathbf{S}_{i,j}\log(\frac{\mathbf{r}_i^{\mathbf{m}_j}}{\mathbf{S}_{i,j}}) + \sum_{i\in[1,\ell]} \ell_i^{\mathbf{q}}\log \ell_i^{\mathbf{q}} + \sum_{j\in[0,k]} \phi_j \log \phi_j - \sum_{j\in[0,k]} \phi_j = \mathbf{h}(\mathbf{S}) \; .$$

In the above equality we used $\sum_{j\in[0,k]} \mathbf{S}_{i,j} = \ell_i^{\mathbf{q}}$ for all $i \in [1,\ell]$ and for any $\mathbf{S} \in \mathbf{Z}_{\mathbf{R}}^{\mathbf{q},\phi}$. Combining the above inequality with $\log \mathrm{perm}(\mathbf{A}) \geq \mathrm{U}(\mathbf{A},\mathbf{Q}) - N$ we get,

$$\log \mathrm{perm}(\mathbf{A}) \geq \mathbf{h}(\mathbf{S}) \; .$$

The above inequality holds for any $\mathbf{S} \in \mathbf{Z}_{\mathbf{R}}^{\mathbf{q},\phi}$ (and therefore holds for the maximizer as well) and we conclude the proof. ■

We next give an upper bound for the log of permanent of $\mathbf{A}$ in terms of $\mathbf{h}(\mathbf{S})$.

Theorem C.4 *For any discrete pseudo-distribution* $\boldsymbol{q}$ *with respect to* $\boldsymbol{R}$ *and profile* ϕ*, let* $\boldsymbol{A}$ *be the matrix defined (with respect to* $\boldsymbol{q}$ *and* ϕ*) in Equation* (67)*. Then,*

$$\log \mathrm{perm}(\boldsymbol{A}) \leq O(k \log \frac{N}{k}) + \max_{\boldsymbol{S}\in \boldsymbol{Z}_{\boldsymbol{R}}^{\boldsymbol{q},\phi}} \boldsymbol{h}(\boldsymbol{S}) \; .$$

Proof Here we construct a particular matrix $\mathbf{S} \in \mathbf{Z}_{\mathbf{R}}^{\mathbf{q},\phi}$ such that $\log \mathrm{perm}(\mathbf{A}) \leq O(k \log \frac{N}{k}) + \mathbf{h}(\mathbf{S})$, which immediately implies the theorem. Recall by Lemma A.8 and A.9, there exists a matrix $\mathbf{P} \in \mathbb{R}_{\geq 0}^{\mathcal{D}\times(k+1)}$ such that, $\sum_{j\in[0,k]} \mathbf{P}_{x,j} = 1$ for all $x \in \mathcal{D}$ and $\sum_{x\in\mathcal{D}} \mathbf{P}_{x,j} = \phi_j$ for all $j \in [0, k]$, and satisfies $\log \mathrm{perm}(\mathbf{A}) \leq O(k \log \frac{N}{k}) + \mathbf{f}(\mathbf{A}, \mathbf{P})$ [9]. Further using the definition of $\mathbf{f}(\mathbf{A}, \mathbf{P})$ we get,

$$\log \mathrm{perm}(\mathbf{A}) \leq O(k \log \frac{N}{k}) + \sum_{j\in[0,k]} \phi_j \log \phi_j - \sum_{j\in[0,k]} \phi_j + \sum_{(x,j)\in\mathcal{D}\times[0,k]} \mathbf{P}_{x,j} \log \frac{\hat{\mathbf{A}}_{x,j}}{\mathbf{P}_{x,j}} , \quad (76)$$

where for the matrix $\mathbf{A}$ defined (with respect to $\mathbf{q}$ and ϕ) in Equation (67), we have,

$$\hat{\mathbf{A}}_{x,j} = \mathbf{q}_x^{\mathbf{m}_j} .$$

We now define the matrix $\mathbf{S}$ that satisfies the conditions of the lemma.

$$\mathbf{S}_{i,j} \stackrel{\mathrm{def}}{=} \sum_{\{x\in\mathcal{D} \mid \mathbf{q}_x=\mathbf{r}_i\}} \mathbf{P}_{x,j}$$

By Theorem A.11, for any fixed $j \in [0, k]$, all $x \in \mathcal{D}$ such that $\mathbf{q}_x = \mathbf{r}_i$, share the same probability value $\mathbf{P}_{x,j}$ and we use the notation $\mathbf{P}_{i,j}$ to denote this value. Using this definition, we have:

$$\mathbf{S}_{i,j} = \ell_i^{\mathbf{q}} \mathbf{P}_{i,j} \quad (77)$$

Further note that for any $i \in [1, \ell]$, if $x \in \mathcal{D}$ is any element such that $\mathbf{q}_x = \mathbf{r}_i$, then

$$\sum_{j\in[0,k]} \mathbf{P}_{i,j} = \sum_{j\in[0,k]} \mathbf{P}_{x,j} = 1$$

We wish to show that $\mathbf{S} \in \mathbf{Z}_{\mathbf{R}}^{\mathbf{q},\phi}$. We first analyze the row sum constraint. For each $i \in [1, \ell]$,

$$\sum_{j\in[0,k]} \mathbf{S}_{i,j} = \sum_{j\in[0,k]} \ell_i^{\mathbf{q}} \mathbf{P}_{i,j} = \ell_i^{\mathbf{q}}$$

We now analyze the column constraint. For each $j \in [0, k]$,

$$\sum_{i\in[1,\ell]} \mathbf{S}_{i,j} = \sum_{i\in[1,\ell]} \sum_{\{x\in\mathcal{D} \mid \mathbf{q}_x=\mathbf{r}_i\}} \mathbf{P}_{x,j} = \sum_{x\in\mathcal{D}} \mathbf{P}_{x,j} = \phi_j$$

In the remainder of the proof we show that the matrix $\mathbf{S}$ defined earlier satisfies $\log \mathrm{perm}(\mathbf{A}) \leq O(k \log \frac{N}{k}) + \mathbf{h}(\mathbf{S})$. We start by simplifying the term $\sum_{(x,j)\in\mathcal{D}\times[0,k]} \mathbf{P}_{x,j} \log \frac{\hat{\mathbf{A}}_{x,j}}{\mathbf{P}_{x,j}}$ in Equation (76),

$$\begin{aligned}
\sum_{(x,j)\in\mathcal{D}\times[0,k]} \mathbf{P}_{x,j} \log \frac{\hat{\mathbf{A}}_{x,j}}{\mathbf{P}_{x,j}} &= \sum_{j\in[0,k]} \sum_{i\in[1,\ell]} \sum_{\{x\in\mathcal{D} \mid \mathbf{q}_x=\mathbf{r}_i\}} \mathbf{P}_{x,j} \log \frac{\hat{\mathbf{A}}_{x,j}}{\mathbf{P}_{x,j}} = \sum_{j\in[0,k]} \sum_{i\in[1,\ell]} \sum_{\{x\in\mathcal{D} \mid \mathbf{q}_x=\mathbf{r}_i\}} \mathbf{P}_{i,j} \log \frac{\mathbf{r}_i^{\mathbf{m}_j}}{\mathbf{P}_{i,j}} \\
&= \sum_{j\in[0,k]} \sum_{i\in[1,\ell]} \ell_i^{\mathbf{q}} \mathbf{P}_{i,j} \log \frac{\mathbf{r}_i^{\mathbf{m}_j}}{\mathbf{P}_{i,j}} = \sum_{i\in[1,\ell]} \sum_{j\in[0,k]} \mathbf{S}_{i,j} \log \frac{\mathbf{r}_i^{\mathbf{m}_j} \ell_i^{\mathbf{q}}}{\mathbf{S}_{i,j}} \\
&= \sum_{i\in[1,\ell]} \sum_{j\in[0,k]} \mathbf{S}_{i,j} \log \frac{\mathbf{r}_i^{\mathbf{m}_j}}{\mathbf{S}_{i,j}} + \sum_{i\in[1,\ell]} \ell_i^{\mathbf{q}} \log \ell_i^{\mathbf{q}}
\end{aligned} \quad (78)$$

[9] The inequality holds because matrix $\mathbf{A}$ has $k + 1$ distinct columns and $O((k + 1) \log \frac{N}{k+1})$ is asymptotically same as $O(k \log \frac{N}{k})$.

In the second equality, we used $\hat{\mathbf{A}}_{x,j} = \mathbf{r}_i^{\mathbf{m}_j}$ and further by the definition of $\mathbf{P}_{i,j}$ we have $\mathbf{P}_{x,j} = \mathbf{P}_{i,j}$ for all $x \in \mathcal{D}$ that satisfy $\mathbf{q}_x = \mathbf{r}_i$. In the third equality, we used $\sum_{\{x \in \mathcal{D} \mid \mathbf{q}_x = \mathbf{r}_i\}} 1 = \ell_i^{\mathbf{q}}$. In the fourth equality we used Equation (77). In the final equality, we used $\sum_{i \in [1,\ell]} \sum_{j \in [0,k]} \mathbf{S}_{i,j} \log \frac{\mathbf{r}_i^{\mathbf{m}_j} \ell_i^{\mathbf{q}}}{\mathbf{S}_{i,j}} = \sum_{i \in [1,\ell]} \sum_{j \in [0,k]} \mathbf{S}_{i,j} \log \frac{\mathbf{r}_i^{\mathbf{m}_j}}{\mathbf{S}_{i,j}} + \sum_{i \in [1,\ell]} \sum_{j \in [0,k]} \mathbf{S}_{i,j} \log \ell_i^{\mathbf{q}}$ and the final term further simplifies to the following, $\sum_{i \in [1,\ell]} \sum_{j \in [0,k]} \mathbf{S}_{i,j} \log \ell_i^{\mathbf{q}} = \sum_{i \in [1,\ell]} \log \ell_i^{\mathbf{q}} \sum_{j \in [0,k]} \mathbf{S}_{i,j} = \sum_{i \in [1,\ell]} \ell_i^{\mathbf{q}} \log \ell_i^{\mathbf{q}}$.

We conclude the proof by combining equations 76 and 78 and using $\sum_{j \in [0,k]} \mathbf{S}_{i,j} = \ell_i^{\mathbf{q}}$ for any $\mathbf{S} \in \mathbf{Z}_{\mathbf{R}}^{\mathbf{q},\phi}$. ■

Note that using Theorem C.3 and C.4, for matrix $\mathbf{A}$ defined (with respect to $\mathbf{q}$ and ϕ) in Equation (67), we showed the following,

$$\max_{\mathbf{S} \in \mathbf{Z}_{\mathbf{R}}^{\mathbf{q},\phi}} \mathbf{h}(\mathbf{S}) \leq \log \mathrm{perm}(\mathbf{A}) \leq O(k \log \frac{N}{k}) + \max_{\mathbf{S} \in \mathbf{Z}_{\mathbf{R}}^{\mathbf{q},\phi}} \mathbf{h}(\mathbf{S}) \; . \tag{79}$$

Our final goal of this section is to maximize $\mathbb{P}(\mathbf{q}, \phi) \propto \frac{1}{\phi_0 !}\mathrm{perm}(\mathbf{A})$ over discrete pseudo-distributions $\mathbf{q}$ but let us take a step back and just focus on writing an upper bound. Consider the term $\max_{\mathbf{S} \in \mathbf{Z}_{\mathbf{R}}^{\mathbf{q},\phi}} \mathbf{h}(\mathbf{S})$ in the expression above, it depends on discrete pseudo-distribution $\mathbf{q}$ at two different places. The first is the constraint set $\mathbf{Z}_{\mathbf{R}}^{\mathbf{q},\phi}$ and the second is the function $\mathbf{h}(\mathbf{S})$ (because it contains the ϕ_0 term in its expression). We address the first issue by defining the following new set that encodes the constraint set $\mathbf{Z}_{\mathbf{R}}^{\mathbf{q},\phi}$ for all discrete pseudo-distributions simultaneously.

Definition C.5 *Let $\mathbf{Z}_{\mathbf{R}}^{\phi} \subset \mathbb{R}_{\geq 0}^{\ell \times (k+1)}$ be the set of non-negative matrices, such that any $\mathbf{S} \in \mathbf{Z}_{\mathbf{R}}^{\phi}$ satisfies,*

$$\sum_{i \in [1,\ell]} \mathbf{S}_{i,j} = \phi_j \textit{ for all } j \in [1,k], \sum_{j \in [0,k]} \mathbf{S}_{i,j} \in \mathbb{Z}_+ \textit{ for all } i \in [1,\ell] \textit{ and } \sum_{i \in [1,k]} \mathbf{r}_i \sum_{j \in [0,k]} \mathbf{S}_{i,j} \leq 1 \; . \tag{80}$$

Note that in the definition of $\mathbf{Z}_{\mathbf{R}}^{\phi}$ we removed the constraint related to ϕ_0 and recall ϕ_0 denotes the number of unseen domain elements. Not having constraint with respect to ϕ_0 helps us encode discrete pseudo-distributions (with respect to $\mathbf{R}$) of different domain sizes. Further for any $\mathbf{S} \in \mathbf{Z}_{\mathbf{R}}^{\phi}$, there is a discrete pseudo-distribution associated with it and we define it next.

Definition C.6 *For any $\mathbf{S} \in \mathbf{Z}_{\mathbf{R}}^{\phi}$, the discrete pseudo-distribution $\mathbf{q}_{\mathbf{S}}$ associated with $\mathbf{S}$ is defined as follows: For any arbitrary $\sum_{j \in [0,k]} \mathbf{S}_{i,j}$ number of domain elements assign probability $\mathbf{r}_i$.*

Note that in the definition above $\mathbf{q}_{\mathbf{S}}$ is a valid pseudo-distribution because of the third condition in Equation (80). Further note that, for any discrete pseudo-distribution $\mathbf{q}$ and $\mathbf{S} \in \mathbf{Z}_{\mathbf{R}}^{\mathbf{q},\phi}$, the distribution $\mathbf{q}_{\mathbf{S}}$ associated with respect to $\mathbf{S}$ is a permutation of distribution $\mathbf{q}$. Since the probability of a profile is invariant under permutations of distribution, we treat all these distributions the same and do not distinguish between them.

We now handle the second issue that corresponds to removing the dependency of discrete pseudo-distribution $\mathbf{q}$ from the function $\mathbf{h}(\mathbf{S})$. To handle this issue, we define a new function $\mathbf{g}(\mathbf{S})$ that when maximized over the set $\mathbf{Z}_{\mathbf{R}}^{\mathbf{q},\phi}$ and $\mathbf{Z}_{\mathbf{R}}^{\phi}$ approximates the value $\mathbb{P}(\mathbf{q}, \phi)$ and $\max_{\mathbf{q} \in \Delta_{\mathbf{R}}^{\mathcal{D}}} \mathbb{P}(\mathbf{q}, \phi)$

respectively (See next theorem for the formal statement). For any $\mathbf{S} \in \mathbb{R}_{\geq 0}^{\ell \times (k+1)}$, the function $\mathbf{g}(\mathbf{S})$ is defined as follows,

$$\mathbf{g}(\mathbf{S}) \overset{\text{def}}{=} \exp\left(\sum_{i\in[1,\ell]}\sum_{j\in[0,k]}\left[\mathbf{S}_{i,j}\log(\frac{\mathbf{r}_i^{\mathbf{m}_j}}{\mathbf{S}_{i,j}})\right] + \sum_{i\in[1,\ell]}\left(\sum_{j\in[0,k]}\mathbf{S}_{i,j}\right)\log\left(\sum_{j\in[0,k]}\mathbf{S}_{i,j}\right)\right) . \tag{81}$$

Note that we switch gears and define the function $\mathbf{g}(\mathbf{S})$ as an exponential function. $\mathbf{g}(\mathbf{S})$ approximates the value $\mathbb{P}(\mathbf{q},\phi)$ instead of log of it and it helps with proof readability. The following theorem summarizes this result.

Theorem C.7 *Let $\boldsymbol{R}$ be a probability discretization set. Given a profile ϕ and discrete pseudo-distribution $\boldsymbol{q}$ with respect to $\boldsymbol{R}$. The following inequality holds,*

$$\exp\left(-O(k\log(N+n))\right)\cdot C_\phi \cdot \max_{\boldsymbol{S}\in\boldsymbol{Z}_{\boldsymbol{R}}^{\boldsymbol{q},\phi}} \boldsymbol{g}(\boldsymbol{S}) \leq \mathbb{P}(\boldsymbol{q},\phi) \leq \exp\left(O\left(k\log\frac{N}{k}\right)\right)\cdot C_\phi \cdot \max_{\boldsymbol{S}\in\boldsymbol{Z}_{\boldsymbol{R}}^{\boldsymbol{q},\phi}} \boldsymbol{g}(\boldsymbol{S}) \tag{82}$$

Further,

$$\exp\left(-O(k\log(N+n))\right)\cdot C_\phi \cdot \max_{\boldsymbol{S}\in\boldsymbol{Z}_{\boldsymbol{R}}^{\phi}} \boldsymbol{g}(\boldsymbol{S}) \leq \max_{\boldsymbol{q}\in\Delta_{\boldsymbol{R}}^{\mathcal{D}}} \mathbb{P}(\boldsymbol{q},\phi) \leq \exp\left(O\left(k\log\frac{N}{k}\right)\right)\cdot C_\phi \cdot \max_{\boldsymbol{S}\in\boldsymbol{Z}_{\boldsymbol{R}}^{\phi}} \boldsymbol{g}(\boldsymbol{S}) \tag{83}$$

Proof For any discrete pseudo-distribution $\mathbf{q}$ with respect to $\mathbf{R}$ and profile ϕ, let $\mathbf{A}$ be the matrix defined (with respect to $\mathbf{q}$ and ϕ) in Equation (67). Then, by Equation (79) we have,

$$\max_{\mathbf{S}\in\mathbf{Z}_{\mathbf{R}}^{\mathbf{q},\phi}} \mathbf{h}(\mathbf{S}) \leq \log \operatorname{perm}(\mathbf{A}) \leq O(k\log\frac{N}{k}) + \max_{\mathbf{S}\in\mathbf{Z}_{\mathbf{R}}^{\mathbf{q},\phi}} \mathbf{h}(\mathbf{S}) .$$

Further by Equation (68) we have,

$$\mathbb{P}(\mathbf{q},\phi) = C_\phi \cdot \left(\prod_{j\in[0,k]}\frac{1}{\phi_j!}\right)\cdot \operatorname{perm}(\mathbf{A}^{\mathbf{q},\phi}) .$$

Combining the above two equations we have,

$$C_\phi\cdot\left(\prod_{j\in[0,k]}\frac{1}{\phi_j!}\right)\cdot \max_{\mathbf{S}\in\mathbf{Z}_{\mathbf{R}}^{\mathbf{q},\phi}} \exp\left(\mathbf{h}(\mathbf{S})\right) \leq \mathbb{P}(\mathbf{q},\phi) \leq \exp\left(O\left(k\log\frac{N}{k}\right)\right)\cdot C_\phi\cdot\left(\prod_{j\in[0,k]}\frac{1}{\phi_j!}\right)\cdot \max_{\mathbf{S}\in\mathbf{Z}_{\mathbf{R}}^{\mathbf{q},\phi}} \exp\left(\mathbf{h}(\mathbf{S})\right) \tag{84}$$

We now simplify the term $\left(\prod_{j\in[0,k]}\frac{1}{\phi_j!}\right)\cdot\exp\left(\mathbf{h}(\mathbf{S})\right)$ in the above expression. First note that for any $\mathbf{S}\in\mathbf{Z}_{\mathbf{R}}^{\mathbf{q},\phi}$,

$$\exp\left(\mathbf{h}(\mathbf{S})\right) = \mathbf{g}(\mathbf{S})\cdot\exp\left(\sum_{j\in[0,k]}\phi_j\log\phi_j - \sum_{j\in[0,k]}\phi_j\right) .$$

Therefore,

$$\left(\prod_{j\in[0,k]}\frac{1}{\phi_j!}\right)\cdot\exp\left(\mathbf{h}(\mathbf{S})\right) = \left(\prod_{j\in[0,k]}\frac{1}{\phi_j!}\right)\cdot\mathbf{g}(\mathbf{S})\cdot\exp\left(\sum_{j\in[0,k]}\phi_j\log\phi_j - \sum_{j\in[0,k]}\phi_j\right) . \tag{85}$$

By Lemma 3.1 (Stirling's approximation) we have,

$$\exp\left(-O\left(k\log(N+n)\right)\right) \leq \left(\prod_{j\in[0,k]}\frac{1}{\phi_j!}\right)\cdot\exp\left(\sum_{j\in[0,k]}\phi_j\log\phi_j-\sum_{j\in[0,k]}\phi_j\right)\leq 1\,. \tag{86}$$

The first inequality follows because for each $j\in[0,k]$, we have $\frac{1}{\phi_j!}\exp\left(\phi_j\log\phi_j-\phi_j\right)\geq\Omega(\frac{1}{\sqrt{\phi_j+1}})$, which by using $\phi_j\leq N+n$ is further lower bounded by $\Omega(\frac{1}{\sqrt{N+n}})\geq\exp\left(-O(\log(N+n))\right)$. Equation (86) follows by taking product over all $j\in[0,k]$. Now combining Equation (86) and Equation (85) we have,

$$\exp\left(-O(k\log(N+n))\right)\cdot\mathbf{g}(\mathbf{S})\leq\left(\prod_{j\in[0,k]}\frac{1}{\phi_j!}\right)\cdot\exp\left(\mathbf{h}(\mathbf{S})\right)\leq\mathbf{g}(\mathbf{S})\,. \tag{87}$$

The first statement of the lemma follows by combining the above Equation (87) with Equation (84), that is we have,

$$\exp\left(-O(k\log(N+n))\right)\cdot C_\phi\cdot\max_{\mathbf{S}\in\mathbf{Z}_{\mathbf{R}}^{\mathbf{q},\phi}}\mathbf{g}(\mathbf{S})\leq\mathbb{P}(\mathbf{q},\phi)\leq\exp\left(O\left(k\log\frac{N}{k}\right)\right)\cdot C_\phi\cdot\max_{\mathbf{S}\in\mathbf{Z}_{\mathbf{R}}^{\mathbf{q},\phi}}\mathbf{g}(\mathbf{S})\,. \tag{88}$$

Given a profile ϕ, for any discrete pseudo-distribution $\mathbf{q}\in\Delta_{\mathbf{R}}^{\mathcal{D}}$ we have $\mathbf{Z}_{\mathbf{R}}^{\mathbf{q},\phi}\subseteq\mathbf{Z}_{\mathbf{R}}^{\phi}$ and further combining it with above inequality we get,

$$\max_{\mathbf{q}\in\Delta_{\mathbf{R}}^{\mathcal{D}}}\mathbb{P}(\mathbf{q},\phi)\leq\exp\left(O\left(k\log\frac{N}{k}\right)\right)\cdot C_\phi\cdot\max_{\mathbf{S}\in\mathbf{Z}_{\mathbf{R}}^{\phi}}\mathbf{g}(\mathbf{S})\,.$$

Note that for any $\mathbf{S}\in\mathbf{Z}_{\mathbf{R}}^{\phi}$, we also have $\mathbf{S}\in\mathbf{Z}_{\mathbf{R}}^{\phi,\mathbf{q}_{\mathbf{S}}}$, where $\mathbf{q}_{\mathbf{S}}$ is the discrete pseudo-distribution associated with respect to $\mathbf{S}$ (See Definition C.6). Therefore,

$$\exp\left(-O(k\log(N+n))\right)\cdot C_\phi\cdot\max_{\mathbf{S}\in\mathbf{Z}_{\mathbf{R}}^{\phi}}\mathbf{g}(\mathbf{S})\leq\exp\left(-O(k\log(N+n))\right)\cdot C_\phi\cdot\max_{\mathbf{q}\in\Delta_{\mathbf{R}}^{\mathcal{D}}}\max_{\mathbf{S}\in\mathbf{Z}_{\mathbf{R}}^{\mathbf{q},\phi}}\mathbf{g}(\mathbf{S})\leq\max_{\mathbf{q}\in\Delta_{\mathbf{R}}^{\mathcal{D}}}\mathbb{P}(\mathbf{q},\phi)\,.$$

For the last inequality in the above derivation we used Equation (88). Now combining the previous two inequalities we conclude the proof. ■

The previous theorem provides an upper bound for the probability of profile with respect to any discrete pseudo-distribution. However one issue with this upper bound is that it is not efficiently computable because the set $\mathbf{Z}_{\mathbf{R}}^{\phi}$ is not a convex set (because of the integrality constraints). We relax these integrality constraints and define the following new set.

Definition C.8 *Let $\mathbf{Z}_{\mathbf{R}}^{\phi,frac}\subseteq\mathbb{R}_{\geq 0}^{\ell\times(k+1)}$ be the set of non-negative matrices, such that any $\mathbf{S}\in\mathbf{Z}_{\mathbf{R}}^{\phi,frac}$ satisfies,*

$$\sum_{i\in[1,\ell]}\mathbf{S}_{i,j}=\phi_j \textit{ for all } j\in[1,k] \textit{ and } \sum_{i\in[1,k]}\mathbf{r}_i\sum_{j\in[0,k]}\mathbf{S}_{i,j}\leq 1\,. \tag{89}$$

Lemma C.9 *Let $\boldsymbol{R}$ be a probability discretization set. Given a profile ϕ, the following holds,*

$$\max_{\boldsymbol{q}\in\Delta_{\boldsymbol{R}}^{\mathcal{D}}} \mathbb{P}(\boldsymbol{q},\phi) \leq \exp\left(O\left(k\log\frac{N}{k}\right)\right)\cdot C_{\phi}\cdot \max_{\boldsymbol{S}\in\boldsymbol{Z}_{\boldsymbol{R}}^{\phi,frac}} \boldsymbol{g}(\boldsymbol{S}) \tag{90}$$

Proof By Theorem C.7 we already have,

$$\max_{\mathbf{q}\in\Delta_{\mathbf{R}}^{\mathcal{D}}} \mathbb{P}(\mathbf{q},\phi) \leq \exp\left(O\left(k\log\frac{N}{k}\right)\right)\cdot C_{\phi}\cdot \max_{\mathbf{S}\in\mathbf{Z}_{\mathbf{R}}^{\phi}} \mathbf{g}(\mathbf{S})\ .$$

The lemma holds because $\mathbf{Z}_{\mathbf{R}}^{\phi}\subseteq\mathbf{Z}_{\mathbf{R}}^{\phi,frac}$. ■

Note that in the above lemma, the upper bound only depends on the profile [10] and we removed all dependencies related to distributions (and also ϕ_0). Next we show that this upper bound can be efficiently computed by using the result that function $\mathbf{g}(\mathbf{S})$ is log concave in $\mathbf{S}$.

Lemma C.10 (Lemma 4.16 in Charikar et al. (2019a)) *Function $\boldsymbol{g}(\boldsymbol{S})$ is log concave in $\boldsymbol{S}$.*

Theorem C.11 (Theorem 4.17 in Charikar et al. (2019a)) *Given a profile $\phi\in\Phi^n$, the optimization problem $\max_{\boldsymbol{S}\in\boldsymbol{Z}_{\boldsymbol{R}}^{\phi,frac}}\log\boldsymbol{g}(\boldsymbol{S})$ can be solved in time $\widetilde{O}(k^2\ell)$.* [11]

C.2. Rounding Algorithm

In the previous section we provided an efficiently computable upper bound for the probability of profile ϕ with respect to any discrete pseudo-distribution $\mathbf{q}\in\Delta_{\mathbf{R}}^{\mathcal{D}}$. This upper bound returns a solution $\mathbf{S}\in\mathbf{Z}_{\mathbf{R}}^{\phi,frac}$ and we need to round this solution to construct a discrete pseudo-distribution that approximates this upper bound. In this section we provide a rounding algorithm that takes as input $\mathbf{S}\in\mathbf{Z}_{\mathbf{R}}^{\phi,frac}$ and returns a solution $\mathbf{S}^{\text{ext}}\in\mathbf{Z}_{\mathbf{R}^{\text{ext}}}^{\phi}$, where $\mathbf{R}^{\text{ext}}$ is an extended probability discretization set. Further using $\mathbf{S}^{\text{ext}}\in\mathbf{Z}_{\mathbf{R}^{\text{ext}}}^{\phi}$, we construct a discrete pseudo-distribution $\mathbf{q}_{\mathbf{S}^{\text{ext}}}$ with respect to $\mathbf{R}^{\text{ext}}$ such that $\mathbb{P}(\mathbf{q}_{\mathbf{S}^{\text{ext}}},\phi)$ approximates the upper bound and therefore is an approximate PML distribution. Our rounding algorithm is technical and we next provide a overview to better understand it.

Overview of the rounding algorithm: The goal of the rounding algorithm is to take a fractional solution $\mathbf{S}\stackrel{\text{def}}{=}\arg\max_{\mathbf{S}'\in\mathbf{Z}_{\mathbf{R}}^{\phi,frac}}\log\mathbf{g}(\mathbf{S}')$ as input and round each row sum to an integral value while preserving the column sums and $\mathbf{g}(\mathbf{S})$ value. Our rounding algorithm proceeds in three steps:

Step 1: Consider the fractional solution $\mathbf{S}\in\mathbb{R}_{\geq 0}^{\ell\times(k+1)}$ and recall the rows are indexed by the elements of set $\mathbf{R}$ (which represent probability values). We first round the rows corresponding to the higher probability values by simply taking the floor (rounding down to the nearest integer) of each entry. This procedure ensures the integrality of the row sums (corresponding to higher probability values) but violates the column sum constraints. To satisfy the column sum constraints and the distributional constraint (i.e. last condition in Equation (80)) simultaneously, we create rows corresponding to new probability values using Algorithm 2. However to ensure that all these new

[10] C_{ϕ} has no dependency on ϕ_0.

[11] Note that here we hide the logarithmic dependence on n, the size of sample.

rows also have integral row sums, we modify the (old) rows corresponding to lower probability values accordingly. Let $\mathbf{S}^{(1)}$ be the solution returned by the first step of the rounding algorithm. Algorithm 2 ensures that the $\mathbf{g}(\mathbf{S}^{(1)})$ value is not much smaller than $\mathbf{g}(\mathbf{S})$. In $\mathbf{S}^{(1)}$, all the new rows and (old) rows corresponding to higher probability values have integral row sums and we round the remaining rows corresponding to smaller probability values next.

Step 2: In this step, we round all the rows corresponding to the smaller probability values. For each of these rows, we scale all the entries in a particular row by the same factor to ensure that the row sum is rounded down to the nearest integer. Similar to the step 1, using Algorithm 2 we create rows corresponding to new probability values to maintain the column sum constraints and the distributional constraint; all these new rows further correspond to small probability values. Unlike in the previous step, the new rows created in step two may not have integral row sums but these rows have a nice diagonal structure. Let $\mathbf{S}^{(2)}$ be this intermediate solution created in step 2. Algorithm 2 ensures that the $\mathbf{g}(\mathbf{S}^{(2)})$ value is not much smaller than $\mathbf{g}(\mathbf{S}^{(1)})$ (and hence $\mathbf{g}(\mathbf{S})$). Note that all the row sums in $\mathbf{S}^{(2)}$ are integral except the new rows created in step 2 that all have small probability values and have diagonal structure.

Step 3: In this final step, using Algorithm 1 we round the new rows created in step 2. Algorithm 1 exploits the low probability and diagonal structure in these rows. The diagonal structure ensures that there is just one non-zero entry in any particular row and we modify the solution $\mathbf{S}^{(2)}$ (from the previous step) as follows. We transfer the mass from a non-integral lower probability value row to an immediate higher probability value row until the (lower probability value) row sum is integral. This process might violate the distributional constraint and we rescale the probability values accordingly to satisfy this constraint. Let $\mathbf{S}^{\mathrm{ext}}$ be the solution returned by step 3. We ensure that all column sums are preserved, all row sums are integral and the $\mathbf{g}(\mathbf{S}^{\mathrm{ext}})$ value is not much smaller than $\mathbf{g}(\mathbf{S}^{(2)})$ (and hence not much smaller than $\mathbf{g}(\mathbf{S})$).

In the remainder of this section we state all three algorithms and the results corresponding to them. For continuity of reading, we defer the proofs of these results to Appendix C.4. For convenience, we first state Algorithm 1 that rounds the rows corresponding to the low probability values in step 3 of our main rounding algorithm (Algorithm 3). We follow up this algorithm with a lemma that summarizes the guarantees provided by it. Later we state Algorithm 2 that creates rows corresponding to new probability values to preserve the column sums and the distributional constraint. This algorithm is invoked as a subroutine in both step 1 and 2 of Algorithm 3. Finally, we state our main rounding algorithm that consists of three different steps. We then state results analyzing each of these steps separately. The final result (Theorem C.16), is the main theorem of this subsection that summarizes the final guarantees promised by our rounding algorithm.

Algorithm 1 Structured Rounding Algorithm

```
Procedure StructuredRounding($(x, w, a)$)
    Input: $x \in (0,1)_{\mathbb{R}}^{[0,k]}$, $w \in \mathbb{R}^{[0,k]}$ and $a = \sum_{j \in [0,k]} x_j \in \mathbb{Z}_{+}$.
    Output: $z \in \mathbb{R}^{[0,k] \times [0,k]}$ and $s \in \mathbb{R}^{a}$.
    Initialize $z = \mathbf{0}^{[0,k] \times [0,k]}$.
    For each $i \in [1,a]$, let $s_i$ denote the smallest index such that $\sum_{j \leq s_i} x_j > i-1$ and let $s_{a+1} = k$.
    for $i \in [1,a]$ do
        $z_{s_i,j} = \begin{cases} x_j & \text{if } s_i < j < s_{i+1} \text{ ,} \\ \sum_{j' \leq s_i} x_{j'} - (i-1) & \text{if } j = s_i \text{ ,} \\ 1 - \sum_{s_i \leq j' < s_{i+1}} z_{s_i,j'} & \text{if } j = s_{i+1} \text{ .} \end{cases}$    (91)
    end
    return $z$ and $s$.
```

The next lemma summarizes the quality of the solution produced by Algorithm 1.

Lemma C.12 *Given a set of reals $x_j \in (0,1)$ for all $j \in [0,k]$ such that $\sum_{j \in [0,k]} x_j \in \mathbb{Z}_{+}$, weights w_j for all $j \in [0,k]$ and exponents $m_j \in \mathbb{Z}_{+}$ for all $j \in [0,k]$* [12]. *Using Algorithm 1, we can efficiently compute a matrix $z \in [0,1]_{\mathbb{R}}^{[0,k] \times [0,k]}$ such that the following conditions hold,*

1. *$\sum_{j \in [0,k]} z_{i,j} \in \{0,1\}$ for all $i \in [0,k]$ and $\sum_{i \in [0,k]} z_{i,j} = x_j$ for all $j \in [0,k]$.*

2. *$\sum_{i \in [0,k]} \left(\sum_{j \in [0,k]} z_{i,j} \right) w_i \leq \sum_{j \in [0,k]} x_j w_j + \max_{j \in [0,k]} w_j$.*

3. *$\prod_{j \in [0,k]} w_j^{m_j x_j} \leq \prod_{i \in [0,k]} \prod_{j \in [0,k]} w_i^{m_j z_{i,j}}$.*

We next provide description of Algorithm 2. The algorithm takes input $(\mathbf{B}, \mathbf{C}, \mathbf{R}, \phi)$ and creates a new probability discretization set $\mathbf{R}'$ (lines 6-10). The solution $\mathbf{B}'$ outputted by the algorithm belongs to $\mathbf{Z}_{\mathbf{R}'}^{\phi,frac}$, has same column sums as $\mathbf{B}$ and the value $\mathbf{g}(\mathbf{B}')$ is lower bounded by $\mathbf{g}(\mathbf{B})$.

[12] Here m_0 need not be equal to zero.

Algorithm 2 Create New Probability Values

```
Procedure CreateNewProbabilityValues ($\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{R}, \phi$)
    Input: Probability discretization set $\mathbf{R}$ ($|\mathbf{R}| = t$), profile $\phi$ (let $k$ be the number of distinct frequencies) and $\mathbf{B} \in \mathbf{Z}_{\mathbf{R}}^{\phi,frac} \subseteq \mathbb{R}^{[1,t]\times[0,k]}$ and $\mathbf{C} \in \mathbb{R}^{[1,t]\times[0,k]}$ such that $\mathbf{C}_{i,j} \leq \mathbf{B}_{i,j}$ for all $i \in [1,t]$ and $j \in [0,k]$. Let $\mathbf{r}_i$ be the $i$'th element of $\mathbf{R}$.
    Output: Probability discretization set $\mathbf{R}'$ and $\mathbf{B}' \in \mathbb{R}^{[1,t+(k+1)]\times[0,k]}$.
    Initialize $\mathbf{B}' = \mathbf{0}^{[1,t+(k+1)]\times[0,k]}$.
    $\mathbf{B}'_{ij} = \mathbf{C}_{ij}$ for all $i \in [1,t], j \in [0,k]$ .
    for $j \in [0,k]$ do
        Create a new row with probability value $\mathbf{r}_{t+1+j} = \frac{\sum_{i\in[1,t]}(\mathbf{B}_{ij}-\mathbf{C}_{ij})\mathbf{r}_i}{\sum_{i\in[1,t]}(\mathbf{B}_{ij}-\mathbf{C}_{ij})}$.
        Assign $\mathbf{B}'_{t+1+j,j} = \sum_{i\in[1,t]}(\mathbf{B}_{ij} - \mathbf{C}_{ij})$.
    end
    Define $\mathbf{R}' \stackrel{\mathrm{def}}{=} \mathbf{R} \cup \{\mathbf{r}_{t+1+j}\}_{j\in[0,k]}$.
    return $\mathbf{R}'$ and $\mathbf{B}'$.
```

The next lemma summarizes the quality of the solution produced by Algorithm 2.

Lemma C.13 *The solution* $(\boldsymbol{R}', \boldsymbol{B}')$ *returned by Algorithm 2 satisfies the following conditions:*

1. $\sum_{j\in[0,k]} \boldsymbol{B}'_{i,j} = \sum_{j\in[0,k]} \boldsymbol{C}_{i,j}$ *for all* $i \in [1,t]$.

2. *For any* $i \in [t+1, t+(k+1)]$ *let* $j \in [0,k]$ *be such that* $i = t+1+j$ *then* $\boldsymbol{B}'_{t+1+j,j'} = 0$ *for all* $j' \in [0,k]$ *and* $j' \neq j$. *(Diagonal Structure)*

3. *For any* $i \in [t+1, t+(k+1)]$ *let* $j \in [0,k]$ *be such that* $i = t+1+j$, *then* $\sum_{j'\in[0,k]} \boldsymbol{B}'_{i,j'} = \boldsymbol{B}'_{t+1+j,j} = \phi_j - \sum_{i'\in[1,t]} \boldsymbol{C}_{i',j}$.

4. $\boldsymbol{B}' \in \boldsymbol{Z}_{\boldsymbol{R}'}^{\phi,frac}$ *and* $\sum_{i\in[1,t+(k+1)]} \sum_{j\in[0,k]} \boldsymbol{B}'_{i,j} = \sum_{i\in[1,t]} \sum_{j\in[0,k]} \boldsymbol{B}_{i,j}$.

5. *Let* $\alpha_i \stackrel{\mathrm{def}}{=} \sum_{j\in[0,k]} \boldsymbol{B}_{i,j} - \sum_{j\in[0,k]} \boldsymbol{C}_{i,j}$ *for all* $i \in [1,t]$ *and* $\Delta \stackrel{\mathrm{def}}{=} \max(\sum_{i\in[1,t]}(\boldsymbol{B}\vec{1})_i, t \times k)$, *then* $\boldsymbol{g}(\boldsymbol{B}') \geq \exp\left(-O\left(\sum_{i\in[1,t]} \alpha_i \log\Delta\right)\right)\boldsymbol{g}(\boldsymbol{B})$.

6. *For each* $j \in [0,k]$, *the new row corresponds to the probability value,* $\boldsymbol{r}_{t+1+j} = \frac{\sum_{i\in[1,t]}(\boldsymbol{B}_{ij}-\boldsymbol{C}_{ij})\boldsymbol{r}_i}{\sum_{i\in[1,t]}(\boldsymbol{B}_{ij}-\boldsymbol{C}_{ij})}$.

In the remainder of this section, we state and analyze our rounding algorithm. Our algorithm works in three steps, and we show that all the solutions produced during the intermediate and final steps all have the desired approximation guarantee. We divide the analysis into three lemmas. Each of the lemmas C.14, C.15 and C.16 analyze the guarantees provided by the intermediate solutions $\mathbf{S}^{(1)}$, $\mathbf{S}^{(2)}$ and final solution $\mathbf{S}^{\mathrm{ext}}$ respectively.

Algorithm 3 Rounding Algorithm

```
Procedure Rounding($S$)
    Input: Probability discretization set $\mathbf{R}$, profile $\phi \in \Phi^n$ and $\mathbf{S} \in \mathbf{Z}^{\phi,frac}_{\mathbf{R}} \subseteq \mathbb{R}^{[1,\ell]\times[0,k]}$.
    Output: Probability discretization set $\mathbf{R}^{\text{ext}}$ and $\mathbf{S}^{\text{ext}}$.
    Step 1:
    Initialize $\mathbf{A} = \mathbf{0}^{[1,\ell]\times[0,k]}$. Let $\mathbf{r}_i$ be the $i$'th element of $\mathbf{R}$.
    Define $\mathbf{H} \overset{\text{def}}{=} \{i \in [1,\ell] \mid \mathbf{r}_i > \gamma\}$ and $\mathbf{L} \overset{\text{def}}{=} \{i \in [1,\ell] \mid \mathbf{r}_i \leq \gamma\}$.
    $\mathbf{A}_{ij} = \lfloor \mathbf{S}_{ij} \rfloor$ for all $i \in \mathbf{H}, j \in [0,k]$ .
    $\mathbf{A}_{ij} = \mathbf{S}_{i,j}\frac{\lfloor \sum_{i\in\mathbf{L}} \mathbf{S}_{i,j} \rfloor}{\sum_{i\in\mathbf{L}} \mathbf{S}_{i,j}}$ for all $i \in \mathbf{L}, j \in [0,k]$ .
    $(\mathbf{S}^{(1)}, \mathbf{R}^{(1)}) = \text{CreateNewProbabilityValues}(\mathbf{S}, \mathbf{A}, \mathbf{R})$.
    Step 2:
    Note that $|\mathbf{R}^{(1)}| = \ell + (k+1)$ and $\mathbf{S}^{(1)} \subseteq \mathbb{R}^{[1,\ell+(k+1)]\times[0,k]}$. Let $\mathbf{r}^{(1)}_i$ be the $i$'th element of $\mathbf{R}^{(1)}$.
    Let $\mathbf{H}^{(1)} \overset{\text{def}}{=} \{i \in [1,\ell+(k+1)] \mid \mathbf{r}^{(1)}_i > \gamma\}$ and $\mathbf{L}^{(1)} \overset{\text{def}}{=} \{i \in [1,\ell+(k+1)] \mid \mathbf{r}^{(1)}_i \leq \gamma\}$.
    Define $\mathbf{A}^{(1)} = \mathbf{0}^{[1,\ell+(k+1)]\times[0,k]}$.
    $\mathbf{A}^{(1)}_{ij} = \mathbf{S}^{(1)}_{ij}$    for all $i \in \mathbf{H}^{(1)}, j \in [0,k]$ .
    $\mathbf{A}^{(1)}_{ij} = \mathbf{S}^{(1)}_{ij}\frac{\lfloor (\mathbf{S}^{(1)}\vec{1})_i \rfloor}{(\mathbf{S}^{(1)}\vec{1})_i}$    for all $i \in \mathbf{L}^{(1)}, j \in [0,k]$ .
    $(\mathbf{S}^{(2)}, \mathbf{R}^{(2)}) = \text{CreateNewProbabilityValues}(\mathbf{S}^{(1)}, \mathbf{A}^{(1)}, \mathbf{R}^{(1)})$.
    Step 3:
    Note that $|\mathbf{R}^{(2)}| = \ell + 2(k+1)$ and $\mathbf{S}^{(2)} \subseteq \mathbb{R}^{[1,\ell+2(k+1)]\times[0,k]}$. Let $\mathbf{r}^{(2)}_i$ be the $i$'th element of $\mathbf{R}^{(2)}$.
    Let $w, x \in \mathbb{R}^{[0,k]}$, such that $w_j \overset{\text{def}}{=} \mathbf{r}^{(2)}_{\ell+(k+1)+1+j}$ and $x_j \overset{\text{def}}{=} \mathbf{S}^{(2)}_{\ell+(k+1)+1+j} - \lfloor \mathbf{S}^{(2)}_{\ell+(k+1)+1+j} \rfloor$ for all $j \in [0,k]$. Define $a \overset{\text{def}}{=} \sum_{j\in[0,k]} x_j$.
    Let $(z, s) \overset{\text{def}}{=} \text{StructuredRounding}(x, w, a)$.
    Initialize $\mathbf{S}^{\text{ext}} = 0^{[1,\ell+2(k+1)]\times[0,k]}$.
    Assign $\mathbf{S}^{\text{ext}}_{i,j} = \mathbf{S}^{(2)}_{i,j}$ for all $i \in [1,\ell+(k+1)]$ and $j \in [0,k]$.
    Assign $\mathbf{S}^{\text{ext}}_{\ell+(k+1)+1+j,j'} = \lfloor \mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j'} \rfloor + z_{j,j'}$ for all $j, j' \in [0,k]$.
    Define $\mathbf{R}^{\text{ext}} \overset{\text{def}}{=} \{\frac{\mathbf{r}^{(2)}_i}{1+\gamma} \mid \text{for all } i \in [1,\ell+2(k+1)]\}$.
    return $\mathbf{R}^{\text{ext}}$ and $\mathbf{S}^{\text{ext}}$.
```

The next lemma summarizes the quality of the intermediate solution $(\mathbf{S}^{(1)}, \mathbf{R}^{(1)})$ produced by Step 1 of Algorithm 3.

Lemma C.14 *The solution $(\boldsymbol{S}^{(1)}, \boldsymbol{R}^{(1)})$ returned by the step 1 of Algorithm 3 satisfies the following:*

1. $(\boldsymbol{S}^{(1)}\vec{1})_i \in \mathbb{Z}_+$ *for all* $i \in \boldsymbol{H}^{(1)}$.

2. $\boldsymbol{S}^{(1)} \in \boldsymbol{Z}^{\phi,frac}_{\boldsymbol{R}^{(1)}}$ *and* $\sum_{i\in[1,\ell+(k+1)]}\sum_{j\in[0,k]} \boldsymbol{S}^{(1)}_{i,j} = \sum_{i\in[1,\ell]}\sum_{j\in[0,k]} \boldsymbol{S}_{i,j}$.

3. $\boldsymbol{g}(\boldsymbol{S}^{(1)}) \geq \exp\left(-O\left(\left(\frac{1}{\gamma} + k\right)\log\Delta\right)\right)\boldsymbol{g}(\boldsymbol{S})$, *where* $\Delta = \max(\sum_{i\in[1,\ell]}(\boldsymbol{S}\vec{1})_i, \ell \times k)$.

Using Lemma C.14 we now provide the guarantees for the solution $\mathbf{S}^{(2)}$ returned by the step 2 of Algorithm 3.

Lemma C.15 *The solution $(\boldsymbol{S}^{(2)}, \boldsymbol{R}^{(2)})$ returned by the step 2 of Algorithm 3 satisfies the following,*

1. $(\boldsymbol{S}^{(2)}\overrightarrow{1})_i \in \mathbb{Z}_+$ *for all* $i \in [1, \ell + (k+1)]$.
2. $\boldsymbol{S}^{(2)}_{\ell+(k+1)+1+j,j'} = 0$ *for all* $j, j' \in [0, k]$ *and* $j \neq j'$ *(Diagonal Structure).*
3. $\boldsymbol{S}^{(2)} \in \boldsymbol{Z}^{\phi,frac}_{\boldsymbol{R}^{(2)}}$ *and* $\sum_{i\in[1,\ell+2(k+1)]}\sum_{j\in[0,k]}\boldsymbol{S}^{(2)}_{i,j} = \sum_{i\in[1,\ell+(k+1)]}\sum_{j\in[0,k]}\boldsymbol{S}^{(1)}_{i,j}$.
4. $\sum_{i\in[\ell+(k+1)+1,\ell+2(k+1)]}(\boldsymbol{S}^{(2)}\overrightarrow{1})_i \in \mathbb{Z}_+$.
5. *For any* $j \in [0, k]$, $\boldsymbol{r}^{(2)}_{\ell+(k+1)+1+j} \leq \gamma$.
6. $\boldsymbol{g}(\boldsymbol{S}^{(2)}) \geq \exp\left(-O\left(\left(\frac{1}{\gamma} + \ell + k\right)\log\Delta\right)\right)\boldsymbol{g}(\boldsymbol{S})$.

Using Lemma C.15 we now provide the guarantees for the final solution $\mathbf{S}^{\mathrm{ext}}$ returned by Algorithm 3.

Theorem C.16 *The final solution returned $(\boldsymbol{S}^{\mathrm{ext}}, \boldsymbol{R}^{\mathrm{ext}})$ by Algorithm 3 satisfies the following,*

1. $\boldsymbol{S}^{\mathrm{ext}} \in \boldsymbol{Z}^{\phi}_{\boldsymbol{R}^{\mathrm{ext}}}$.
2. $\boldsymbol{g}(\boldsymbol{S}^{\mathrm{ext}}) \geq \exp\left(-O\left(\left(\frac{1}{\gamma} + \ell + k + \gamma n\right)\log\Delta\right)\right)\boldsymbol{g}(\boldsymbol{S})$.

C.3. Combining everything together

Here we combine the analysis from previous two sections to provide an efficient algorithm to compute an $\exp(\sqrt{n}\log n)$ approximate PML distribution. The main contribution of this section is to define a probability discretization set $\mathbf{R}$ that guarantees existence of a discrete pseudo-distribution $\mathbf{q}$ with respect to $\mathbf{R}$ that is also an $\exp(\sqrt{n}\log n)$ approximate PML pseudo-distribution. We further use this probability discretization set $\mathbf{R}$ and combine it with results from the previous two sections to finally output an $\exp(\sqrt{n}\log n)$ approximate PML distribution. In this direction, we first provide definition of $\mathbf{R}$ that has desired guarantees and such a set $\mathbf{R}$ was already constructed in Charikar et al. (2019a) and we formally state results from Charikar et al. (2019a) that help us define such a set $\mathbf{R}$.

Lemma C.17 (Lemma 4.1 in Charikar et al. (2019a)) *For any profile $\phi \in \Phi^n$, there exists a distribution $\boldsymbol{q}'' \in \Delta^{\mathcal{D}}$ such that $\boldsymbol{q}''$ is an $\exp(-6)$-approximate PML distribution and* $\min_{x\in\mathcal{D}:\boldsymbol{q}''_x\neq 0}\boldsymbol{q}''_x \geq \frac{1}{2n^2}$.

The above lemma allows to define a region in which our approximate PML takes all its probability values and we use idea similar to Charikar et al. (2019a) to define this region.

Let $\mathbf{R} \stackrel{\mathrm{def}}{=} \{(1+\epsilon)^{1-i}\}_{i\in[\ell]}$ be a discretization of probability space, where $\ell = O(\frac{\log n}{\epsilon})$ is the smallest integer such that $\frac{1}{4n^2} \leq (1+\epsilon)^{1-\ell} \leq \frac{1}{2n^2}$ for some $\epsilon \in (0,1)$. Fix any arbitrary order for the elements of set $\mathbf{R}$, we use $\mathbf{r}_i$ to denote the i'th element of this set. We next state a result in Charikar et al. (2019a) that captures the effect of this discretization.

Lemma C.18 (Lemma 4.4 in Charikar et al. (2019a)) *For any profile $\phi \in \Phi^n$ and distribution $\boldsymbol{p} \in \Delta^{\mathcal{D}}$, its discrete pseudo-distribution $\boldsymbol{q} = \mathrm{disc}(\boldsymbol{p}) \in \Delta^{\mathcal{D}}_{\boldsymbol{R}}$ satisfies:*

$$\mathbb{P}(\boldsymbol{p},\phi) \geq \mathbb{P}(\boldsymbol{q},\phi) \geq \exp\left(-\epsilon n\right) \mathbb{P}(\boldsymbol{p},\phi)\,.$$

We are now ready to state our final algorithm. Following this algorithm, we prove that it returns an approximate PML distribution.

Algorithm 4 Algorithm for approximate PML

```
Procedure Approximate PML($\phi$, $\boldsymbol{R}$)
    Input: Profile $\phi \in \Phi^n$ and probability discretization set $\mathbf{R}$.
    Output: A distribution $\mathbf{p}_{\mathrm{approx}}$.
    Solve $\mathbf{S} = \arg\max_{\mathbf{A} \in \mathbf{Z}^{\phi,frac}_{\mathbf{R}}} \log \mathbf{g}(\mathbf{A})$.
    Use Algorithm 3 to round the fractional solution $\mathbf{S}$ to integral solution $\mathbf{S}^{\mathrm{ext}} \in \mathbf{Z}^{\phi}_{\mathbf{R}^{\mathrm{ext}}}$.
    Construct discrete pseudo-distribution $\mathbf{q}_{\mathbf{S}^{\mathrm{ext}}}$ with respect to $\mathbf{S}^{\mathrm{ext}}$ (See Definition C.6).
    return $\mathbf{p}_{\mathrm{approx}} \stackrel{\mathrm{def}}{=} \frac{\mathbf{q}_{\mathbf{S}^{\mathrm{ext}}}}{\|\mathbf{q}_{\mathbf{S}^{\mathrm{ext}}}\|_1}$.
```

Theorem 4.1 ($\exp(\sqrt{n}\log n)$-approximate PML) *For any given profile $\phi \in \Phi^n$, Algorithm 4 computes an $\exp\left(-O(\sqrt{n}\log n)\right)$-approximate PML distribution in $\widetilde{O}(n^{1.5})$ time.*

Proof Choose $\epsilon = \frac{\log n}{\sqrt{n}}$ and let the probability discretization space $\mathbf{R} \stackrel{\mathrm{def}}{=} \{(1+\frac{1}{\sqrt{n}})^{1-i}\}_{i \in [\ell]}$ and $\ell \stackrel{\mathrm{def}}{=} |\mathbf{R}|$ be the smallest integer such that $\frac{1}{2n^2} \geq (1+\frac{1}{\sqrt{n}})^{1-\ell} \geq \frac{1}{4n^2}$ and therefore $\ell \in O(\sqrt{n})$. Let $\mathbf{r}_i$ be the i'th element of set $\mathbf{R}$ and we have $\mathbf{r}_i \geq \frac{1}{4n^2}$.

Given profile ϕ, let $\mathbf{p}_{\mathrm{pml}}$ be the PML distribution. Define $\mathbf{q}_{\mathrm{pml}} \stackrel{\mathrm{def}}{=} \lfloor \mathbf{p}_{\mathrm{pml}} \rfloor_{\mathbf{R}}$ and by Lemma C.18 (and choice of ϵ) we have,

$$\mathbb{P}(\mathbf{q}_{\mathrm{pml}},\phi) \geq \exp\left(-O(\sqrt{n}\log n)\right) \mathbb{P}(\mathbf{p}_{\mathrm{pml}},\phi)\,. \tag{92}$$

Let $\mathbf{S} \stackrel{\mathrm{def}}{=} \arg\max_{\mathbf{A} \in \mathbf{Z}^{\phi,frac}_{\mathbf{R}}} \mathbf{g}(\mathbf{A})$, then by Lemma C.9 we have,

$$\max_{\mathbf{q} \in \Delta^{\mathcal{D}}_{\mathbf{R}}} \mathbb{P}(\mathbf{p},\phi) \leq \exp\left(O\left(k \log \frac{N}{k}\right)\right) \cdot C_{\phi} \cdot \mathbf{g}(\mathbf{S})\,. \tag{93}$$

Note that $\mathbf{q}_{\mathrm{pml}} \in \Delta^{\mathcal{D}}_{\mathbf{R}}$, therefore $\mathbb{P}(\mathbf{q}_{\mathrm{pml}},\phi) \leq \max_{\mathbf{q} \in \Delta^{\mathcal{D}}_{\mathbf{R}}} \mathbb{P}(\mathbf{q},\phi)$ and further combined with equations 92 and 93 we have,

$$\mathbb{P}(\mathbf{p}_{\mathrm{pml}},\phi) \leq \exp\left(O\left(k \log \frac{N}{k} + \sqrt{n}\log n\right)\right) \cdot C_{\phi} \cdot \mathbf{g}(\mathbf{S})\,. \tag{94}$$

Let $\mathbf{S}^{\mathrm{ext}}$ and $\mathbf{R}^{\mathrm{ext}}$ be the solution returned by Algorithm 3, then by the second condition of Theorem C.16 we have,

$$\mathbf{g}(\mathbf{S}^{\mathrm{ext}}) \geq \exp\left(-O\left(\left(\frac{1}{\gamma} + \ell + k + \gamma n\right) \log \Delta\right)\right) \mathbf{g}(\mathbf{S}) \tag{95}$$

Combining equations 94 and 95 we have,

$$\mathbb{P}(\mathbf{p}_{\mathrm{pml}},\phi) \leq \exp\left(O\left(k\log\frac{N}{k}+\sqrt{n}\log n+\left(\frac{1}{\gamma}+\ell+k+\gamma n\right)\log\Delta\right)\right)\cdot C_{\phi}\cdot\mathbf{g}(\mathbf{S}^{\mathrm{ext}})\;. \quad (96)$$

We now simplify the above expression by providing the bounds and values for parameters k,ℓ,γ,N and Δ. We choose $\gamma=\frac{1}{\sqrt{n}}$ and recall $\ell\in O(\sqrt{n})$. Given n samples, the number of distinct frequencies in upper bounded by $\sqrt{n}$ and therefore $k\leq\sqrt{n}$. By Lemma C.17, up to constant multiplicative loss we can assume that the minimum non-zero probability value of our approximate PML distribution is at least $\frac{1}{4n^2}$ and therefore the support $N\leq 4n^2$. Recall by the third condition of Lemma C.14, we have $\Delta=\max(\sum_{i\in[1,\ell]}(\mathbf{S}\vec{1})_i,\ell\times k)$. The condition $\mathbf{S}\in\mathbf{Z}_{\mathbf{R}}^{\phi,frac}$ implies $\sum_{i\in[1,\ell]}\mathbf{r}_i(\mathbf{S}\vec{1})_i\leq 1$ and further using $\mathbf{r}_i\geq\frac{1}{4n^2}$ for all $i\in[1,\ell]$ we have $\sum_{i\in[1,\ell]}(\mathbf{S}\vec{1})_i\leq 4n^2$. Therefore $\Delta\leq\max(4n^2,\ell\times k)\in O(n^2)$.

Substituting these values in Equation (96) we get,

$$\mathbb{P}(\mathbf{p}_{\mathrm{pml}},\phi)\leq\exp\left(O\left(\sqrt{n}\log n\right)\right)\cdot C_{\phi}\cdot\mathbf{g}(\mathbf{S}^{\mathrm{ext}})\;. \quad (97)$$

By the first condition of Theorem C.16 we have $\mathbf{S}^{\mathrm{ext}}\in\mathbf{Z}^{\phi}_{\mathbf{R}^{\mathrm{ext}}}$. Let $\mathbf{q}_{\mathbf{S}^{\mathrm{ext}}}$ be the discrete pseudo-distribution with respect to $\mathbf{S}^{\mathrm{ext}}$, then the condition $\mathbf{S}^{\mathrm{ext}}\in\mathbf{Z}^{\phi}_{\mathbf{R}^{\mathrm{ext}}}$ further implies $\mathbf{S}^{\mathrm{ext}}\in\mathbf{Z}^{\mathbf{q}_{\mathbf{S}^{\mathrm{ext}}},\phi}_{\mathbf{R}^{\mathrm{ext}}}$ and combined with Theorem C.7 we have,

$$\exp\left(-O(k\log(N+n))\right)\cdot C_{\phi}\cdot\mathbf{g}(\mathbf{S}^{\mathrm{ext}})\leq\mathbb{P}(\mathbf{q}_{\mathbf{S}^{\mathrm{ext}}},\phi) \quad (98)$$

Combining equations 97, 98, $k\leq\sqrt{n}$ and $N\leq 4n^2$ we have,

$$\mathbb{P}(\mathbf{q}_{\mathbf{S}^{\mathrm{ext}}},\phi)\geq\exp\left(-O\left(\sqrt{n}\log n\right)\right)\mathbb{P}(\mathbf{p}_{\mathrm{pml}},\phi)\;. \quad (99)$$

Define $\mathbf{p}_{\mathrm{approx}}\stackrel{\mathrm{def}}{=}\frac{\mathbf{q}_{\mathbf{S}^{\mathrm{ext}}}}{\|\mathbf{q}_{\mathbf{S}^{\mathrm{ext}}}\|_1}$, then $\mathbf{p}_{\mathrm{approx}}$ is a distribution, $\mathbb{P}(\mathbf{p}_{\mathrm{approx}},\phi)\geq\mathbb{P}(\mathbf{q}_{\mathbf{S}^{\mathrm{ext}}},\phi)$ (because $\mathbf{q}_{\mathbf{S}^{\mathrm{ext}}}$ is a pseudo-distribution and $\|\mathbf{q}_{\mathbf{S}^{\mathrm{ext}}}\|_1\leq 1$) and combined with Equation (99) we get,

$$\mathbb{P}(\mathbf{p}_{\mathrm{approx}},\phi)\geq\exp\left(-O\left(\sqrt{n}\log n\right)\right)\mathbb{P}(\mathbf{p}_{\mathrm{pml}},\phi)\;. \quad (100)$$

Therefore $\mathbf{p}_{\mathrm{approx}}$ is an $\exp\left(-O\left(\sqrt{n}\log n\right)\right)$-approximate PML distribution.

In the remainder of the proof we argue about the running time of our final algorithm for approximate PML. Step 4 of the algorithm, that is the convex program $\arg\max_{\mathbf{A}\in\mathbf{Z}_{\mathbf{R}}^{\phi,frac}}\log\mathbf{g}(\mathbf{A})$ can be solved in $\widetilde{O}(k^2\ell)$ time (See Theorem C.11). Algorithm 2 (CreateNewProbabilityValues) and Algorithm 1 (StructuredRounding) can be implemented in $\widetilde{O}(k\ell)$ and $\widetilde{O}(k^2)$ time respectively; therefore, the Algorithm 3 (Rounding algorithm) can be implemented in $\widetilde{O}(k\ell)$ time. Combining everything together our final algorithm (Algorithm 4) can be implemented in $\widetilde{O}(k^2\ell)$ time. Further using $k,\ell\in O(\sqrt{n})$, we conclude the proof. ■

C.4. Missing Proofs from Appendix C.2

Here we provide the proofs for all the lemmas and theorems in Appendix C.2

Proof [Proof of Lemma C.12] Without loss of generality assume $w_0 \geq w_1 \geq w_2 \cdots \geq w_k$. Let $a \stackrel{\text{def}}{=} \sum_{j \in [0,k]} x_j$, we invoke Algorithm 1 with inputs (x, w, a). Let $s \in \mathbb{Z}_+^a$ and $z \in \mathbb{R}^{[0,k] \times [0,k]}$ be the output of Algorithm 1. We now provide the proof for the three conditions in the lemma.

Condition 1: By construction of Algorithm 1, for any $s \in \{s_i\}_{i \in [1,a]}$ we have $\sum_{j \in [0,k]} z_{s,j} = 1$ (Line 6) and for any other $s \in [0,k] \backslash \{s_i\}_{i \in [1,a]}$ we have $\sum_{j \in [0,k]} z_{s,j} = 0$. Therefore the first part of condition 1 holds.

For any $j \in [0,k]$, one of the following two cases holds,

1. If $j \in \{s_i\}_{i \in [1,a]}$ and in this case let $i \in [1,a]$ be such that $s_i = j$. By line 6 (third case) of the algorithm we have,

$$z_{s_{i-1},j} = 1 - \left(\sum_{j' \leq s_{i-1}} x_{j'} - (i-2) + \sum_{s_{i-1} < j' < s_i} x_{j'} \right) = (i-1) - \sum_{j' < s_i} x_{j'} \ . \tag{101}$$

We now analyze the term $\sum_{i' \in [0,k]} z_{i',j}$,

$$\sum_{i' \in [0,k]} z_{i',j} = z_{s_i,j} + z_{s_{i-1},j} = \sum_{j' \leq s_i} x_{j'} - (i-1) + (i-1) - \sum_{j' < s_i} x_{j'} = x_{s_i} = x_j \ .$$

The first equality follows because for $i' \in [0,k] \backslash \{s_i, s_{i-1}\}$ we have $z_{i',j} = 0$ and this follows by the second and third case in line 6 of the algorithm. In the second equality we substituted values for z_{s_i,s_i} and z_{s_{i-1},s_i} using second case (Line 6) and Equation (101) respectively.

2. Else $j \in [0,k] \backslash \{s_i\}_{i \in [1,a]}$, and in this case let $i \in [1,a]$ be such that $s_i < j < s_{i+1}$. Then by the first case in line 6 of the algorithm we have,

$$\sum_{i' \in [0,k]} z_{i',j} = z_{s_i,j} = x_j \ .$$

Condition 2: Consider $\sum_{i \in [0,k]} \left(\sum_{j \in [0,k]} z_{i,j} \right) w_i$,

$$\begin{aligned}
\sum_{i \in [0,k]} \left(\sum_{j \in [0,k]} z_{i,j} \right) w_i &= \sum_{i \in [1,a]} \left(\sum_{s_i \leq j \leq s_{i+1}} z_{s_i,j} \right) w_{s_i} \leq \sum_{i \in [1,a]} \sum_{s_i \leq j \leq s_{i+1}} z_{s_i,j} (w_j + w_{s_i} - w_{s_{i+1}}) \\
&\leq \sum_{i \in [1,a]} \sum_{s_i \leq j \leq s_{i+1}} z_{s_i,j} w_j + \sum_{i \in [1,a]} \sum_{s_i \leq j \leq s_{i+1}} z_{s_i,j} (w_{s_i} - w_{s_{i+1}}) \\
&= \sum_{i \in [1,a]} \sum_{j \in [0,k]} z_{s_i,j} w_j + \sum_{i \in [1,a]} \sum_{s_i \leq j \leq s_{i+1}} z_{s_i,j} (w_{s_i} - w_{s_{i+1}}) \ .
\end{aligned} \tag{102}$$

The first equality follows because rest of the other entries are zero. In the second inequality we used $j \leq s_{i+1}$ and therefore $w_j \geq w_{s_{i+1}}$ by our assumption at the beginning of the proof. In the remainder, we simplify both the terms. Consider the first term in the final expression above,

$$\sum_{i \in [1,a]} \sum_{j \in [0,k]} z_{s_i,j} w_j = \sum_{j \in [0,k]} w_j \sum_{i \in [1,a]} z_{s_i,j} = \sum_{j \in [0,k]} w_j x_j \ . \tag{103}$$

In the first equality we interchanged the summations. In the second equality we used $\sum_{i\in[1,a]} z_{s_i,j} = \sum_{i'\in[0,k]} z_{i',j}$ and further invoked condition 1 of the lemma. Now consider the second term in the final expression of Equation (102),

$$\sum_{i\in[1,a]} \sum_{s_i \leq j \leq s_{i+1}} z_{s_i,j}(w_{s_i} - w_{s_{i+1}}) = \sum_{i\in[1,a]} (w_{s_i} - w_{s_{i+1}}) \sum_{s_i \leq j \leq s_{i+1}} z_{s_i,j} = \sum_{i\in[1,a]} (w_{s_i} - w_{s_{i+1}}) \\ = (w_{s_1} - w_{s_{x+1}}) \leq \max_{j\in[0,k]} w_j \ . \tag{104}$$

The second equality follows by line 6 of the algorithm. Condition 2 follows by combining equations 102, 103 and 104.

Condition 3: First we show that $z_{i,j} > 0$ implies $i \leq j$. Consider $j \in [0,k]$,

1. If $j \in \{s_i\}_{i\in[1,a]}$, in this case let $i \in [1,a]$ be such that $s_i = j$. Then by the second and third case in line 6 of the algorithm we have, $z_{i',j} > 0$ implies $i' \in \{s_i, s_{i-1}\}$. Further, using $s_{i-1} < s_i$ and $s_i = j$ we have $i' \leq j$.

2. Else $j \in [0,k] \backslash \{s_i\}_{i\in[1,a]}$ and in this case let $i \in [1,a]$ be such that $s_i < j < s_{i+1}$. Then by the first case in line 6 of the algorithm we have, $z_{i',j} > 0$ implies $i' = s_i$. Further, using $s_i < j$ we have $i' < j$.

Using the above implication we have,

$$\prod_{j\in[0,k]} w_j^{m_j x_j} = \prod_{j\in[0,k]} w_j^{m_j \sum_{i\in[0,k]} z_{i,j}} = \prod_{i\in[0,k]} \prod_{j\in[0,k]} w_j^{m_j z_{i,j}} \leq \prod_{i\in[0,k]} \prod_{j\in[0,k]} w_i^{m_j z_{i,j}} \tag{105}$$

In the first equality we used $x_j = \sum_{i\in[0,k]} z_{i,j}$ for all $j \in [0,k]$ (Condition 1). In the final inequality, we used the result $z_{i,j} > 0$ implies $i \leq j$ and further combined it with the assumption at the beginning of the proof, that is, $w_i \geq w_j$ for all $i, j \in [0,k]$ and $i \leq j$. ■

Proof [Proof of Lemma C.13] Define $\phi_0 \stackrel{\text{def}}{=} \sum_{i\in[1,t]} \mathbf{B}_{i,0}$. In the following we provide the proof for each case.

Condition 1: For each $i \in [1,t]$, $\mathbf{B}'_{i,j} = \mathbf{C}_{i,j}$ for all $j \in [0,k]$ and the first condition holds.

Condition 2: Note that $\mathbf{B}'$ is initialized to a zero matrix (Line 4). Further for any $i \in [t+1, t+(k+1)]$ let $j \in [0,k]$ be such that $i = t+1+j$, then the algorithm only updates the $\mathbf{B}'_{t+1+j,j}$'th entry in the i'th row and keeps rest of the entries unchanged. Therefore the second condition holds.

Condition 3: For each $i \in [t+1, t+(k+1)]$ let $j \in [0,k]$ be such that $i = t+1+j$, then $\sum_{j'\in[0,k]} \mathbf{B}'_{i,j'} = \mathbf{B}'_{t+1+j,j} = \sum_{i'\in[1,t]} (\mathbf{B}_{i',j} - \mathbf{C}_{i',j}) = \phi_j - \sum_{i'\in[1,t]} \mathbf{C}_{i',j}$. The first equality holds because of the Condition 2. The third equality follows from the Line 8 of the algorithm. The last equality holds because $\mathbf{B} \in \mathbf{Z}_{\mathbf{R}}^{\phi,frac}$ and we have $\sum_{i\in[1,\ell]} \mathbf{B}_{i,j} = \phi_j$.

Condition 4: Here we provide the proof for $\mathbf{B}' \in \mathbf{Z}_{\mathbf{R}'}^{\phi,frac}$. For any $j \in [0,k]$, we first show that $\sum_{i\in[1,t+(k+1)]} \mathbf{B}'_{i,j} = \phi_j$.

$$\sum_{i\in[1,t+(k+1)]} \mathbf{B}'_{i,j} = \sum_{i\in[1,t]} \mathbf{B}'_{i,j} + \sum_{i\in[t+1,t+(k+1)]} \mathbf{B}'_{i,j} = \sum_{i\in[1,t]} \mathbf{C}_{i,j} + \mathbf{B}'_{t+1+j,j} \\ = \sum_{i\in[1,t]} \mathbf{C}_{i,j} + \phi_j - \sum_{i\in[1,t]} \mathbf{C}_{i,j} = \phi_j$$

The second equality follows because $\mathbf{B}'_{i,j} = \mathbf{C}_{i,j}$ for all $i \in [1,t]$ and $j \in [0,k]$ (Line 6) and $\sum_{i \in [t+1,t+(k+1)]} \mathbf{B}'_{i,j} = \mathbf{B}'_{t+1+j,j}$ (Condition 2). The third equality follows from the Condition 3.

We next show that $\sum_{i \in [1,t+(k+1)]} \mathbf{r}_i \left(\sum_{j \in [0,k]} \mathbf{B}'_{i,j}\right) \leq 1$.

$$\begin{aligned}
\sum_{i \in [1,t+(k+1)]} \mathbf{r}_i \left(\sum_{j \in [0,k]} \mathbf{B}'_{i,j}\right) &= \sum_{i \in [1,t]} \mathbf{r}_i \left(\sum_{j \in [0,k]} \mathbf{B}'_{i,j}\right) + \sum_{j \in [0,k]} \mathbf{r}_{t+1+j} \mathbf{B}'_{t+1+j,j} \\
&= \sum_{i \in [1,t]} \mathbf{r}_i \left(\sum_{j \in [0,k]} \mathbf{C}_{i,j}\right) + \sum_{j \in [0,k]} \frac{\sum_{i \in [1,t]} (\mathbf{B}_{ij} - \mathbf{C}_{ij}) \mathbf{r}_i}{\sum_{i \in [1,t]} (\mathbf{B}_{ij} - \mathbf{C}_{ij})} \left(\sum_{i \in [1,t]} (\mathbf{B}_{i,j} - \mathbf{C}_{i,j})\right) \\
&= \sum_{i \in [1,t]} \mathbf{r}_i \left(\sum_{j \in [0,k]} \mathbf{C}_{i,j}\right) + \sum_{j \in [0,k]} \sum_{i \in [1,t]} (\mathbf{B}_{ij} - \mathbf{C}_{ij}) \mathbf{r}_i \\
&= \sum_{i \in [1,t]} \mathbf{r}_i \left(\sum_{j \in [0,k]} \mathbf{B}_{i,j}\right) \leq 1
\end{aligned} \tag{106}$$

In the first equality, we divided the summation into two parts and for the second part we used Condition 3. In the second equality we used Line 7 and 8 of the algorithm. In the third and fourth equality we simplified the expression. In the final inequality we used $\mathbf{B} \in \mathbf{Z}_{\mathbf{R}}^{\phi,frac}$.

Combining all the conditions together we have $\mathbf{B}' \in \mathbf{Z}_{\mathbf{R}'}^{\phi,frac}$. In the remainder we show that $\sum_{i \in [1,t+(k+1)]} \sum_{j \in [0,k]} \mathbf{B}'_{i,j} = \sum_{i \in [1,t]} \sum_{j \in [0,k]} \mathbf{B}_{i,j}$.

Recall we already showed that $\sum_{i \in [1,t+(k+1)]} \mathbf{B}'_{i,j} = \phi_j$ for all $j \in [0,k]$. Recall $\phi_0 = \sum_{i \in [1,t]} \mathbf{B}_{i,0}$ and $\mathbf{B} \in \mathbf{Z}_{\mathbf{R}}^{\phi,frac}$ implies $\phi_j = \sum_{i \in [1,t]} \mathbf{B}_{i,j}$ for all $j \in [1,k]$. Therefore we have,

$$\sum_{i \in [1,t+(k+1)]} \sum_{j \in [0,k]} \mathbf{B}'_{i,j} = \sum_{i \in [1,t]} \sum_{j \in [0,k]} \mathbf{B}_{i,j}$$

Condition 5: We first provide the explicit expressions for $\mathbf{g}(\mathbf{B}')$ and $\mathbf{g}(\mathbf{B})$ below:

$$\mathbf{g}(\mathbf{B}') = \left(\prod_{i \in [1,t]} \mathbf{r}_i^{(\mathbf{B}'\vec{\mathrm{m}})_i} \frac{\exp\left((\mathbf{B}'\vec{1})_i \log(\mathbf{B}'\vec{1})_i\right)}{\prod_{j \in [0,k]} \exp\left(\mathbf{B}'_{ij} \log \mathbf{B}'_{ij}\right)}\right) \left(\prod_{j \in [0,k]} \mathbf{r}_{t+1+j}^{\vec{\mathrm{m}}_j \mathbf{B}'_{t+1+j,j}} \cdot 1\right)$$

$$\mathbf{g}(\mathbf{B}) = \prod_{i \in [1,t]} \left(\mathbf{r}_i^{(\mathbf{B}\vec{\mathrm{m}})_i} \frac{\exp\left((\mathbf{B}\vec{1})_i \log(\mathbf{B}\vec{1})_i\right)}{\prod_{j \in [0,k]} \exp\left(\mathbf{B}_{ij} \log \mathbf{B}_{ij}\right)}\right)$$

Note that in the expression for $\mathbf{g}(\mathbf{B}')$ we used Condition 2. In the above two definitions for $\mathbf{g}(\mathbf{B}')$ and $\mathbf{g}(\mathbf{B})$, we refer to the expression involving $\mathbf{r}_i$'s as the probability term and the rest as the counting term. We start the analysis of Condition 5 by first bounding the probability term:

$$\begin{aligned}
\prod_{i\in[1,t]}\mathbf{r}_i^{(\mathbf{B}\overrightarrow{\mathrm{m}})_i} &= \left(\prod_{i\in[1,t]}\mathbf{r}_i^{(\mathbf{B}'\overrightarrow{\mathrm{m}})_i}\right)\left(\prod_{i\in[1,t]}\mathbf{r}_i^{\sum_{j\in[0,k]}\overrightarrow{\mathrm{m}}_j(\mathbf{B}_{ij}-\mathbf{B}'_{ij})}\right) = \left(\prod_{i\in[1,t]}\mathbf{r}_i^{(\mathbf{B}'\overrightarrow{\mathrm{m}})_i}\right)\left(\prod_{j\in[0,k]}\prod_{i\in[1,t]}\mathbf{r}_i^{\overrightarrow{\mathrm{m}}_j(\mathbf{B}_{ij}-\mathbf{B}'_{ij})}\right)\\
&= \left(\prod_{i\in[1,t]}\mathbf{r}_i^{(\mathbf{B}'\overrightarrow{\mathrm{m}})_i}\right)\left(\prod_{j\in[0,k]}\left(\prod_{i\in[1,t]}\mathbf{r}_i^{(\mathbf{B}_{ij}-\mathbf{B}'_{ij})}\right)^{\overrightarrow{\mathrm{m}}_j}\right) = \left(\prod_{i\in[1,t]}\mathbf{r}_i^{(\mathbf{B}'\overrightarrow{\mathrm{m}})_i}\right)\left(\prod_{j\in[0,k]}\left(\prod_{i\in[1,t]}\mathbf{r}_i^{(\mathbf{B}_{ij}-\mathbf{C}_{ij})}\right)^{\overrightarrow{\mathrm{m}}_j}\right)\\
&\leq \left(\prod_{i\in[1,t]}\mathbf{r}_i^{(\mathbf{B}'\overrightarrow{\mathrm{m}})_i}\right)\left(\prod_{j\in[0,k]}\left(\frac{\sum_{i\in[1,t]}\mathbf{r}_i(\mathbf{B}_{ij}-\mathbf{C}_{ij})}{\sum_{i\in[1,t]}(\mathbf{B}_{ij}-\mathbf{C}_{ij})}\right)^{\overrightarrow{\mathrm{m}}_j\sum_{i\in[1,t]}(\mathbf{B}_{ij}-\mathbf{C}_{ij})}\right)\\
&\leq \left(\prod_{i\in[1,t]}\mathbf{r}_i^{(\mathbf{B}'\overrightarrow{\mathrm{m}})_i}\right)\left(\prod_{j\in[0,k]}\mathbf{r}_{t+1+j}^{\overrightarrow{\mathrm{m}}_j\mathbf{B}'_{t+1+j,j}}\right)
\end{aligned}\tag{107}$$

The first three inequalities simplify the expression. The fourth equality follows because $\mathbf{B}'_{i,j}=\mathbf{C}_{i,j}$ for all $i\in[1,t]$ and $j\in[0,k]$. The fifth inequality follows from AM-GM inequality. The final expression above is the probability term associated with $\mathbf{B}'$ and the equation above shows that our rounding procedure only increases the probability term and it remains to bound the counting term.

$$\begin{aligned}
\frac{\mathbf{g}(\mathbf{B}')}{\mathbf{g}(\mathbf{B})} &\geq \prod_{i\in[1,t]}\frac{\exp\left((\mathbf{B}'\overrightarrow{1})_i\log(\mathbf{B}'\overrightarrow{1})_i-(\mathbf{B}\overrightarrow{1})_i\log(\mathbf{B}\overrightarrow{1})_i\right)}{\prod_{j\in[0,k]}\exp\left(\mathbf{B}'_{ij}\log\mathbf{B}'_{ij}-\mathbf{B}_{ij}\log\mathbf{B}_{ij}\right)}\\
&= \prod_{i\in[1,t]}\frac{\exp\left((\mathbf{C}\overrightarrow{1})_i\log(\mathbf{C}\overrightarrow{1})_i-(\mathbf{B}\overrightarrow{1})_i\log(\mathbf{B}\overrightarrow{1})_i\right)}{\prod_{j\in[0,k]}\exp\left(\mathbf{C}_{ij}\log\mathbf{C}_{ij}-\mathbf{B}_{ij}\log\mathbf{B}_{ij}\right)}\,.
\end{aligned}\tag{108}$$

Consider the numerator in the above expression, for each $i\in[1,t]$ let $s_i\overset{\mathrm{def}}{=}(\mathbf{C}\overrightarrow{1})_i$, then

$$\begin{aligned}
\prod_{i\in[1,t]}\exp\left((\mathbf{C}\overrightarrow{1})_i\log(\mathbf{C}\overrightarrow{1})_i-(\mathbf{B}\overrightarrow{1})_i\log(\mathbf{B}\overrightarrow{1})_i\right) &= \prod_{i\in[1,t]}\exp\left(s_i\log s_i-(s_i+\alpha_i)\log(s_i+\alpha_i)\right)\\
&= \prod_{i\in[1,t]}\exp\left(s_i\log\frac{s_i}{s_i+\alpha_i}-\alpha_i\log(s_i+\alpha_i)\right)\\
&\geq \prod_{i\in[1,t]}\exp\left(s_i\frac{-\alpha_i}{s_i}-\alpha_i\log(s_i+\alpha_i)\right)\\
&\geq \exp\left(-O\left(\log(\sum_{i\in[1,t]}s_i)\sum_{i\in[1,t]}\alpha_i\right)\right)\\
&\geq \exp\left(-O\left(\sum_{i\in[1,t]}\alpha_i\log\Delta\right)\right)\,.
\end{aligned}\tag{109}$$

In the third inequality we used $\log(1+x) \geq \frac{x}{1+x}$ for all $x \geq -1$. The final inequality follows because $\sum_{i\in[1,t]} s_i \leq \sum_{i\in[1,t]}(\mathbf{B}\overrightarrow{1})_i \leq \Delta$. Now consider the denominator in the above expression, let $\alpha_{i,j} = \mathbf{B}_{i,j} - \mathbf{C}_{i,j}$ for all $i \in [1,t]$ and $j \in [0,k]$, then

$$\begin{aligned}
\prod_{i\in[1,t]}\prod_{j\in[0,k]} \exp\left(\mathbf{C}_{ij}\log\mathbf{C}_{ij} - \mathbf{B}_{ij}\log\mathbf{B}_{ij}\right) &= \prod_{i\in[1,t]}\prod_{j\in[0,k]} \exp\left(\mathbf{C}_{ij}\log\mathbf{C}_{ij} - (\mathbf{C}_{ij}+\alpha_{i,j})\log(\mathbf{C}_{ij}+\alpha_{i,j})\right) \\
&= \prod_{i\in[1,t]}\prod_{j\in[0,k]} \exp\left(\mathbf{C}_{ij}\log\frac{\mathbf{C}_{ij}}{\mathbf{C}_{ij}+\alpha_{i,j}} - \alpha_{i,j}\log(\mathbf{C}_{ij}+\alpha_{i,j})\right) \\
&\leq \prod_{i\in[1,t]}\prod_{j\in[0,k]} \exp\left(-\alpha_{i,j}\log(\mathbf{C}_{ij}+\alpha_{i,j})\right) \\
&\leq \prod_{i\in[1,t]}\prod_{j\in[0,k]} \exp\left(-\alpha_{i,j}\log\alpha_{i,j}\right) \leq \exp\left(O\big(\log(t\times k)\sum_{i\in[1,t]}\alpha_i\big)\right) \\
&\leq \exp\left(O\left(\sum_{i\in[1,t]}\alpha_i\log\Delta\right)\right) .
\end{aligned} \tag{110}$$

In the third inequality we used $\alpha_{i,j} \geq 0$ and therefore $\mathbf{C}_{ij}\log\frac{\mathbf{C}_{ij}}{\mathbf{C}_{ij}+\alpha_{i,j}} \leq 0$. In the fourth inequality we used $\log(\mathbf{C}_{ij}+\alpha_{i,j}) \geq \log\alpha_{i,j}$. In the fifth inequality we used $\sum_{j\in[0,k]}\alpha_{i,j} = \alpha_i$ for all $i \in [1,t]$ and further $\sum_{i\in[1,t]}\sum_{j\in[0,k]} -\alpha_{i,j}\log\alpha_{i,j} = \sum_{i\in[1,t]}\alpha_i\left(\sum_{j\in[0,k]} -\frac{\alpha_{i,j}}{\alpha_i}\log\frac{\alpha_{i,j}}{\alpha_i} - \log\alpha_i\right) \leq \log(k+1)\sum_{i\in[1,t]}\alpha_i - \sum_{i\in[1,t]}\alpha_i\log\alpha_i$. Now consider the term $-\sum_{i\in[1,t]}\alpha_i\log\alpha_i$ and note that $-\sum_{i\in[1,t]}\alpha_i\log\alpha_i = (\sum_{i\in[1,t]}\alpha_i)\left(-\sum_{i\in[1,t]}\frac{\alpha_i}{\sum_{i\in[1,t]}\alpha_i}\log\frac{\alpha_i}{\sum_{i\in[1,t]}\alpha_i} - \log\sum_{i\in[1,t]}\alpha_i\right) \leq (1+\log t)\sum_{i\in[1,t]}\alpha_i$. The fifth inequality in Equation (110) follows by combining the previous two derivations together. The final inequality follows because $t \times k \leq \Delta$.

Condition 6: This condition follows immediately from Line 7 of the algorithm. ■

Proof [Proof of Lemma C.14] In the following we provide the proof for the claims in the lemma.

Condition 1: Note that $\mathbf{H}^{(1)} \subseteq \mathbf{H} \cup [\ell+1, \ell+(k+1)]$, where $[\ell+1, \ell+(k+1)]$ are the indices corresponding to the new rows created by the procedure $\mathrm{CreateNewProbabilityValues}$ (Algorithm 2). Consider any $i \in \mathbf{H}^{(1)}$, then one the following two cases hold,

1. If $i \in \mathbf{H}$, then by the first condition of Lemma C.13 we have $(\mathbf{S}^{(1)}\overrightarrow{1})_i = (\mathbf{A}\overrightarrow{1})_i = \sum_{j\in[0,k]}\mathbf{A}_{i,j} = \sum_{j\in[0,k]}\lfloor\mathbf{S}_{i,j}\rfloor \in \mathbb{Z}_+$.
2. Else $i \in [\ell+1, \ell+(k+1)]$ and in this case we have $\sum_{i\in[1,\ell]}\mathbf{A}_{i,j} = \sum_{i\in\mathbf{H}}\mathbf{A}_{i,j} + \sum_{i\in\mathbf{L}}\mathbf{A}_{i,j} = \sum_{i\in\mathbf{H}}\lfloor\mathbf{S}_{i,j}\rfloor + \lfloor\sum_{i\in\mathbf{L}}\mathbf{S}_{i,j}\rfloor \in \mathbb{Z}_+$. The second equality in the previous derivation follows from Line 7 and 8 of the algorithm. The previous derivation combined with third condition of Lemma C.13 we get, $(\mathbf{S}^{(1)}\overrightarrow{1})_i = \phi_j - \sum_{i\in[1,\ell]}\mathbf{A}_{i,j} \in \mathbb{Z}_+$.

$(\mathbf{S}^{(1)}\overrightarrow{1})_i \in \mathbb{Z}_+$ in both the cases and the condition 1 follows.

Condition 2: This condition follows immediately from the fourth condition of Lemma C.13.

Condition 3: Let $\alpha_i = \sum_{j\in[0,k]} \mathbf{S}_{i,j} - \sum_{j\in[0,k]} \mathbf{A}_{i,j}$ for all $i \in [1,\ell]$. First we upper bound the term $\sum_{i\in\mathbf{H}} \alpha_i$. Consider $\sum_{i\in\mathbf{H}} \alpha_i \leq \sum_{i\in\mathbf{H}} \sum_{j\in[0,k]} \mathbf{S}_{i,j} \leq \frac{1}{\gamma}$. The last inequality follows because of the constraint $\sum_{i\in[1,\ell]} \mathbf{r}_i \sum_{j\in[0,k]} \mathbf{S}_{i,j} \leq 1$ ($\mathbf{S} \in \mathbf{Z}^{\phi,frac}_{\mathbf{R}}$) and $\mathbf{r}_i > \gamma$ for all $i \in \mathbf{H}$.

We now upper bound the term $\sum_{i\in\mathbf{L}} \alpha_i$. Consider $\sum_{i\in\mathbf{L}} \alpha_i = \sum_{i\in\mathbf{L}} \left(\sum_{j\in[0,k]} \mathbf{S}_{i,j} - \sum_{j\in[0,k]} \mathbf{A}_{i,j}\right) = \sum_{j\in[0,k]} \left(\sum_{i\in\mathbf{L}} \mathbf{S}_{i,j} - \sum_{i\in\mathbf{L}} \mathbf{A}_{i,j}\right)$. Further $\sum_{i\in\mathbf{L}} \mathbf{A}_{i,j} = \lfloor \sum_{i\in\mathbf{L}} \mathbf{S}_{i,j} \rfloor$ for all $j \in [0,k]$ (Line 8 of the algorithm) and we get $\sum_{i\in\mathbf{L}} \alpha_i \leq k+1$.

Therefore $\sum_{i\in[\ell]} \alpha_i = \sum_{i\in\mathbf{H}} \alpha_i + \sum_{i\in\mathbf{L}} \alpha_i \leq \frac{1}{\gamma} + k + 1$ and combined with fifth condition Lemma C.13 we have,

$$\mathbf{g}(\mathbf{S}^{(1)}) \geq \exp\left(-O\left(\left(\frac{1}{\gamma}+k\right)\log\Delta\right)\right)\mathbf{g}(\mathbf{S})\ .$$

■

Proof [Proof of Lemma C.15] In the following we provide proof for all the conditions in the lemma.

Condition 1: For all $i \in [1, \ell+(k+1)]$, one of the following two conditions hold,

1. If $i \in \mathbf{H}^{(1)}$, then by the first condition of Lemma C.13 we have $(\mathbf{S}^{(2)}\overrightarrow{1})_i = (\mathbf{A}^{(1)}\overrightarrow{1})_i = (\mathbf{S}^{(1)}\overrightarrow{1})_i \in \mathbb{Z}_+$. The last expression follows from first condition of Lemma C.14.
2. Else $i \in \mathbf{L}^{(1)}$, then again by the first condition of Lemma C.13 we have $(\mathbf{S}^{(2)}\overrightarrow{1})_i = (\mathbf{A}^{(1)}\overrightarrow{1})_i = \lfloor(\mathbf{S}^{(1)}\overrightarrow{1})_i\rfloor \in \mathbb{Z}_+$. The last equality follows from Line 15 of the algorithm.

For all $i \in [1, \ell+(k+1)]$, we have $(\mathbf{S}^{(2)}\overrightarrow{1})_i \in \mathbb{Z}_+$ and therefore condition 1 holds.

Condition 2: This condition follows immediately from the second condition of Lemma C.13.

Condition 3: This condition follows immediately from the fourth condition of Lemma C.13.

Condition 4: Consider the term $\sum_{i\in[\ell+(k+1)+1,\ell+2(k+1)]} (\mathbf{S}^{(2)}\overrightarrow{1})_i$,

$$\begin{aligned}
\sum_{i\in[\ell+(k+1)+1,\ell+2(k+1)]} (\mathbf{S}^{(2)}\overrightarrow{1})_i &= \sum_{i\in[1,\ell+2(k+1)]} (\mathbf{S}^{(2)}\overrightarrow{1})_i - \sum_{i\in[1,\ell+(k+1)]} (\mathbf{S}^{(2)}\overrightarrow{1})_i \\
&= \sum_{j\in[0,k]} \phi_j - \sum_{i\in[1,\ell+(k+1)]} (\mathbf{A}^{(1)}\overrightarrow{1})_i \\
&= \sum_{j\in[0,k]} \phi_j - \left(\sum_{i\in\mathbf{H}^{(1)}} (\mathbf{A}^{(1)}\overrightarrow{1})_i + \sum_{i\in\mathbf{L}^{(1)}} (\mathbf{A}^{(1)}\overrightarrow{1})_i\right) \\
&= \sum_{j\in[0,k]} \phi_j - \left(\sum_{i\in\mathbf{H}^{(1)}} (\mathbf{S}^{(1)}\overrightarrow{1})_i + \sum_{i\in\mathbf{L}^{(1)}} \lfloor(\mathbf{S}^{(1)}\overrightarrow{1})_i\rfloor\right) \in \mathbb{Z}_+
\end{aligned} \tag{111}$$

In the first equality we add and subtract $\sum_{i\in[1,\ell+(k+1)]} (\mathbf{S}^{(2)}\overrightarrow{1})_i$ term. The first term in the second equality follows because $\sum_{i\in[1,\ell+2(k+1)]} (\mathbf{S}^{(2)}\overrightarrow{1})_i = \sum_{j\in[0,k]} \sum_{i\in[1,\ell+2(k+1)]} \mathbf{S}^{(2)}_{i,j} = \sum_{j\in[0,k]} \phi_j$ and the last equality follows because $\mathbf{S}^{(2)} \in \mathbf{Z}^{\phi,frac}_{\mathbf{R}^{(2)}}$ (Condition 3). The second term in the second equality follows by the first condition of Lemma C.13. In the third equality we divided the summation terms over $\mathbf{H}^{(1)}$ and $\mathbf{L}^{(1)}$. In the fourth equality we used Line 14 of the algorithm and further for any $i \in \mathbf{L}^{(1)}$ Line 15 implies $(\mathbf{A}^{(1)}\overrightarrow{1})_i = \sum_{j\in[0,k]} \mathbf{S}^{(1)}_{ij}\frac{\lfloor(\mathbf{S}^{(1)}\overrightarrow{1})_i\rfloor}{(\mathbf{S}^{(1)}\overrightarrow{1})_i} = \lfloor(\mathbf{S}^{(1)}\overrightarrow{1})_i\rfloor$. Finally by first condition of Lemma C.14 we have $(\mathbf{S}^{(1)}\overrightarrow{1})_i \in \mathbb{Z}_+$ for all $i \in \mathbf{H}^{(1)}$ and $\phi_j \in \mathbb{Z}_+$ for all $j \in [0,k]$. Therefore, $\sum_{i\in[\ell+(k+1)+1,\ell+2(k+1)]}(\mathbf{S}^{(2)}\overrightarrow{1})_i \in \mathbb{Z}_+$ and the condition 4 holds.

Condition 5: For any $j \in [0,k]$ we have,

$$\begin{aligned}\mathbf{r}^{(2)}_{\ell+(k+1)+1+j} &= \frac{\sum_{i\in[1,\ell+(k+1)]}(\mathbf{S}^{(1)}_{ij}-\mathbf{A}^{(1)}_{ij})\mathbf{r}^{(1)}_i}{\sum_{i\in[1,\ell+(k+1)]}(\mathbf{S}^{(1)}_{ij}-\mathbf{A}^{(1)}_{ij})} = \frac{\sum_{i\in\mathbf{L}^{(1)}}(\mathbf{S}^{(1)}_{ij}-\mathbf{A}^{(1)}_{ij})\mathbf{r}^{(1)}_i}{\sum_{i\in\mathbf{L}^{(1)}}(\mathbf{S}^{(1)}_{ij}-\mathbf{A}^{(1)}_{ij})} \\ &\leq \gamma\frac{\sum_{i\in\mathbf{L}^{(1)}}(\mathbf{S}^{(1)}_{ij}-\mathbf{A}^{(1)}_{ij})}{\sum_{i\in\mathbf{L}^{(1)}}(\mathbf{S}^{(1)}_{ij}-\mathbf{A}^{(1)}_{ij})} \leq \gamma.\end{aligned} \tag{112}$$

The first equality follows from the sixth condition of Lemma C.13. The second equality follows because $\mathbf{S}^{(1)}_{i,j} = \mathbf{A}^{(1)}_{i,j}$ for all $i \in \mathbf{H}^{(1)}$ and $j \in [0,k]$ (Line 14). The third inequality follows because $\mathbf{S}^{(1)}_{i,j} \geq \mathbf{A}^{(1)}_{i,j}$ for all $i \in \mathbf{L}^{(1)}$ and $j \in [0,k]$ (Line 15) and further $\mathbf{r}^{(1)}_i \leq \gamma$ for all $i \in \mathbf{L}^{(1)}$ (Line 12).

Condition 6: For any $i \in [1, \ell+(k+1)]$, let $\alpha_i = \sum_{j\in[0,k]}\mathbf{S}^{(1)}_{i,j} - \sum_{j\in[0,k]}\mathbf{A}^{(1)}$. Note that $\alpha_i = 0$ for all $i \in \mathbf{H}^{(1)}$ (Line 14) and $\alpha_i = (\mathbf{S}^{(1)}\overrightarrow{1})_i - \lfloor(\mathbf{S}^{(1)}\overrightarrow{1})_i\rfloor \leq 1$ for all $i \in \mathbf{L}^{(1)}$ (Line 15). Therefore $\sum_{i\in[1,\ell+(k+1)]}\alpha_i \leq |\mathbf{L}^{(1)}| \leq \ell+(k+1)$ and further combined with the fifth condition of Lemma C.13 we have $\mathbf{g}(\mathbf{S}^{(2)}) \geq \exp\left(-O\left((\ell+k)\log\Delta\right)\right)\mathbf{g}(\mathbf{S}^{(1)})$. Note that by the third condition of Lemma C.14 we have $\mathbf{g}(\mathbf{S}^{(1)}) \geq \exp\left(-O\left(\left(\frac{1}{\gamma}+k\right)\log\Delta\right)\right)\mathbf{g}(\mathbf{S})$. Combining the previous two inequalities we get $\mathbf{g}(\mathbf{S}^{(2)}) \geq \exp\left(-O\left((\ell+k+\frac{1}{\gamma})\log\Delta\right)\right)\mathbf{g}(\mathbf{S})$ and condition 6 holds. ■

Proof [Proof of Theorem C.16] In the following we provide proof for the two conditions of the theorem.

Condition 1: Here we provide the proof for the condition $\mathbf{S}^{\text{ext}} \in \mathbf{Z}^{\phi}_{\mathbf{R}^{\text{ext}}}$.

1. For all $i \in [1, \ell+2(k+1)]$, consider $(\mathbf{S}^{\text{ext}}\overrightarrow{1})_i$. If $i \in [1,\ell+(k+1)]$, then $(\mathbf{S}^{\text{ext}}\overrightarrow{1})_i = (\mathbf{S}^{(2)}\overrightarrow{1})_i \in \mathbb{Z}_+$. The first equality follows by line 22 of the algorithm and the last expression follows by first condition of Lemma C.15. Else $i \in [\ell+(k+1)+1, \ell+2(k+1)]$, let j be such that $i = \ell+(k+1)+1+j$, then $(\mathbf{S}^{\text{ext}}\overrightarrow{1})_i = \sum_{j'\in[0,k]}\mathbf{S}^{\text{ext}}_{\ell+(k+1)+1+j,j'} = \sum_{j'\in[0,k]}\left(\lfloor\mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j'}\rfloor + z_{j,j'}\right) = \lfloor\mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j}\rfloor + \sum_{j'\in[0,k]} z_{j,j'} \in \mathbb{Z}_+$. The second equality follows by line 23 of the algorithm. The third equality follows from the second condition of Lemma C.15. Finally by the first condition of Lemma C.12 we have $\sum_{j'\in[0,k]} z_{j,j'} \in \{0,1\}$ for all $j \in [0,k]$ and therefore $(\mathbf{S}^{\text{ext}}\overrightarrow{1})_i \in \mathbb{Z}_+$ for any $i \in [\ell+(k+1)+1, \ell+2(k+1)]$. Combining the analysis of cases $i \in [1,\ell+(k+1)]$ and $i \in [\ell+(k+1)+1,\ell+2(k+1)]$ the condition 1 holds.

2. For all $j \in [0,k]$,

$$\begin{aligned}\sum_{i\in[1,\ell+2(k+1)]}\mathbf{S}^{\text{ext}}_{i,j} &= \sum_{i\in[1,\ell+(k+1)]}\mathbf{S}^{\text{ext}}_{i,j} + \sum_{i\in[\ell+(k+1)+1,\ell+2(k+1)]}\mathbf{S}^{\text{ext}}_{i,j} \\ &= \sum_{i\in[1,\ell+(k+1)]}\mathbf{S}^{(2)}_{i,j} + \sum_{j'\in[0,k]}\left(\lfloor\mathbf{S}^{(2)}_{\ell+(k+1)+1+j',j}\rfloor + z_{j',j}\right).\end{aligned} \tag{113}$$

The second equality follows because $\mathbf{S}^{\text{ext}}_{i,j} = \mathbf{S}^{(2)}_{i,j}$ for all $i \in [1, \ell + (k+1)]$ (Line 22) and $\mathbf{S}^{\text{ext}}_{i,j} = \lfloor \mathbf{S}^{(2)}_{\ell+(k+1)+1+j',j} \rfloor + z_{j',j}$ for all $i \in [\ell + (k+1) + 1, \ell + 2(k+1)]$ (Line 23). We next simplify the second term in the above expression.

$$\sum_{j' \in [0,k]} \left(\lfloor \mathbf{S}^{(2)}_{\ell+(k+1)+1+j',j} \rfloor + z_{j',j} \right) = \lfloor \mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j} \rfloor + \sum_{j' \in [0,k]} z_{j',j} = \lfloor \mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j} \rfloor + x_j$$
$$= \mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j} = \sum_{i \in [\ell+(k+1)+1, \ell+2(k+1)]} \mathbf{S}^{(2)}_{i,j} . \tag{114}$$

In the first and final equality we used the second condition of Lemma C.15 (Diagonal Structure). In the second equality we used the first condition of Lemma C.12. In the third equality we used the definition of x_j (Line 19). Combining equations 113 and 114 we get,

$$\sum_{i \in [1, \ell+2(k+1)]} \mathbf{S}^{\text{ext}}_{i,j} = \sum_{i \in [1, \ell+2(k+1)]} \mathbf{S}^{(2)}_{i,j} = \phi_j$$

In the last inequality we used $\mathbf{S}^{(2)} \in \mathbf{Z}^{\phi, frac}_{\mathbf{R}^{(2)}}$.

3. Let $\mathbf{r}^{\text{ext}}_i$ for all $i \in [1, \ell+2(k+1)]$ be the i'th element of $\mathbf{R}^{\text{ext}}$. Consider $\sum_{i \in [1, \ell+2(k+1)]} \mathbf{r}^{\text{ext}}_i (\mathbf{S}^{\text{ext}} \overrightarrow{1})_i$, we have,

$$\sum_{i \in [1, \ell+2(k+1)]} \mathbf{r}^{\text{ext}}_i (\mathbf{S}^{\text{ext}} \overrightarrow{1})_i = \sum_{i \in [1, \ell+2(k+1)]} \frac{\mathbf{r}^{(2)}_i}{1+\gamma} (\mathbf{S}^{\text{ext}} \overrightarrow{1})_i$$
$$= \frac{1}{1+\gamma} \sum_{i \in [1, \ell+(k+1)+1]} \mathbf{r}^{(2)}_i (\mathbf{S}^{(2)} \overrightarrow{1})_i + \frac{1}{1+\gamma} \sum_{i \in [\ell+(k+1)+1, \ell+2(k+1)]} \mathbf{r}^{(2)}_i (\mathbf{S}^{\text{ext}} \overrightarrow{1})_i . \tag{115}$$

The first equality follows from Line 24 of the algorithm. In the second equality we divided the summation into two parts and used $\mathbf{S}^{\text{ext}}_{i,j} = \mathbf{S}^{(2)}_{i,j}$ for all $i \in [1, \ell + (k+1) + 1]$ and $j \in [0, k]$ (Line 22) for the first part. We now simplify the second part of the above expression.

$$\sum_{i \in [\ell+(k+1)+1, \ell+2(k+1)]} \mathbf{r}^{(2)}_i (\mathbf{S}^{\text{ext}} \overrightarrow{1})_i = \sum_{j \in [0,k]} \mathbf{r}^{(2)}_{\ell+(k+1)+1+j} \sum_{j' \in [0,k]} \left(\lfloor \mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j'} \rfloor + z_{j,j'} \right)$$
$$= \sum_{j \in [0,k]} w_j \left(\mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j} - x_j \right) + \sum_{j \in [0,k]} w_j \sum_{j' \in [0,k]} z_{j,j'}$$
$$\leq \sum_{j \in [0,k]} w_j \left(\mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j} - x_j \right) + \sum_{j \in [0,k]} w_j x_j + \max_{j \in [0,k]} w_j$$
$$= \sum_{i \in [\ell+(k+1)+1, \ell+2(k+1)]} \mathbf{r}^{(2)}_i (\mathbf{S}^{(2)} \overrightarrow{1})_i + \gamma . \tag{116}$$

In the first equality we expanded the $(\mathbf{S}^{\text{ext}}\overrightarrow{1})_i$ term. Further we used $\mathbf{S}^{\text{ext}}_{\ell+(k+1)+1+j,j'} = \lfloor \mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j'} \rfloor + z_{j,j'}$ for all $j, j' \in [0,k]$ (Line 23). In the second equality we used the second condition of Lemma C.15 (Diagonal Structure) and further combined it with definitions of w_j and x_j from Line 19 of the algorithm. The third inequality follows from second condition of Lemma C.12. In the final inequality we used $\max_{j \in [0,k]} w_j \leq \gamma$ that follows from the definition of w_j and fifth condition of Lemma C.15. Further we combined it with $\mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j} = (\mathbf{S}^{(2)}\overrightarrow{1})_i$ that follows from the second condition of Lemma C.15.

Combining equations 115 and 116 we have,

$$\sum_{i \in [1,\ell+2(k+1)]} \mathbf{r}^{\text{ext}}_i (\mathbf{S}^{\text{ext}}\overrightarrow{1})_i \leq \frac{1}{1+\gamma}\left(\sum_{i \in [1,\ell+2(k+1)]} \mathbf{r}^{(2)}_i (\mathbf{S}^{(2)}\overrightarrow{1})_i + \gamma\right) \leq 1\,.$$

In the final inequality we used $\mathbf{S}^{(2)} \in \mathbf{Z}^{\phi,frac}_{\mathbf{R}^{(2)}}$ and therefore $\sum_{i \in [1,\ell+2(k+1)]} \mathbf{r}^{(2)}_i (\mathbf{S}^{(2)}\overrightarrow{1})_i \leq 1$.

The condition 1 holds by combining the analysis of all the above three cases.

Condition 2: Recall the definition of $\mathbf{g}(\mathbf{S}^{\text{ext}})$,

$$\mathbf{g}(\mathbf{S}^{\text{ext}}) = \prod_{i \in [1,\ell+2(k+1)]} \left(\mathbf{r}_i^{\text{ext}\,(\mathbf{S}^{\text{ext}}\overrightarrow{\mathrm{m}})_i} \frac{\exp\left((\mathbf{S}^{\text{ext}}\overrightarrow{1})_i \log(\mathbf{S}^{\text{ext}}\overrightarrow{1})_i\right)}{\prod_{j \in [0,k]} \exp\left(\mathbf{S}^{\text{ext}}_{ij} \log \mathbf{S}^{\text{ext}}_{ij}\right)} \right)$$

In the above expression consider the probability term,

$$\begin{aligned}
\prod_{i \in [1,\ell+2(k+1)]} \mathbf{r}_i^{\text{ext}\,(\mathbf{S}^{\text{ext}}\overrightarrow{\mathrm{m}})_i} &= \prod_{i \in [1,\ell+2(k+1)]} \left(\frac{\mathbf{r}^{(2)}_i}{1+\gamma}\right)^{(\mathbf{S}^{\text{ext}}\overrightarrow{\mathrm{m}})_i} \\
&\geq \exp\left(-O(\gamma n)\right) \left(\prod_{i \in [1,\ell+(k+1)]} \mathbf{r}_i^{(2)\,(\mathbf{S}^{\text{ext}}\overrightarrow{\mathrm{m}})_i} \right) \left(\prod_{i \in [\ell+(k+1)+1,\ell+2(k+1)]} \mathbf{r}_i^{(2)\,(\mathbf{S}^{\text{ext}}\overrightarrow{\mathrm{m}})_i} \right) \\
&= \exp\left(-O(\gamma n)\right) \left(\prod_{i \in [1,\ell+(k+1)]} \mathbf{r}_i^{(2)\,(\mathbf{S}^{(2)}\overrightarrow{\mathrm{m}})_i} \right) \left(\prod_{i \in [\ell+(k+1)+1,\ell+2(k+1)]} \mathbf{r}_i^{(2)\,(\mathbf{S}^{\text{ext}}\overrightarrow{\mathrm{m}})_i} \right).
\end{aligned} \tag{117}$$

In the first equality we used line 24 of the algorithm. In the second inequality we used $\sum_{i \in [1,\ell+2(k+1)]} (\mathbf{S}^{\text{ext}}\overrightarrow{\mathrm{m}})_i = n$ that further implies $(1+\gamma)^{-\sum_{i \in [1,\ell+2(k+1)]} (\mathbf{S}^{\text{ext}}\overrightarrow{\mathrm{m}})_i} \geq \exp\left(-O(\gamma n)\right)$. In the third equality we used $\mathbf{S}^{\text{ext}}_{i,j} = \mathbf{S}^{(2)}_{i,j}$ for all $i \in [1, \ell+(k+1)]$ and $j \in [0,k]$ (Line 22). We now analyze the second product term in the final expression above,

$$\begin{aligned}\prod_{i\in[\ell+(k+1)+1,\ell+2(k+1)]} {\mathbf{r}_i^{(2)}}^{(\mathbf{S}^{\text{ext}}\vec{\mathrm{m}})_i} &= \prod_{j\in[0,k]} {\mathbf{r}_{\ell+(k+1)+1+j}^{(2)}}^{\sum_{j'\in[0,k]}\mathbf{S}^{\text{ext}}_{\ell+(k+1)+1+j,j'}\vec{\mathrm{m}}_{j'}}\\ &= \prod_{j\in[0,k]} {\mathbf{r}_{\ell+(k+1)+1+j}^{(2)}}^{\sum_{j'\in[0,k]}\left(\lfloor\mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j'}\rfloor+z_{j,j'}\right)\vec{\mathrm{m}}_{j'}}\\ &= \left(\prod_{j\in[0,k]} {\mathbf{r}_{\ell+(k+1)+1+j}^{(2)}}^{\lfloor\mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j}\rfloor}\right)\left(\prod_{j\in[0,k]} {\mathbf{r}_{\ell+(k+1)+1+j}^{(2)}}^{\sum_{j'\in[0,k]}z_{j,j'}\vec{\mathrm{m}}_{j'}}\right).\end{aligned} \tag{118}$$

The second equality follows from line 23 of the algorithm. The third equality follows from the second condition of Lemma C.15 (Diagonal Structure).

Now consider the second product term in the above expression.

$$\prod_{j\in[0,k]} {\mathbf{r}_{\ell+(k+1)+1+j}^{(2)}}^{\sum_{j'\in[0,k]}z_{j,j'}\vec{\mathrm{m}}_{j'}} = \prod_{j\in[0,k]} w_j^{\sum_{j'\in[0,k]}z_{j,j'}\vec{\mathrm{m}}_{j'}} \geq \prod_{j\in[0,k]} w_j^{x_j\vec{\mathrm{m}}_j}. \tag{119}$$

In the first equality we used the definition of w_j (Line 19). The second inequality follows from the third condition of Lemma C.12.

Combining equations 118, 119 and further using $x_j = \mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j} - \lfloor\mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j}\rfloor$ for all $j \in [0,k]$ (Line 19) we have,

$$\prod_{i\in[\ell+(k+1)+1,\ell+2(k+1)]} {\mathbf{r}_i^{(2)}}^{(\mathbf{S}^{\text{ext}}\vec{\mathrm{m}})_i} \geq \prod_{j\in[0,k]} {\mathbf{r}_{\ell+(k+1)+1+j}^{(2)}}^{\mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j}\vec{\mathrm{m}}_j} = \prod_{i\in[\ell+(k+1)+1,\ell+2(k+1)]} {\mathbf{r}_i^{(2)}}^{(\mathbf{S}^{(2)}\vec{\mathrm{m}})_i}. \tag{120}$$

In the final inequality we used the second condition of Lemma C.15 (Diagonal Structure).

Combining equations 117 and 120 we have,

$$\prod_{i\in[1,\ell+2(k+1)]} {\mathbf{r}_i^{\text{ext}}}^{(\mathbf{S}^{\text{ext}}\vec{\mathrm{m}})_i} \geq \exp\left(-O(\gamma n)\right) \prod_{i\in[1,\ell+2(k+1)]} {\mathbf{r}_i^{(2)}}^{(\mathbf{S}^{(2)}\vec{\mathrm{m}})_i}$$

Using the above expression we have,

$$\begin{aligned}\frac{\mathbf{g}(\mathbf{S}^{\text{ext}})}{\mathbf{g}(\mathbf{S}^{(2)})} &\geq \exp\left(-O(\gamma n)\right) \prod_{i\in[1,\ell+2(k+1)]} \left(\frac{\exp\left((\mathbf{S}^{\text{ext}}\vec{1})_i\log(\mathbf{S}^{\text{ext}}\vec{1})_i - (\mathbf{S}^{(2)}\vec{1})_i\log(\mathbf{S}^{(2)}\vec{1})_i\right)}{\prod_{j'\in[0,k]}\exp\left(\mathbf{S}^{\text{ext}}_{i,j'}\log\mathbf{S}^{\text{ext}}_{i,j'} - \mathbf{S}^{(2)}_{i,j'}\log\mathbf{S}^{(2)}_{i,j'}\right)}\right)\\ &= \exp\left(-O(\gamma n)\right) \prod_{i\in[\ell+(k+1)+1,\ell+2(k+1)]} \left(\frac{\exp\left((\mathbf{S}^{\text{ext}}\vec{1})_i\log(\mathbf{S}^{\text{ext}}\vec{1})_i - (\mathbf{S}^{(2)}\vec{1})_i\log(\mathbf{S}^{(2)}\vec{1})_i\right)}{\prod_{j'\in[0,k]}\exp\left(\mathbf{S}^{\text{ext}}_{i,j'}\log\mathbf{S}^{\text{ext}}_{i,j'} - \mathbf{S}^{(2)}_{i,j'}\log\mathbf{S}^{(2)}_{i,j'}\right)}\right)\\ &= \exp\left(-O(\gamma n)\right) \prod_{i\in[\ell+(k+1)+1,\ell+2(k+1)]} \exp\left((\mathbf{S}^{\text{ext}}\vec{1})_i\log(\mathbf{S}^{\text{ext}}\vec{1})_i - \sum_{j'\in[0,k]}\mathbf{S}^{\text{ext}}_{i,j'}\log\mathbf{S}^{\text{ext}}_{i,j'}\right).\end{aligned} \tag{121}$$

In the second equality we used $\mathbf{S}^{\mathrm{ext}}_{i,j} = \mathbf{S}^{(2)}_{i,j}$ for all $i \in [1, \ell + (k+1)]$ and $j \in [0,k]$ (Line 22). The third inequality follows by the second condition of Lemma C.15 (Diagonal Structure). In the remainder of the proof we lower bound the term in the final expression.

For each $i \in [\ell + (k+1) + 1, \ell + 2(k+1)]$ let $j \in [0,k]$ be such that $i = \ell + (k+1) + 1 + j$, then $(\mathbf{S}^{\mathrm{ext}}\overrightarrow{1})_i = \sum_{j' \in [0,k]} (\lfloor \mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j'} \rfloor + z_{j,j'}) = \lfloor \mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j} \rfloor + \sum_{j' \in [0,k]} z_{j,j'}$. The first equality follows from line 23 of the algorithm. The second equality follows by the second condition of Lemma C.15 (Diagonal Structure). Using first condition of Lemma C.12, one of the following two cases hold,

1. If $\sum_{j' \in [0,k]} z_{j,j'} = 0$, then $z_{j,j'} = 0$ for all $j' \in [0,k]$. Using second condition of Lemma C.15 (Diagonal Structure), we have $\mathbf{S}^{\mathrm{ext}}_{\ell+(k+1)+1+j,j'} = \lfloor \mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j'} \rfloor + z_{j,j'} = 0$ for all $j' \in [0,k]$ and $j' \neq j$. Further note, $(\mathbf{S}^{\mathrm{ext}}\overrightarrow{1})_i = \lfloor \mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j} \rfloor + \sum_{j' \in [0,k]} z_{j,j'} = \mathbf{S}^{\mathrm{ext}}_{\ell+(k+1)+1+j,j}$. Combining previous two equalities we have, $(\mathbf{S}^{\mathrm{ext}}\overrightarrow{1})_i \log(\mathbf{S}^{\mathrm{ext}}\overrightarrow{1})_i - \sum_{j' \in [0,k]} \mathbf{S}^{\mathrm{ext}}_{i,j'} \log \mathbf{S}^{\mathrm{ext}}_{i,j'} = 0$. Therefore,
$$\exp\left((\mathbf{S}^{\mathrm{ext}}\overrightarrow{1})_i \log(\mathbf{S}^{\mathrm{ext}}\overrightarrow{1})_i - \sum_{j' \in [0,k]} \mathbf{S}^{\mathrm{ext}}_{i,j'} \log \mathbf{S}^{\mathrm{ext}}_{i,j'} \right) \geq 1 \, . \tag{122}$$

2. If $\sum_{j' \in [0,k]} z_{j,j'} = 1$, then $z_{j,j'} \in [0,1]_{\mathbb{R}}$ for all $j' \in [0,k]$. Using second condition of Lemma C.15 (Diagonal Structure), we have $\mathbf{S}^{\mathrm{ext}}_{i,j'} = \mathbf{S}^{\mathrm{ext}}_{\ell+(k+1)+1+j,j'} = \lfloor \mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j'} \rfloor + z_{j,j'} = z_{j,j'}$ for all $j' \in [0,k]$ and $j' \neq j$. Therefore, $\sum_{j' \in [0,k]} \mathbf{S}^{\mathrm{ext}}_{i,j'} \log \mathbf{S}^{\mathrm{ext}}_{i,j'} = (\lfloor \mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j} \rfloor + z_{j,j}) \log(\lfloor \mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j} \rfloor + z_{j,j}) + \sum_{j' \neq j} z_{j,j'} \log z_{j,j'} \leq (\lfloor \mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j} \rfloor + z_{j,j}) \log(\lfloor \mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j} \rfloor + z_{j,j})$. The final inequality follows because $z_{j,j'} \in [0,1]_{\mathbb{R}}$ and $z_{j,j'} \log z_{j,j'} \leq 0$ for all $j' \in [0,k]$.

   Further note, $(\mathbf{S}^{\mathrm{ext}}\overrightarrow{1})_i = \lfloor \mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j} \rfloor + \sum_{j' \in [0,k]} z_{j,j'} = \lfloor \mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j} \rfloor + 1$. Combining previous two inequalities we have, $(\mathbf{S}^{\mathrm{ext}}\overrightarrow{1})_i \log(\mathbf{S}^{\mathrm{ext}}\overrightarrow{1})_i - \sum_{j' \in [0,k]} \mathbf{S}^{\mathrm{ext}}_{i,j'} \log \mathbf{S}^{\mathrm{ext}}_{i,j'} \geq (\lfloor \mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j} \rfloor + 1) \log(\lfloor \mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j} \rfloor + 1) - (\lfloor \mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j} \rfloor + z_{j,j}) \log(\lfloor \mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j} \rfloor + z_{j,j}) \geq 0$. The last inequality follows because of the following: If $\lfloor \mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j} \rfloor = 0$, then the inequality follows because $z_{j,j} \in [0,1]_{\mathbb{R}}$ and $z_{j,j} \log z_{j,j} \leq 0$. Else $\lfloor \mathbf{S}^{(2)}_{\ell+(k+1)+1+j,j} \rfloor \geq 1$, in this case we use the fact that $x \log x$ is a monotonically increasing for $x \geq 1$.

   Therefore
$$\exp\left((\mathbf{S}^{\mathrm{ext}}\overrightarrow{1})_i \log(\mathbf{S}^{\mathrm{ext}}\overrightarrow{1})_i - \sum_{j' \in [0,k]} \mathbf{S}^{\mathrm{ext}}_{i,j'} \log \mathbf{S}^{\mathrm{ext}}_{i,j'} \right) \geq 1 \, . \tag{123}$$

Combining equations 122 and 123, for all $i \in [\ell + (k+1) + 1, \ell + 2(k+1)]$ we have,
$$\exp\left((\mathbf{S}^{\mathrm{ext}}\overrightarrow{1})_i \log(\mathbf{S}^{\mathrm{ext}}\overrightarrow{1})_i - \sum_{j' \in [0,k]} \mathbf{S}^{\mathrm{ext}}_{i,j'} \log \mathbf{S}^{\mathrm{ext}}_{i,j'} \right) \geq 1 \, .$$

Substituting previous inequality in Equation (121) we get,

$$\frac{\mathbf{g}(\mathbf{S}^{\text{ext}})}{\mathbf{g}(\mathbf{S}^{(2)})} \geq \exp\left(-O(\gamma n)\right) \ .$$

Further the condition 2 of the theorem follows by combining the above inequality with the sixth condition of Lemma C.15. ■

Appendix D. Alternative proof for the distinct column case.

Here we provide an alternative and simpler proof for Lemma A.1 which was pointed to us by an anonymous reviewer. This alternative proof is derived using Corollary 3.4.5 in Barvinok's book Barvinok (2017) (which is further derived using the Bregman-Minc inequality) and we formally state it below.

Lemma 1 (Corollary 3.4.5 from Barvinok (2017)) *Suppose that $\boldsymbol{Q}$ is a $N \times N$ doubly stochastic matrix that satisfies,*

$$\boldsymbol{Q}_{i,j} \leq \frac{1}{b_i} \textit{ for all } i \in [N], j \in [N]$$

for some positive integers $b_1, \dots b_N$. Then,

$$\text{perm}(\boldsymbol{Q}) \leq \prod_{i \in [N]} \frac{(b_i!)^{1/b_i}}{b_i} \ .$$

Using the above result, we now prove Lemma A.1 and we restate it for convenience.

Lemma A.1 (Scaled Sinkhorn permanent approximation) *For any matrix $\boldsymbol{A} \in \mathbb{R}_{\geq 0}^{\mathcal{D} \times \mathcal{D}}$ with at most k distinct columns, the following holds,*

$$\text{scaledsinkhorn}(\boldsymbol{A}) \leq \text{perm}(\boldsymbol{A}) \leq \exp\left(O\left(k \log \frac{N}{k}\right)\right) \text{scaledsinkhorn}(\boldsymbol{A}) \ . \tag{11}$$

Proof [Alternative proof for Lemma A.1] The lower bound follows from Corollary A.5 and in the remainder we prove the upper bound. Let $\mathbf{Q}$ be the maximizer of the scaled Sinkhorn objective, then it is a well know fact that $\mathbf{Q}$ satisfies,

$$\mathbf{Q} = \mathbf{LAR} \ ,$$

where matrices $\mathbf{L}$ and $\mathbf{R}$ are the left and right non-negative diagonal matrices. Further by the symmetry of the objective, there exists an optimum solution $\mathbf{Q}$ that has at most k distinct columns and we work with such an optimum solution. As $\mathbf{L}$ and $\mathbf{R}$ are diagonal matrices, the following two inequalities are trivial,

$$\text{perm}(\mathbf{Q}) = \text{perm}(\mathbf{L})\text{perm}(\mathbf{A})\text{perm}(\mathbf{R}) \ , \tag{124}$$

$$\text{scaledsinkhorn}(\mathbf{Q}) = \text{perm}(\mathbf{L}) \ \text{scaledsinkhorn}(\mathbf{A}) \ \text{perm}(\mathbf{R}), \tag{125}$$

Further note that for all doubly stochastic matrices $\mathbf{Q}$ we always have,

$$\exp(-N) \leq \text{scaledsinkhorn}(\mathbf{Q}) \ . \tag{126}$$

Therefore combining Equations (124) to (126), to prove the upper bound it is enough to show that,

$$\mathrm{perm}(\mathbf{Q}) \leq \exp\left(O\left(k \log \frac{N}{k}\right)\right) \cdot \exp(-N) \,.$$

As matrix $\mathbf{Q}$ has at most k distinct columns, let the multiplicities of these distinct columns be $\phi_1, \dots, \phi_k$. Note that if a column has multiplicity ϕ_i, the maximal element in this column is at most $1/\phi_i$. Now by Theorem 1 (Corollary 3.4.5. in Barvinok (2017)), we have

$$\mathrm{perm}(\mathbf{Q}) \leq \prod_{i=1}^{k} \frac{\phi_i!}{\phi_i^{\phi_i}} \leq \exp\left(O\left(k \log \frac{N}{k}\right)\right) \cdot \exp\left(-N\right) \,,$$

where the last inequality follows because the term $\prod_{i=1}^{k} \frac{\phi_i!}{\phi_i^{\phi_i}}$ is maximized when all ϕ_i's are equal and take value N/k. Therefore we conclude the proof. ■