LARGE RANDOM MATRICES WITH GIVEN MARGINS

HANBAEK LYU AND SUMIT MUKHERJEE

ABSTRACT. We study large random matrices with i.i.d. entries conditioned to have prescribed row and column sums (margins), a problem connected to relative entropy minimization, Schrödinger bridges, contingency tables, and random graphs with given degree sequences. Our central result is a *transference principle*: a complex margin-conditioned matrix can be closely approximated by a simpler matrix whose entries are independent and drawn from an exponential tilting of the original model. The tilt parameters are determined by the sum of two potentials. We establish phase diagrams for *tame margins*, where these potentials are uniformly bounded. This framework resolves a 2011 conjecture by Chatterjee, Diaconis, and Sly on δ-tame degree sequences and generalizes a sharp phase transition in contingency tables obtained by Dittmer, Lyu, and Pak in 2020. For tame margins, we show that a generalized Sinkhorn algorithm can compute the potentials at a dimension-free exponential rate. Our limit theory further establishes that for a convergent sequence of tame margins, the potentials converge as fast as the margins converge.

We apply this framework and obtain several key results for the margin-conditioned matrix: The marginal distribution of any single entry is asymptotically an exponential tilting of the base measure, resolving a 2010 conjecture by Barvinok on contingency tables. The conditioned matrix concentrates in cut norm around the *typical table* (the expectation of the tilted model), which acts as a static Schrödinger bridge between the margins. The empirical singular value distribution of the rescaled matrix converges to an explicit law determined by the variance profile of the tilted model. In particular, we confirm the universality of the Marchenko-Pastur law for constant linear margins.

CONTENTS

2010 *Mathematics Subject Classification.* 60B20, 60C05, 62H17, 49Q22.

Key words and phrases. Random matrices, margins, contingency tables, Schrödinger bridge, Sinkhorn algorithm, transference, concentration, empirical singular value distribution.

1. Introduction and main results

In this paper, we are interested in the structure of random matrices with i.i.d. entries conditioned to have prescribed row and column sums. Let μ be a σ-finite Borel measure on $\mathbb{R}$ and let

$$A := \inf\{\mathrm{supp}(\mu)\} \le \sup\{\mathrm{supp}(\mu)\} =: B. \tag{1.1}$$

We allow for the possibility that $A = -\infty$ or $B = \infty$, or both. The *base model* is the product measure $\mu^{\otimes(m\times n)}$ on the set of $m \times n$ random matrices. When μ is a probability measure, then the entries in the base model are independent and identically distributed as μ.

Let $\mathbf{x} = (x_{ij})$ be an $m \times n$ matrix of real entries. We define the *row margin* of $\mathbf{x}$ as the vector $r(\mathbf{x}) := (r_1(\mathbf{x}), \dots, r_m(\mathbf{x}))$ with $r_i(\mathbf{x}) := \sum_{j=1}^n x_{ij}$; the *column margin* of $\mathbf{x}$ is the vector $c(\mathbf{x}) := (c_1(\mathbf{x}), \dots, c_n(\mathbf{x}))$ with $c_j(\mathbf{x}) := \sum_{i=1}^m x_{ij}$. We call the pair $(r(\mathbf{x}), c(\mathbf{x}))$ the *margin* of $\mathbf{x}$. For each $\rho \ge 0$, we let

$$\mathcal{T}_\rho(\mathbf{r}, \mathbf{c}) := \left\{\mathbf{x} \in \mathbb{R}^{m\times n} : \|r(\mathbf{x}) - \mathbf{r}\|_1 \le \rho \text{ and } \|c(\mathbf{x}) - \mathbf{c}\|_1 \le \rho\right\}$$

denote the set of all $m \times n$ real matrices whose margin is within L^1 distance ρ from the prescribed margin $(\mathbf{r}, \mathbf{c})$. The *transportation polytope* with margin $(\mathbf{r}, \mathbf{c})$ is the set $\mathcal{T}(\mathbf{r}, \mathbf{c}) := \mathcal{T}_0(\mathbf{r}, \mathbf{c})$.

A fundamental question we investigate in this work is how a random matrix drawn from the base model behaves if we condition its margin to take prescribed values. Namely,

$$\textit{If we condition } X \sim \mu^{\otimes(m\times n)} \textit{ on being in } \mathcal{T}_\rho(\mathbf{r}, \mathbf{c}), \textit{ what does it look like?} \tag{1.2}$$

Since the margin is a fundamental observable for a matrix, it is natural to ask about the most likely structure of a random matrix after we observe its margin. This natural question connects to various important problems across diverse fields. When μ is the counting measure on the set $\mathbb{Z}_{\ge 0}$ of nonnegative integers, X is the uniformly random contingency table with given margin $(\mathbf{r}, \mathbf{c})$, which is a fundamental object in statistics and combinatorics [Goo63, DG95]. Counting the exact number of contingency tables is known to be #P-complete [DKM97] even for the $2 \times n$ case. There is extensive literature in combinatorics on approximately counting the number of contingency tables (see, e.g., [CM07, CM10, Bar09, BLSY10, Bar10a, Bar10b, BH10, LP22]). The number of contingency tables is also closely related to the Littlewood-Richardson coefficients [CEW22]. When μ is the counting measure on $\{0, 1\}$, X is a uniformly random bipartite graph

with given degree sequence [GM09, Bar10a, Wu20]. Further imposing symmetry and zero diagonal entries, it represents a uniformly random simple graphs with a given degree sequence [MW90a, MW90b, GIM21, CDS11, BH13]. When μ is the Lebesgue measure on $\mathbb{R}_{\geq 0}$, then X is a uniformly random nonnegative matrix from the transportation polytope $\mathscr{T}(\mathbf{r},\mathbf{c})$, which specializes to the uniformly random doubly stochastic matrix [CDS14, Ngu14]. When μ is the Poisson distribution with unit mean, X follows the multivariate hypergeometric distribution (or Fisher-Yates) with margin $(\mathbf{r},\mathbf{c})$ [DSC95]. We will see that in this case, the structure of X is closely related to the static Schrödinger bridge and entropic optimal transport [For40, PTT21].

A moment's thought reveals that it is not at all easy to sample such margin-conditioned random matrices. This task is often nontrivial, as margin conditioning can induce complex correlations between the entries. Specifically, the problem of sampling uniformly random contingency tables has been extensively studied (see, e.g., [DSC95, DKM97, KTV99, DG00, Mor02, Dye03, CDG$^+$06]). In this literature, the Diaconis-Gangolli Markov chain [DE85] is an important sampling algorithm, but obtaining the cutoff for it remains an open problem [NN20]. In 2020, Dittmer, Lyu, and Pak [DLP20] showed that the structure of uniformly random contingency tables exhibits a sharp phase transition as the margin varies continuously, a phenomenon conjectured by Barvinok [Bar10a] in 2010. This partially explains the hardness of the sampling problem.

We propose to approximate X by a random matrix Y with *independent entries* and establish the following *transference principle*:

(1.3) **Transference**: *Events that are sufficiently rare under Y are also rare under X.*

Hence, as far as such sufficiently rare events are concerned, one can avoid analyzing (and even sampling) X altogether and simply use the comparison model Y. The point is that the independence of the entries in Y makes it much easier to analyze than X. We propose two equivalent constructions of Y from dual perspectives of maximum likelihood (parametric) and minimum relative entropy (non-parametric):

1. (*Maximum Likelihood Perspective*): Y is the random matrix with independent entries obtained by a rank-one tilting of the base model $\mu^{\otimes(m\times n)}$, where the tilt parameters are obtained by a maximum-likelihood procedure based on the observable margin $(\mathbf{r},\mathbf{c})$.
2. (*Minimum Relative Entropy Perspective*): Among the class of random matrices with independent entries and expected margin $(\mathbf{r},\mathbf{c})$, Y has the minimum relative entropy from the base model $\mu^{\otimes(m\times n)}$.

The equivalent characterizations of Y above are in *Kantorovich duality* as in the static Schrödinger bridge/entropic optimal transport and their dual formulations. We give precise statements of our main results in the following subsections.

1.1. **Construction of the conditional probability measures.** We construct the law of the margin-conditioned random matrix X in (1.2), which we denote $\lambda_{\mathbf{r},\mathbf{c},\rho}$, under either of the following two (non-exclusive) assumptions. This covers all examples that we consider in this work and also all instances in the literature that we are aware of.

Assumption 1.1. $\mu^{\otimes(m\times n)}(\mathscr{T}_\rho(\mathbf{r},\mathbf{c}))\in(0,\infty)$.

Assumption 1.2. $\rho=0$ *or* $\mu^{\otimes(m\times n)}(\mathscr{T}_\rho(\mathbf{r},\mathbf{c}))=0$. *Furthermore, let* $\pi:\mathbb{R}^{m\times n}\to\mathbb{R}^{m+n}$ *denote the map that sends a matrix* $\mathbf{x}$ *to its margin* $(r(\mathbf{x}),c(\mathbf{x}))$. *Let* $\nu=\pi_\#(\mu^{\otimes(m\times n)})$ *denote the pushforward* $\mu^{\otimes(m\times n)}$ *under* π. *Then* ν *is* σ*-finite.*

Under Assumption 1.1, we simply define $\lambda_{\mathbf{r},\mathbf{c},\rho}$ by normalizing the product measure $\mu^{\otimes(m\times n)}$ restricted to $\mathscr{T}_\rho(\mathbf{r},\mathbf{c})$. Under Assumption 1.2, we need to condition on an event of measure zero

under $\mu^{\otimes(m\times n)}$. To handle this, we use the 'disintegration approach' for constructing regular conditional probabilities (e.g., see Chang and Pollard [CP97]). Namely, we set $\lambda_{\mathbf{r},\mathbf{c},\rho} := \lambda_{\mathbf{r},\mathbf{c},0}$, where the conditional probability measures $\lambda_{\mathbf{r},\mathbf{c}} := \lambda_{\mathbf{r},\mathbf{c},0}$ are constructed for ν-almost all margins $(\mathbf{r},\mathbf{c})$ via disintegrating the product measure $\mu^{\otimes(m\times n)}$ using the margin map $\mathbf{x} \mapsto (r(\mathbf{x}), c(\mathbf{x}))$. (See Sec. 5.2 for more details.) Frequently we will regard the pushforward measure ν on the margins as a measure on $\mathbb{R}^{m+n-1}$ by identifying a margin $(\mathbf{r},\mathbf{c})$ with the $m+n-1$-dimensional vector $(\mathbf{r}(1),\cdots,\mathbf{r}(m-1),\mathbf{c}(1),\cdots,\mathbf{c}(n))$ omitting $\mathbf{r}(m)$.

1.2. **Exponential tilting and the typical table.** Let us introduce some notation on exponential tilting and the parameterized comparison model. Throughout this paper, we use 'increasing' (resp., 'decreasing') and 'non-decreasing' (resp., 'non-increasing') interchangeably.

Definition 1.3 (Exponential tilting)**.** Define the set Θ of all allowed tilt parameters for μ:

$$\Theta := \Big\{\theta \in \mathbb{R} : \int e^{\theta x} d\mu(x) < \infty\Big\}. \tag{1.4}$$

Let Θ° be the interior of Θ. For any $\theta \in \Theta^\circ$, let μ_θ denote the tilted probability measure given by

$$\frac{d\mu_\theta}{d\mu}(x) = e^{\theta x - \psi(\theta)}, \quad \psi(\theta) := \log \int e^{\theta x} d\mu(x). \tag{1.5}$$

Then we have $\mathbb{E}_{\mu_\theta}[X] = \psi'(\theta)$ and $\mathrm{Var}_{\mu_\theta}(X) = \psi''(\theta) > 0$, so the function $\psi'(\cdot) : \Theta^\circ \mapsto (A,B)$ is strictly increasing, and has a strictly increasing inverse $\phi(\cdot) : (A,B) \mapsto \Theta^\circ$ satisfying $\phi(\psi'(\theta)) = \theta$ for all $\theta \in \Theta^\circ$. Throughout this paper, we assume the measure μ on $\mathbb{R}$ is such that the set Θ of tilt parameters in (1.4) is nonempty and not a singleton. Then it follows from Hölder's inequality that Θ is a nonempty interval and $\Theta^\circ = (\phi(A), \phi(B))$.

Now we introduce a random matrix model obtained by applying exponential tilting to the base model where the tilt parameters are parameterized by the direct sum of two vectors. These parameters will be tuned to achieve the prescribed margin as the expected margin.

Throughout this paper, we identify an n-dimensional vector $\mathbf{a}$ with the function $\mathbf{a} : \mathbb{R}^n \to \mathbb{R}$ so the ith coordinate of $\mathbf{a}$ is denoted by $\mathbf{a}(i)$. For two vectors $\mathbf{a},\mathbf{b}$, let $\mathbf{a}\oplus\mathbf{b}$ denote the matrix whose (i,j) coordinate is $\mathbf{a}(i)+\mathbf{b}(j)$. For a univariate function $\varphi : \mathbb{R} \to \mathbb{R}$, we denote by $\varphi(\mathbf{a}\oplus\mathbf{b})$ the matrix whose (i,j) entry is $\varphi(\mathbf{a}(i)+\mathbf{b}(j))$.

Definition 1.4 (The $(\boldsymbol{\alpha},\boldsymbol{\beta})$-model)**.** Let $\{\mu_\theta\}_{\theta\in\Theta^\circ}$ be probability measures on $\mathbb{R}$, as introduced in (1.5), and let vectors $\boldsymbol{\alpha} \in \mathbb{R}^m$ and $\boldsymbol{\beta} \in \mathbb{R}^m$ be such that $\boldsymbol{\alpha}(i) + \boldsymbol{\beta}(j) \in \Theta^\circ$ for all $i \in [m]$, $j \in [n]$. The $(\boldsymbol{\alpha},\boldsymbol{\beta})$-*model* is an $m \times n$ random matrix $Y = (Y_{ij})$ where the entries are independent and $Y_{ij} \sim \mu_{\boldsymbol{\alpha}(i)+\boldsymbol{\beta}(j)}$. In this case, we write $Y \sim \mu_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}$.

We begin by observing that the likelihood of observing a matrix $\mathbf{x} \in \mathbb{R}^{m\times n}$ under the $(\boldsymbol{\alpha},\boldsymbol{\beta})$-model depends only on the margin of $\mathbf{x}$. Indeed, the log-likelihood of observing $\mathbf{x}$ under $Y \sim \mu_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}$, with respect to the base measure $\mu^{\otimes(m\times n)}$ is

$$\sum_{i=1}^m \sum_{j=1}^n \Big[\mathbf{x}_{ij}(\boldsymbol{\alpha}(i)+\boldsymbol{\beta}(j)) - \psi(\boldsymbol{\alpha}(i)+\boldsymbol{\beta}(j))\Big] = \langle \mathbf{r},\boldsymbol{\alpha}\rangle + \langle \mathbf{c},\boldsymbol{\beta}\rangle - \sum_{i,j} \psi(\boldsymbol{\alpha}(i)+\boldsymbol{\beta}(j)),$$

where the right-hand side depends only on the margin of $\mathbf{x}$ but not on the specific entries of $\mathbf{x}$. Here, $\langle\cdot,\cdot\rangle$ denotes the standard dot product. Hence, given a realization $\mathbf{x}$ of the $(\boldsymbol{\alpha},\boldsymbol{\beta})$-model, the row and column sums of $\mathbf{x}$ form a sufficient statistic for the parameters $\boldsymbol{\alpha},\boldsymbol{\beta}$. This is analogous to the fact that the degree sequence of an observed graph under the $\boldsymbol{\beta}$-model is a sufficient statistic for $\boldsymbol{\beta}$, see [CDS11]. Given a single observation $\mathbf{x} \in \mathscr{T}(\mathbf{r},\mathbf{c})$, the maximum likelihood estimate (MLE) of $(\boldsymbol{\alpha},\boldsymbol{\beta})$ is obtained by maximizing the log-likelihood function above.

Definition 1.5 (MLE for margin $(\mathbf{r},\mathbf{c})$)**.** Fix an $m\times n$ margin $(\mathbf{r},\mathbf{c})$. The *MLE* of $(\boldsymbol{\alpha},\boldsymbol{\beta})$ for margin $(\mathbf{r},\mathbf{c})$ is a solution to the following concave maximization problem:

$$\sup_{\boldsymbol{\alpha},\boldsymbol{\beta}}\left(g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha},\boldsymbol{\beta}):=\langle\mathbf{r},\boldsymbol{\alpha}\rangle+\langle\mathbf{c},\boldsymbol{\beta}\rangle-\sum_{i,j}\psi(\boldsymbol{\alpha}(i)+\boldsymbol{\beta}(j))\right), \tag{1.6}$$

where we optimize over the open set where $\boldsymbol{\alpha}(i)+\boldsymbol{\beta}(j)\in\Theta^\circ$ for all $i\in[m]$, $j\in[n]$. An MLE $(\boldsymbol{\alpha},\boldsymbol{\beta})$ for $(\mathbf{r},\mathbf{c})$ is said to be a *standard MLE* if $\langle\boldsymbol{\alpha},\mathbf{1}\rangle=0$ and is denoted as $(\boldsymbol{\alpha}^{\mathbf{r},\mathbf{c}},\boldsymbol{\beta}^{\mathbf{r},\mathbf{c}})$.

A careful reader may wonder if the $(\boldsymbol{\alpha},\boldsymbol{\beta})$-model is a bit too restrictive for describing the structure of margin-conditioned random matrices. For instance, what if we use all mn independent tilt parameters θ_{ij} for each entry instead of the $m+n$ ones in the $(\boldsymbol{\alpha},\boldsymbol{\beta})$-model? Parameterizing θ_{ij} by the mean z_{ij} after tilting through the relation $\theta_{ij}=\phi(z_{ij})$, the optimal such mn tilt parameters are given by solving the following relative entropy minimization problem:

Definition 1.6 (Typical table)**.** Fix a $m\times n$ margin $(\mathbf{r},\mathbf{c})$. The *typical table* $Z^{\mathbf{r},\mathbf{c}}$ for margin $(\mathbf{r},\mathbf{c})$ with respect to the base measure μ is defined by

$$Z^{\mathbf{r},\mathbf{c}}:=\operatorname*{argmin}_{Z=(z_{ij})\in\mathcal{T}(\mathbf{r},\mathbf{c})}\left[H(Z):=\sum_{i,j}D(\mu_{\phi(z_{ij})}\|\mu)\right], \tag{1.7}$$

where $D(\mu_\theta\|\mu)$ is the *relative entropy* from μ to the tilted probability measure μ_θ defined as

$$D(\mu_\theta\|\mu):=\begin{cases}\int_{x\in\mathbb{R}}\log\left(\frac{d\mu_\theta}{d\mu}(x)\right)d\mu_\theta(x)=\theta\psi'(\theta)-\psi(\theta) & \text{if }\theta\in(\phi(A),\phi(B))\\ \infty & \text{otherwise.}\end{cases} \tag{1.8}$$

Note that when μ is a probability measure, the relative entropy above agrees with the Kullback-Leibler (KL) divergence from μ to μ_θ so $D(\mu_\theta\|\mu)\geq 0$. However, this quantity need not be nonnegative in general when μ is not a probability measure. For instance, if μ is the counting measure on nonnegative integers (see Ex. 3.4), then $D(\mu_\theta\|\mu)$ equals the negative entropy of the geometric distribution μ_θ, so it is nonpositive and is not bounded from below.

On the one hand, the MLE problem in (1.6) seeks to estimate the unknown parameters $(\boldsymbol{\alpha},\boldsymbol{\beta})$ that best describe the random matrix with a given margin through the $(\boldsymbol{\alpha},\boldsymbol{\beta})$-model. On the other hand, the typical table problem in (1.7) seeks to find the best $m\times n$ mean matrix in the transportation polytope that achieves the smallest possible relative entropy when the law of each entry is exponentially tilted. A key observation in this work is that these two problems are strongly dual to each other, analogously to the Kantorovich duality in the Schrödinger bridge theory (see Sec. 2.1).

Theorem 1.7 (Strong duality between typical table and MLE)**.** *Let $(\mathbf{r},\mathbf{c})$ be an $m\times n$ margin. Then an MLE $(\boldsymbol{\alpha},\boldsymbol{\beta})$ for $(\mathbf{r},\mathbf{c})$ exists if and only if the typical table $Z^{\mathbf{r},\mathbf{c}}$ exists if and only if $\mathcal{T}(\mathbf{r},\mathbf{c})\cap(A,B)^{m\times n}\neq\emptyset$. Assuming the last condition, the following implications hold:*

$$(\boldsymbol{\alpha},\boldsymbol{\beta})\text{ is an MLE}\iff\mathbb{E}_{\mu_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}}[Y]=\psi'(\boldsymbol{\alpha}\oplus\boldsymbol{\beta})\in\mathcal{T}(\mathbf{r},\mathbf{c})\iff Z^{\mathbf{r},\mathbf{c}}=\psi'(\boldsymbol{\alpha}\oplus\boldsymbol{\beta}). \tag{1.9}$$

Furthermore, the typical table and the MLE problems are in strong duality:

$$\inf_{Z\in\mathcal{T}(\mathbf{r},\mathbf{c})}H(Z)=\sup_{\boldsymbol{\alpha},\boldsymbol{\beta}}g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha},\boldsymbol{\beta}). \tag{1.10}$$

An implication of Theorem 1.7 above is that the $(\boldsymbol{\alpha},\boldsymbol{\beta})$-model is the best possible among all possible entry-wise exponential tiltings of the base model. In fact, when μ is a probability measure, it is the best possible among all random matrix ensembles, as it is the *information projection* of the base measure $\mu^{\otimes(m\times n)}$ onto the set of all probability measures on $m\times n$ real matrices constrained to have expected margin $(\mathbf{r},\mathbf{c})$. See Sec. 2.3 for more discussion.

The behavior of the $(\boldsymbol{\alpha},\boldsymbol{\beta})$-model depends crucially on how far the entries of $\boldsymbol{\alpha}\oplus\boldsymbol{\beta}$ are away from the boundary values $\phi(A)$ and $\phi(B)$. This leads to the following notion of 'tameness' of a margin. With a slight abuse of notation, we say $a \le M \le b$ for a matrix M and scalars a, b if every entry of M lies in $[a,b]$.

Definition 1.8 (Tame margins)**.** Fix $\delta > 0$ and let $\mathscr{M}^\delta = \mathscr{M}^\delta(\mu,m,n)$ denote the set of all $m \times n$ δ*-tame* margins $(\mathbf{r},\mathbf{c})$, that is, the MLE $(\boldsymbol{\alpha},\boldsymbol{\beta})$ exists and

$$A_\delta := \max\Big(A+\delta, -\frac{1}{\delta}\Big) \le \psi'(\boldsymbol{\alpha}\oplus\boldsymbol{\beta}) \le \min\Big(B-\delta, \frac{1}{\delta}\Big) =: B_\delta. \tag{1.11}$$

According to Theorem 1.7, δ-tameness of a margin $(\mathbf{r},\mathbf{c})$ can be equivalently defined as the typical table $Z^{\mathbf{r},\mathbf{c}}$ taking all entries from $[A_\delta, B_\delta]$. Also we remark that, since $\boldsymbol{\alpha}\oplus\boldsymbol{\beta}$ belongs to $(\phi(A),\phi(B))^{m\times n}$ by definition, any margin $(\mathbf{r},\mathbf{c})$ with an MLE (or typical table) is always δ-tame for some $\delta > 0$ that may depend on m and n. For asymptotic analysis, it is important to know whether a sequence of margins is uniformly δ-tame for a fixed $\delta > 0$. 'Cloning' a given margin provides a simple way to generate a family of δ-tame margins.

Example 1.9 (Cloned margins)**.** Let $(\mathbf{r}_0,\mathbf{c}_0)$ be an $a\times b$ margin with an MLE $(\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)$. The k*-cloning* of $(\mathbf{r}_0,\mathbf{c}_0)$ is the $ka\times kb$ margin $(\mathbf{r},\mathbf{c})$ with $\mathbf{r} = k\mathbf{r}_0\otimes\mathbf{1}_k$ and $\mathbf{c} = k\mathbf{c}_0\otimes\mathbf{1}_k$, where $\otimes$ denotes the Kronecker product, i.e., $\mathbf{r}$ and $\mathbf{c}$ repeat $k\mathbf{r}_0$ and $k\mathbf{c}_0$ k times, respectively. Then $(\mathbf{r},\mathbf{c})$ has MLE $(\boldsymbol{\alpha}_0\otimes\mathbf{1}_k,\boldsymbol{\beta}_0\otimes\mathbf{1}_k)$. Hence if $(\mathbf{r}_0,\mathbf{c}_0)$ is δ-tame for some $\delta = \delta(\mathbf{r}_0,\mathbf{c}_0) > 0$, then its k-clonings for all $k \ge 1$ are all δ-tame. For instance, the constant linear margin $(\mathbf{r},\mathbf{r})$ with $\mathbf{r} = na\mathbf{1}_n$ for any $a \in (A,B)$ is the n-cloning of the 1×1 margin (a,a) and is δ-tame for any $\delta > 0$ such that $A_\delta \le a \le B_\delta$.

Establishing δ-tameness for a general margin $(\mathbf{r},\mathbf{c})$ beyond the cloned ones turns out to be quite delicate. In Section 1.5, we establish phase diagrams and characterizations of the δ-tame margins.

1.3. **Transference principles.** In this section, we give precise statements of the transference principle informally stated in (1.3). The following 'weak transference principle' concerns the case when the measure of the conditioning set $\mathscr{T}_\rho(\mathbf{r},\mathbf{c})$ is nonzero as in Assumption 1.1.

Theorem 1.10 (Weak transference)**.** *Let $(\mathbf{r},\mathbf{c})$ be an $m\times n$ margin with an MLE $(\boldsymbol{\alpha},\boldsymbol{\beta})$ and let $X \sim \lambda_{\mathbf{r},\mathbf{c},\rho}$ under Assumption 1.1. Let $\delta > 0$ be small enough so that $(\mathbf{r},\mathbf{c})$ is δ-tame and let $Y \sim \mu_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}$. Then there exists a constant $C_1 = C_1(\mu,\delta) > 0$ such that for each Borel set $\mathscr{E} \subseteq \mathbb{R}^{m\times n}$,*

$$\mathbb{P}(X\in\mathscr{E}) \le \exp\big(C_1\rho\big)\,\mathbb{P}\big(Y\in\mathscr{T}_\rho(\mathbf{r},\mathbf{c})\big)^{-1}\,\mathbb{P}\left(Y\in\mathscr{E}\right). \tag{1.12}$$

Furthermore, if $\rho \ge C_2\sqrt{mn(m+n)}$ and $m,n \ge C_2$ for some constant $C_2 = C_2(\mu,\delta) > 0$, then

$$\mathbb{P}(X\in\mathscr{E}) \le 2\exp\big(C_1\rho\big)\,\mathbb{P}\left(Y\in\mathscr{E}\right). \tag{1.13}$$

For this to be useful, the probability $\mathbb{P}(Y\in\mathscr{T}_\rho(\mathbf{r},\mathbf{c}))$ that the maximum likelihood tilted model Y satisfies the margin constraint up to an L 1 error must be sufficiently large.

The upper bound in (1.12) is useful only if the product of the first two terms is not too large. For this, the probability $\mathbb{P}(Y\in\mathscr{T}_\rho(\mathbf{r},\mathbf{c}))$ that the maximum likelihood tilted model Y satisfying margin $(\mathbf{r},\mathbf{c})$ up to the L^1-error ρ must be sufficiently large. For the second part in (1.13), we show $\mathbb{P}(Y\in\mathscr{T}_\rho(\mathbf{r},\mathbf{c})) \ge 1/2$ if $\rho \approx \sqrt{mn(m+n)}$. The reason is that the expected margin of Y is $(\mathbf{r},\mathbf{c})$ and the entries of Y are independent with comparably tilted sub-exponential distributions $\mu_{\boldsymbol{\alpha}(i)+\boldsymbol{\beta}(j)}$. Hence by Bernstein's inequality, any fixed signed sum of its row/column sum fluctuates by $O(\sqrt{mn})$ with an exponential tail. Since there are at most 2^{m+n} such signed sums, the L^1 fluctuation of the margin of Y is of order $O(\sqrt{mn(m+n)})$. Hence $Y\in\mathscr{T}_\rho(\mathbf{r},\mathbf{c})$ with a large enough

probability. However, the overall transference cost $2\exp(C_1\rho)$ is still large, which comes from approximating the margin of Y by $(\mathbf{r},\mathbf{c})$ when we only know it is within L^1-distance ρ away from $(\mathbf{r},\mathbf{c})$.

Next, we establish several 'strong transference principles' for exact margin conditioning $\rho = 0$ or $\mathscr{T}_\rho(\mathbf{r},\mathbf{c})$ having measure zero (Assumption 1.2). In order to state our results, we need to introduce some notations. For each matrix $\mathbf{z} \in \mathbb{R}^{m\times n}$, let $\bar{\mathbf{z}}$ denote its $(m-1)\times(n-1)$ submatrix obtained by deleting the last row and column from $\mathbf{z}$, and let $\check{\mathbf{z}}$ denote the $m+n-1$ entries of $\mathbf{z}$ in its last row and column. Note that X is completely determined by $\overline{X}$ due to the exact margin constraint. More precisely, define the 'completion map' $\Gamma_{\mathbf{r},\mathbf{c}} : \mathbb{R}^{(m-1)\times(n-1)} \to \mathbb{R}^{m\times n}$ by

$$\Gamma_{\mathbf{r},\mathbf{c}}(\bar{\mathbf{x}})_{ij} := \begin{cases} \bar{\mathbf{x}}_{ij} & \text{if } 1 \le i < m \text{ and } 1 \le j < n \\ \mathbf{r}(i) - \bar{\mathbf{x}}_{i\bullet} & \text{for } 1 \le i < m \text{ and } j = n \\ \mathbf{c}(j) - \bar{\mathbf{x}}_{\bullet j} & \text{for } i = m \text{ and } 1 \le j < n \\ \bar{\mathbf{x}}_{\bullet\bullet} - \left(\sum_{i=1}^{m-1} \mathbf{r}(i)\right) + \mathbf{c}(n) & \text{if } (i,j) = (m,n). \end{cases} \tag{1.14}$$

Here, $\bar{\mathbf{x}}_{i\bullet}$, $\bar{\mathbf{x}}_{\bullet j}$, and $\bar{\mathbf{x}}_{\bullet\bullet}$ denote the ith row sum, the jth column sum, and the total sum of $\bar{\mathbf{x}}$, respectively. Then given an $(m-1)\times(n-1)$ matrix $\bar{\mathbf{x}}$, $\Gamma_{\mathbf{r},\mathbf{c}}(\bar{\mathbf{x}})$ is the unique $m\times n$ matrix with margin $(\mathbf{r},\mathbf{c})$ such that $\overline{\Gamma_{\mathbf{r},\mathbf{c}}(\bar{\mathbf{x}})} = \bar{\mathbf{x}}$. In particular, $X = \Gamma_{\mathbf{r},\mathbf{c}}(\overline{X})$. We let $\check{\Gamma}_{\mathbf{r},\mathbf{c}}(\cdot)$ denote the entries in the last row and column of $\Gamma_{\mathbf{r},\mathbf{c}}(\cdot)$.

The result below is our general transference principle for exactly margin-conditioned matrices.

Theorem 1.11 (Strong transference)**.** *Suppose Assumption 1.2 holds. For ν-almost all margins $(\mathbf{r},\mathbf{c})$ with an MLE $(\boldsymbol{\alpha},\boldsymbol{\beta})$, the following hold:*

(i) *Let $X \sim \lambda_{\mathbf{r},\mathbf{c}}$ and $Y \sim \mu_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}$. Then X has the same distribution as Y conditional on $Y \in \mathscr{T}(\mathbf{r},\mathbf{c})$.*

(ii) *Let $\nu_{\bar{\mathbf{y}}}(\cdot)$ denote the law of $(r(Y),c(Y))$ given $\overline{Y} = \bar{\mathbf{y}}$ and let $p_{\bar{\mathbf{y}}}(\cdot)$ denote the Radon-Nikodym derivative of $\nu_{\bar{\mathbf{y}}}$ w.r.t. the unconditional law ν_Y of $(r(Y),c(Y))$. Then for each bounded measurable function $h : \mathbb{R}^{(m-1)\times(n-1)} \to \mathbb{R}$,*

$$\mathbb{E}[h(\overline{X})] = \mathbb{E}\left[p_{\overline{Y}}(\mathbf{r},\mathbf{c})\,h(\overline{Y})\right] \le \left(\sup_{\bar{\mathbf{y}}} p_{\bar{\mathbf{y}}}(\mathbf{r},\mathbf{c})\right)\mathbb{E}[h(\overline{Y})]. \tag{1.15}$$

Furthermore, if $\bar{\mathbf{y}} \mapsto p_{\bar{\mathbf{y}}}(\mathbf{r},\mathbf{c})$ is proportional to some function $q_{\bar{\mathbf{y}}}(\mathbf{r},\mathbf{c})$, then

$$\sup_{\bar{\mathbf{y}}} p_{\bar{\mathbf{y}}}(\mathbf{r},\mathbf{c}) = \left(\sup_{\bar{\mathbf{y}}} q_{\bar{\mathbf{y}}}(\mathbf{r},\mathbf{c})\right)\mathbb{E}[q_{\overline{Y}}(\mathbf{r},\mathbf{c})]^{-1}. \tag{1.16}$$

Following our high-level transference principle in (1.3), we expect to establish results describing the structure of X based on the approximation scheme $X \approx Y$, where Y is the (unconditional) maximum-likelihood $(\boldsymbol{\alpha},\boldsymbol{\beta})$-model for the margin $(\mathbf{r},\mathbf{c})$. However, one should not expect this to give a useful approximation for arbitrary base measure and margin for the following counterexample. Consider $\mu = \mathrm{Uniform}(\{0,1,\sqrt{2}\})$ and symmetric constant margin $(\mathbf{r},\mathbf{r})$ with $\mathbf{r} = n\mathbf{1}_n$. Then $(\mathbf{r},\mathbf{r})$ has a positive mass under ν, $\overline{X} = \mathbf{1}\mathbf{1}^\top$ almost surely, but $\overline{Y} = \mathbf{1}\mathbf{1}^\top$ with probability exponentially small in n^2. For this example, the supremum of $p_{\bar{\mathbf{y}}}(\mathbf{r},\mathbf{c})$ over $\bar{\mathbf{y}}$ is $\mathbb{P}(Y \in \mathscr{T}(\mathbf{r},\mathbf{c}))^{-1}$, which is exponentially large in n^2. Therefore, in this case, even the large-deviations events under Y cannot be shown to be rare under X using the transference approach. (See Ex. 3.9 for details.)

Next, we deduce a useful corollary of Theorem 1.11 by obtaining a computable form of the proportionality $q_{\bar{\mathbf{y}}}(\mathbf{r},\mathbf{c})$. Note that the law $\nu_{\bar{\mathbf{y}}}$ of the margin $(r(Y),c(Y))$ given $\overline{Y} = \bar{\mathbf{y}}$ is the pullback of the law of $\check{Y}$ under the one-to-one map $(\mathbf{r},\mathbf{c}) \mapsto \check{\Gamma}_{\mathbf{r},\mathbf{c}}(\bar{\mathbf{y}})$:

$$\nu_{\bar{\mathbf{y}}}(d(\mathbf{r},\mathbf{c})) = \bigotimes_{i=m \text{ or } j=n} \mu_{\boldsymbol{\alpha}(i)+\boldsymbol{\beta}(j)}\left(d\check{\Gamma}_{\mathbf{r},\mathbf{c}}(\bar{\mathbf{y}})_{ij}\right). \tag{1.17}$$

Hence locally $\nu_{\overline{\mathbf{y}}}$ is the product of shifts of the tilted measures $\mu_{\boldsymbol{\alpha}(i)+\boldsymbol{\beta}(j)}$ for $i=m$ or $j=n$. Thus, if these shifted and tilted measures have density w.r.t. a common measure, say ζ, on $\mathbb{R}$, then we can easily compute the Radon-Nikodym derivative $d\nu_{\overline{\mathbf{y}}}/d\zeta^{\otimes(m+n-1)}$. In this case, it would be natural to suspect that $\overline{\mathbf{y}}\mapsto p_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c})$ is proportional to this Radon-Nikodym derivative. This conclusion indeed holds if $d\zeta^{\otimes(m+n-1)}/d\nu$ makes sense at least locally near each margin $(\mathbf{r},\mathbf{c})$ in the support of ν. This latter condition holds for a wide range of discrete and continuous measures as stated in Corollary 1.12 below.

Corollary 1.12. *Keep the same setting in Theorem 1.11. Suppose μ has density $p\geq 0$ w.r.t. ζ, which is either the counting measure on $\mathbb{Z}$ or the Lebesgue measure on $\mathbb{R}$. In the continuous case, suppose $\{p>0\}$ is open. Then (1.16) holds with $q_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c})=d\nu_{\overline{\mathbf{y}}}/d\zeta^{\otimes(m+n-1)}$ and*

$$\mathbb{E}[q_{\overline{Y}}(\mathbf{r},\mathbf{c})]=\exp(-g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha},\boldsymbol{\beta}))\int\prod_{i,j}p(\Gamma_{\mathbf{r},\mathbf{c}}(\overline{\mathbf{x}})_{ij})\,\zeta^{\otimes(m-1)\times(n-1)}(d\overline{\mathbf{x}}). \tag{1.18}$$

When μ is the Lebesgue measure on $\mathbb{R}_{\geq 0}$, notice that the integral in (1.18) is the volume of the transportation polytope $\mathcal{T}(\mathbf{r},\mathbf{c})\cap\mathbb{R}_{\geq 0}^{m\times n}$ of nonnegative matrices with margin $(\mathbf{r},\mathbf{c})$. Also when μ is the counting measure on $\mathbb{Z}_{\geq 0}$, then the same integral is the number of contingency tables with margin $(\mathbf{r},\mathbf{c})$. Various upper and lower bounds on these quantities have been extensively studied in the combinatorics literature in the last three decades. For our purpose, we can use the lower bounds on the number of contingency tables [Bar09, BLP23] and of the volume of the transportation polytope [CM07, BH12, BR24]. This gives the following strong transference result that does not require tameness:

Theorem 1.13. *Suppose Assumption 1.2 holds. Assume μ is the counting measure on $[0,b)\cap\mathbb{Z}$ for some $b\in[0,\infty]$ or the Lebesgue measure on $\mathbb{R}_{\geq 0}$. Then there exists an absolute constant $\gamma>0$ such that for ν-almost all margins $(\mathbf{r},\mathbf{c})$ with an MLE $(\boldsymbol{\alpha},\boldsymbol{\beta})$ and $\mathbf{r}(i)\geq 1,\mathbf{c}(j)\geq 1$ for all i,j,*

$$\mathbb{E}[h(\overline{X})]\leq N^{\gamma(m+n)}\,\mathbb{E}[h(\overline{Y})]$$

for each bounded measurable function $h:\mathbb{R}^{(m-1)\times(n-1)}\to\mathbb{R}$, denoting $N=\sum_i\mathbf{r}(i)=\sum_j\mathbf{c}(j)$.

For possibly non-uniform density function $p\geq 0$, the integral in (1.18) can be viewed as a weighted volume of the transportation polytope $\mathcal{T}(\mathbf{r},\mathbf{c})\cap\mathrm{supp}(\mu)^{m\times n}$ where the weight is proportional to the product of $p(\Gamma_{\mathbf{r},\mathbf{c}}(\overline{\mathbf{x}})_{ij})$. It seems that there is no known upper bound on such 'weighted volume' of the transportation polytope for general density p. We circumvent this limitation by lower-bounding the expectation $\mathbb{E}[d\nu_{\overline{Y}}/d\zeta^{\otimes(m+n-1)}]$ through a probabilistic argument. This gives the following strong transference principle that holds for a wide range of discrete and continuous base measures with a looser bound on the transference cost.

Theorem 1.14. *Suppose Assumption 1.2 holds. Assume that μ has a density $p\geq 0$ w.r.t. ζ, which is either the counting measure on $\mathbb{Z}$ or the Lebesgue measure on $\mathbb{R}$. Further, assume:*

1. *There exists a constant $a>0$ such that $(m\wedge n)\geq(m\vee n)^a$ and m,n are sufficiently large so that $m\wedge n\geq 12^a$.*
2. *$\mathrm{supp}(\mu)=[A,B]\cap\mathrm{supp}(\zeta)$ in both cases and the Radon-Nikodym derivatives of μ_θ w.r.t. ζ for $\theta\in[\phi(A_\delta),\phi(B_\delta)]$ for any $\delta>0$ are uniformly upper bounded by some constant.*

Then there exists a constant $C=C(\mu,\delta,a)>0$ such that for ν-almost all δ-tame margins $(\mathbf{r},\mathbf{c})$ and for each bounded measurable function $h:\mathbb{R}^{(m-1)\times(n-1)}\to\mathbb{R}$,

$$\mathbb{E}[h(\overline{X})]\leq\exp(C(m\sqrt{n}\log n+n\sqrt{m}\log m)\log mn)\,\mathbb{E}[h(\overline{Y})].$$

Remark 1.15. For a weaker condition $(m\wedge n)\gg\log(m\vee n)$ on the aspect ratio, a minor modification of our argument for Theorem 1.14 shows that the similar strong transference holds with a

larger transference cost bound $\exp(o(mn))$. Since Theorem 1.13 based on combinatorial estimates holds for arbitrary m, n, we suspect the requirement on m, n in Theorem 1.14 is only an artifact of our proof technique.

Remark 1.16. Our result in Theorem 1.14 obtained by probabilistic arguments imply lower bounds on the volume of the polytope with general densities p other than $p(x) = \mathbf{1}(x \geq 0)$. See Remark 5.8 for details.

1.4. **Limit theory for the comparison model.** Our next goal is to establish a limit theory for the margin-conditioned random matrices as the sequence of $m \times n$ margins $(\mathbf{r}_m, \mathbf{c}_n)$ converges to a limiting 'continuum margin' $(\mathbf{r}, \mathbf{c})$ in a suitable sense, as m, n both tend to infinity. By transference, it should essentially be enough to establish such a result for the corresponding maximum-likelihood tilted models. This is precisely what we establish in this section with an explicit bound on the rate of convergence of the corresponding typical tables and the rescaled MLEs.

First, let us make the convergence of margins precise.

Definition 1.17 (Continuum margin)**.** A *continuum margin* is a pair $(\mathbf{r}, \mathbf{c})$ of integrable functions $\mathbf{r}, \mathbf{c} : (0,1] \to \mathbb{R}$ such that $\int_0^1 \mathbf{r}(x)\,dx = \int_0^1 \mathbf{c}(y)\,dy$. For a $m \times n$ discrete margin $(\mathbf{r}_m, \mathbf{c}_n)$, define the corresponding continuum step margin $(\bar{\mathbf{r}}_m, \bar{\mathbf{c}}_n)$ as

$$\bar{\mathbf{r}}_m(t) := n^{-1}\mathbf{r}_m(\lceil mt \rceil), \qquad \bar{\mathbf{c}}_n(t) := m^{-1}\mathbf{c}_n(\lceil nt \rceil). \tag{1.19}$$

We say a sequence of $m \times n$ margins $(\mathbf{r}_m, \mathbf{c}_n)$ *converges in* L^1 to a continuum margin $(\mathbf{r}, \mathbf{c})$ if

$$\lim_{m,n\to\infty} \|\mathbf{r} - \bar{\mathbf{r}}_m\|_1 + \|\mathbf{c} - \bar{\mathbf{c}}_n\|_1 = 0.$$

It will be convenient to compare matrices of different dimensions in the space of kernels.

Definition 1.18 (Kernels)**.** A *kernel* is an integrable function $W : [0,1]^2 \to \mathbb{R}$. Given an $m \times n$ matrix A, define a function W_A on the unit square as follows: Partition the unit square $(0,1]^2$ into $m \times n$ rectangles of the form $R_{ij} := \left(\frac{i-1}{m}, \frac{i}{m}\right] \times \left(\frac{j-1}{n}, \frac{j}{n}\right]$, for $i \in [m]$ and $j \in [n]$. Set

$$W_A(x, y) := A_{ij} \text{ if } (x, y) \in R_{ij}.$$

The *p-norm* of a kernel W is defined as

$$\|W\|_p := \left(\int_{S\times T} |W(x,y)|^p\,dx\,dy\right)^{1/p}.$$

In Theorem 1.19 below, we show that the typical tables and the rescaled MLEs converge in L^2 to a limiting kernel characterized by a continuum dual variable, provided the discrete margins are uniformly δ-tame for a fixed $\delta > 0$. Furthermore, the rate of convergence (measured in the squared L^2-distance) is at least the rate of convergence of the margins (measured in L^1-distance).

Theorem 1.19 (Scaling limit of typical tables and MLEs)**.** *Fix $\delta > 0$ and let $(\mathbf{r}_m, \mathbf{c}_n)$ be a sequence of $m \times n$ δ-tame margins converging to a continuum margin $(\mathbf{r}, \mathbf{c})$ in L^1 as $m, n \to \infty$.*

(i) *There exists bounded measurable functions $\boldsymbol{\alpha}, \boldsymbol{\beta} : [0,1] \to \mathbb{R}$ s.t. $\int \boldsymbol{\alpha}(x)dx = 0$, $\phi(A_\delta) \leq \boldsymbol{\alpha} \oplus \boldsymbol{\beta} \leq \phi(B_\delta)$, and the kernel $W^{\mathbf{r},\mathbf{c}} := \psi'(\boldsymbol{\alpha} \oplus \boldsymbol{\beta})$ has continuum margin $(\mathbf{r}, \mathbf{c})$.*

(ii) *Let $C_\delta := 4\max\{|\phi(A_\delta)|, |\phi(B_\delta)|\}$, where A_δ, B_δ are as in (1.11). Then*

$$\|W^{\mathbf{r},\mathbf{c}} - W_{Z^{\mathbf{r}_m,\mathbf{c}_n}}\|_2^2 \leq C_\delta \left(\sup_{\phi(A_\delta)\leq w\leq \phi(B_\delta)} \psi''(w)\right) \|(\mathbf{r}, \mathbf{c}) - (\bar{\mathbf{r}}_m, \bar{\mathbf{c}}_n)\|_1.$$

Furthermore, if $(\boldsymbol{\alpha}_m, \boldsymbol{\beta}_n)$ *is the standard MLE for margin* $(\mathbf{r}_m, \mathbf{c}_n)$ *and if we define functions* $\bar{\boldsymbol{\alpha}}_m, \bar{\boldsymbol{\beta}}_n : [0,1] \to \mathbb{R}$ *as* $\bar{\boldsymbol{\alpha}}_m(x) := \boldsymbol{\alpha}_m(\lceil mx \rceil)$ *and* $\bar{\boldsymbol{\beta}}_n(y) := \boldsymbol{\beta}_n(\lceil ny \rceil)$*, then*

$$\|\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}_m\|_2^2 + \|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}_n\|_2^2 \le C_\delta \left(\frac{\sup_{\phi(A_\delta) \le w \le \phi(B_\delta)} \psi''(w)}{\inf_{\phi(A_\delta) \le w \le \phi(B_\delta)} \psi''(w)} \right) \|(\mathbf{r}, \mathbf{c}) - (\bar{\mathbf{r}}_m, \bar{\mathbf{c}}_n)\|_1.$$

In particular, $\|W^{\mathbf{r},\mathbf{c}} - W_{Z^{\mathbf{r}_m,\mathbf{c}_n}}\|_2 \to 0$ *and* $\|(\boldsymbol{\alpha}, \boldsymbol{\beta}) - (\bar{\boldsymbol{\alpha}}_m, \bar{\boldsymbol{\beta}}_n)\|_2 \to 0$ *as* $m, n \to \infty$.

1.5. **Phase diagram for tame margins.** The assumption of δ-tameness of the margin was crucial in the development we described in earlier sections. But since this assumption is implicit, it will only be useful if we can provide a more explicit characterization of the set $\mathscr{M}^\delta$ of all δ-tame margins for each $\delta > 0$. We will seek such conditions depending only on the extreme values of the margin. Namely, for each point $(s,t) \in (A,B)^2$, $s \le t$ in a two-dimensional phase diagram, we ask if an arbitrary $m \times n$ margin $(\mathbf{r}, \mathbf{c})$ satisfying

$$\mathscr{T}(\mathbf{r}, \mathbf{c}) \cap (A,B)^{m \times n} \ne \emptyset \quad \text{and} \quad s \le \mathbf{r}(i)/n, \mathbf{c}(j)/m \le t \quad \text{for all } (i,j) \in [m] \times [n] \tag{1.20}$$

is δ-tame for some $\delta = \delta(\mu, s, t) > 0$. Let $\Omega(\mu) \subseteq (A,B)^2$ denote the set of all such points (s,t) that guarantee such uniform tameness for the base measure μ, which we call the *phase diagram* for tame margins for μ. The goal of this section is to identify the phase diagrams for various measures μ. Our results are summarized in Figure 1.

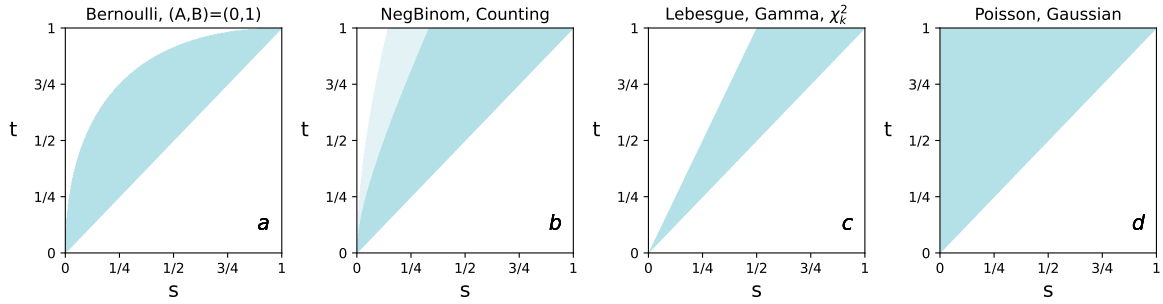

FIGURE 1. Phase diagrams for tame margins for various base measures μ. The upper contours are given by (**a**) $(s+t)^2 < 2s$, (**b**) $t \le 1 + \sqrt{1 + rs^{-1}}$ with $r = 5$ for Negative Binomial (NegBinom) and $r = 1$ for Counting (so $\Omega(\mu)$ for the former measure is the union of the two shaded regions), (**c**) $t \le s/2$, and (**d**) $t = \infty$.

First, when μ has bounded support (i.e., $|A|, |B| < \infty$), we show that $\Omega(\mu)$ is determined by a simple quadratic inequality involving only A and B. Furthermore, the shape of $\Omega(\mu)$ is universal among the class of all measures of bounded support (see Fig. 1**a**) up to translation and scaling. This result is stated below.

Theorem 1.20. *Suppose* $-\infty < A \le B < \infty$*. Then each* $(s,t) \in (A,B)^2$ *with* $s \le t$ *belongs to* $\Omega(\mu)$ *if*

$$(s + t - 2A)^2 < 4(B - A)(s - A). \tag{1.21}$$

Furthermore, (s,t) *does not belong to* $\Omega(\mu)$ *if the reverse inequality in* (1.21) *holds.*

When μ has unbounded support, the situation changes drastically and the dependence of $\Omega(\mu)$ on μ is more sophisticated. We first investigate the phase diagram restricted to the particular class of 2×2 block margins that we call the 'Barvinok margins'. These are symmetric linear margins with two distinct values where a vanishing fraction of the row sums are assigned the larger value (see [Bar10b, Sec. 1.6] and [DLP20, LP22]). In this case, the phase diagram is determined by the inequality $2\phi(t) - \phi(s) < \phi(B)$ as stated below.

Corollary 1.21 (of Proposition 7.3)**.** *Suppose μ is such that $A = 0$ and $B = \infty$. Consider a sequence of symmetric $n \times n$ margins $(\mathbf{r}_n, \mathbf{r}_n)$ such that*

$$\mathbf{r}_n := \big(\underbrace{s_n n, \dots, s_n n}_{n-\lfloor n^a \rfloor}, \underbrace{t_n n, \dots, t_n n}_{\lfloor n^a \rfloor}\big) \in \mathbb{R}^n_{\geq 0} \tag{1.22}$$

where $s_n \to s$ and $t_n \to t$ as $n \to \infty$ for $0 \leq s \leq t$ and $a \in [0,1)$ is fixed. If $2\phi(t) - \phi(s) < \phi(B)$, then for all $n \geq 1$, $(\mathbf{r}_n, \mathbf{r}_n)$ is δ-tame for some $\delta = \delta(\mu, s, t) > 0$. If $2\phi(t) - \phi(s) > \phi(B)$, then the (n,n) entry of the typical table for $(\mathbf{r}_n, \mathbf{r}_n)$ diverges to infinity as $n \to \infty$.

The above result implies that for base measures μ with $(A, B) = (0, \infty)$, the interior of $\Omega(\mu)$ is contained in the region defined by $2\phi(t) - \phi(s) < \phi(B)$ and $s \leq t$. While there is a possibility that $\Omega(\mu)$ can be much smaller than what Corollary 1.21 suggests, below in Theorem 1.22, we establish that these two regions in fact coincide whenever ψ'' is increasing and log-convex. Recall that $\psi''(\theta)$ is the variance of the tilted measure μ_θ. Hence μ must have right-infinite support ($B = \infty$) if ψ'' is increasing. Since the more exponential tilting the more mass on the larger values, it is natural to expect that ψ'' would be increasing. Log-convexity is an additional condition we crucially use to reduce the analysis for general symmetric margins to Barvinok margins (see Lem. 7.2). These conditions hold for a wide range of measures with unbounded support such as Gaussian, Poisson, counting on $\mathbb{Z}_{\geq 0}$, Lebesgue on $\mathbb{R}_{\geq 0}$, and Gamma (see Sec. 3).

Theorem 1.22. *Suppose ψ'' is increasing and log-convex on Θ° (necessarily $B = \infty$). Then each $(s,t) \in (A,B)^2$ with $s \leq t$ belongs to $\Omega(\mu)$ if*

$$\phi(A) < 3\phi(s) - 2\phi(t) \quad \text{and} \quad 2\phi(t) - \phi(s) < \phi(B). \tag{1.23}$$

Furthermore, for measures μ s.t. $\phi(A) = -\infty$, (s,t) does not belong to $\Omega(\mu)$ if $2\phi(t) - \phi(s) > \phi(B)$.

The first inequality in (1.23) is often vacuous (e.g., when $\phi(A) = -\infty$), and only the second one matters. For a wide range of measures, we can solve the second inequality explicitly and the phase diagram $\Omega(\mu)$ is determined by the ratio t/s being strictly less than some critical threshold λ_c, as stated in Corollary 1.23 below.

Corollary 1.23. *An arbitrary $m \times n$ margin $(\mathbf{r}, \mathbf{c})$ satisfying (1.20) with $0 < s < t$ is δ-tame for some $\delta = \delta(\mu, s, t) > 0$ provided $t/s < \lambda_c$, where*

$$\lambda_c := \begin{cases} 1 + \sqrt{1 + r s^{-1}} & \text{if } \mu = r\text{-fold convolution of the counting measure on } \mathbb{Z}_{\geq 0} \text{ for } r \geq 1, \\ 2 & \text{if } \mu \text{ has density } e^{-ax} x^{\gamma - 1} \text{ on } \mathbb{R}_{\geq 0} \text{ w.r.t. Lebesgue measure for } a \geq 0,\ \gamma \geq 1 \\ \infty & \text{if } \mu = \text{Poisson or Gaussian.} \end{cases} \tag{1.24}$$

Furthermore, for μ as above, if $t/s > \lambda_c$, then there exists a sequence of $n \times n$ margins satisfying the hypothesis but for which some entry in the typical table diverging to infinity as $n \to \infty$.

The critical threshold λ_c has not been identified before for any base measure, but there are some previously known bounds on the critical ratio λ_c for the counting and the Lebesgue cases.

For the counting base measure, Barvinok, Luria, Samorodnitsky, and Yong [BLSY10, Thm. 3.5] showed that $\lambda_c \leq \frac{1+\sqrt{5}}{2} \approx 1.618$, the golden ratio. We learned from Barvinok that this bound was later improved to 2 by Luria [Lur08]. Our Corollary 1.23 identifies the critical threshold as $\lambda_c = 1 + \sqrt{1 + s^{-1}}$, which agrees with the sharp phase transition point for Barvinok margins obtained by Dittmer, Lyu, and Pak [DLP20]. Barvinok and Hartigan [BH12] obtained a Gaussian formula for the number of contingency tables that is asymptotically correct for large δ-tame margins. Corollary 1.23 implies this asymptotic formula holds for margins satisfying (1.20) with $t/s < 1 + \sqrt{1 + s^{-1}}$ and this result cannot be improved.

For the Lebesgue bese measure on $\mathbb{R}_{\geq 0}$, Barvinok and Rudelson [BR24] recently observed that $\lambda_c \geq 2$. Namely, they noted that the Barvinok margin $\mathbf{r} = \mathbf{c} = (n, \dots, n, \lambda n)$, now with the Lebesgue base measure (see Ex. 3.6 for the corresponding typical table), is asymptotically δ-tame for $\lambda < 2$ and it is not for $\lambda > 2$ (see Remark 7.4). There was no previously known upper bound on λ_c, and our result above shows that $\lambda_c = 2$.

As we have an almost complete understanding of the phase diagram of tame margins, one may ask a finer question about characterizing a necessary and sufficient condition for a given margin to be δ-tame for some δ independent of the margin. We provide such a result for symmetric margins and general base measures with compact support. The equivalent condition for δ-tameness turns out to be the celebrated Erdős-Gallai condition [EG60] with a quadratic gap.

Theorem 1.24 (δ-tameness and Erdős-Gallai condition for symmetric margins)**.** *Suppose μ has compact support with $A = 0$ and $B < \infty$. If an $n \times n$ symmetric margin $(\mathbf{r}, \mathbf{r})$ with an MLE $(\boldsymbol{\alpha}, \boldsymbol{\alpha})$ is δ-tame for some $\delta > 0$, then there exists constants $c_1, c_2 \in (0, B)$ and $c_3 > 0$ depending only on μ and δ such that $c_1 \leq \mathbf{r}(i)/n \leq c_2$ for all i and*

$$\min_{I \subseteq [n], |I| \geq c_1^2 n} \left(B|I|^2 + \sum_{i \notin I} B|I| \wedge \mathbf{r}(i) - \sum_{i \in I} \mathbf{r}(i) \right) - c_3 |I|^2 \geq 0. \tag{1.25}$$

Conversely, if the above condition holds for constants $c_1, c_2 \in (0, B)$ and $c_3 > 0$, then $(\mathbf{r}, \mathbf{r})$ is δ-tame for some $\delta = \delta(\mu, c_1, c_2, c_3) > 0$.

In 2011, Chatterjee, Diaconis, and Sly showed that the condition (1.25) implies uniform boundedness of the MLE for the $\boldsymbol{\beta}$-model with degree sequence $\mathbf{r}$ [CDS11]. They conjectured that this condition is equivalent to the δ-tameness of Barvinok and Hartigan [BH13] defined in terms of the typical table. The special case of our Theorem 1.24 for Bernoulli base measure in conjunction with the strong duality in Theorem 1.7 establishes this conjecture. See Sec. 2.7 for more discussion.

1.6. **Generalized Sinkhorn algorithm for computing the MLEs.** In most cases, there is no closed-form expression for an MLE for a given margin $(\mathbf{r}, \mathbf{c})$. Based on the classical alternating maximization principle applied to the MLE problem (1.6), we propose the following iterative algorithm for computing an MLE for a given margin $(\mathbf{r}, \mathbf{c})$, which we call the *generalized Sinkhorn* algorithm:

$$\textbf{Generalized Sinkhorn} \quad \begin{cases} \text{For } 1 \leq j \leq n,\ \boldsymbol{\beta}_k(j) \leftarrow \text{unique } \beta \in \mathbb{R} \text{ s.t. } \mathbf{c}(j) = \sum_{i=1}^m \psi'(\boldsymbol{\alpha}_{k-1}(i) + \beta), \\ \text{For } 1 \leq i \leq m,\ \boldsymbol{\alpha}_k(i) \leftarrow \text{unique } \alpha \in \mathbb{R} \text{ s.t. } \mathbf{r}(i) = \sum_{j=1}^n \psi'(\alpha + \boldsymbol{\beta}_k(j)). \end{cases} \tag{1.26}$$

To explain the connection to the Sinkhorn algorithm in the Schrödinger bridge and the optimal transport literature, we first note that the typical table $Z^{\mathbf{r},\mathbf{c}}$ in (1.7) can be viewed as the static Schrödinger bridge between the row and column margins $\mathbf{r}$ and $\mathbf{c}$ with divergence $D(\mu_{\phi(\cdot)} \| \mu)$ and the uniform prior measure. Also, the MLE problem (1.6) corresponds to its Kantorovich dual, and the algorithm (1.26) is exactly the Sinkhorn algorithm that computes the Schrödinger potentials. Taking the Poisson base measure $\mu = \text{Poisson}(1)$, our divergence $D(\mu_{\phi(\cdot)} \| \mu)$ reduces to the KL-divergence, the standard choice in the Schrödinger bridge and the optimal transport literature [FL89, Cut13, Léo13]. See Sec. 2.1 and 2.2 for more discussions.

Even with the connection to the existing literature we mentioned above, convergence analysis of the generalized Sinkhorn (1.26) is limited beyond the KL-divergence case. This is because the updated coordinates $\boldsymbol{\beta}_k(j)$ and $\boldsymbol{\alpha}_k(i)$ in (1.26) in general do not have simple closed-forms as in the KL-divergnece case (see (2.6)). Despite this difficulty, we show that, under mild conditions, the procedure (1.26) for arbitrary base measure μ converges at a linear rate in the sense that the dual objective value gap decays exponentially fast. It also yields that the direct sum of the MLEs,

$\boldsymbol{\alpha}_k \oplus \boldsymbol{\beta}_k$, converges exponentially fast to the uniquely determined matrix of maximum-likelihood tilt parameters $\boldsymbol{\alpha}^* \oplus \boldsymbol{\beta}^*$.

Theorem 1.25 (Linear convergence of the generalized Sinkhorn)**.** *Let $(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k)$, $k \geq 0$ denote the iterates produced by the Sinkhorn algorithm (1.26) for some $m \times n$ δ-tame margin $(\mathbf{r}, \mathbf{c})$ such that $\mathcal{T}(\mathbf{r}, \mathbf{c}) \cap (A, B)^{m\times n}$ is non-empty. Fix an MLE $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ and denote $\Delta_k := g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) - g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k)$. For each $\varepsilon > 0$, let $\sigma_1(\varepsilon)^2$ (resp., $\sigma_2(\varepsilon)^2$) denote the infimum (resp., supremum) of ψ'' on $(\phi(A_\varepsilon), \phi(B_\varepsilon))$.*

(i) *(Asymptotic linear convergence) There exists an integer $k_0 = k_0(\mu, \mathbf{r}, \mathbf{c}) \geq 0$ such that the following holds with $\varepsilon = \delta/2$:*

$$\frac{\sigma_1(\varepsilon)^2}{2} \|(\boldsymbol{\alpha}^* \oplus \boldsymbol{\beta}^*) - (\boldsymbol{\alpha}_k \oplus \boldsymbol{\beta}_k)\|_F^2 \leq \Delta_k \leq \left(1 - \frac{\sigma_1(\varepsilon)^4}{\sigma_2(\varepsilon)^4}\right)^{k-k_0} \Delta_{k_0} \quad \text{for all } k \geq k_0. \tag{1.27}$$

(ii) *(Non-asymptotic linear convergence I) Suppose ψ'' is increasing and $\boldsymbol{\alpha}_0 = \mathbf{0}$. Then (1.27) holds with $k_0 = 1$ and $\varepsilon = \delta$.*

(iii) *(Non-asymptotic linear convergence II) Suppose there exists $\varepsilon > 0$ such that*

$$\phi(A_\varepsilon) + 2\|\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}^*\|_\infty \leq \boldsymbol{\alpha}^* \oplus \boldsymbol{\beta}^* \leq \phi(B_\varepsilon) - 2\|\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}^*\|_\infty. \tag{1.28}$$

Then (1.27) holds with $k_0 = 1$.

The asymptotic linear convergence in **(i)** holds for general base measure μ, margin $(\mathbf{r}, \mathbf{c})$ with an MLE, and initialization $\boldsymbol{\alpha}_0$. If ψ'' is increasing (e.g., when μ is Gaussian, Poisson, Lebesgue measure on $\mathbb{R}_{\geq 0}$, Gamma, and counting measure on $\mathbb{Z}_{\geq 0}$; see Sec. 3), then the linear convergence holds for all iterates with zero initialization by **(ii)**. For ψ'' not necessarily increasing, we can apply **(iii)** for almost all cases of our interest. A sufficient condition for $\varepsilon > 0$ to satisfy (1.28) with zero initialization $\boldsymbol{\alpha}_0 = \mathbf{0}$ is (by shifting the MLE appropriately so that $\|\boldsymbol{\alpha}^*\|_\infty \leq (\phi(B_\delta) - \phi(A_\delta))/2$)

$$\phi(A_\varepsilon) \leq 2\phi(A_\delta) - \phi(B_\delta), \quad 2\phi(B_\delta) - \phi(A_\delta) \leq \phi(B_\varepsilon).$$

One can always find small $\varepsilon > 0$ satisfying the above provided $\phi(A) < 2\phi(A_\delta) - \phi(B_\delta)$ and $2\phi(B_\delta) - \phi(A_\delta) < \phi(B)$. The latter conditions always hold if μ has arbitrary exponential moments so that $\Theta^\circ = \mathbb{R}$ (e.g., μ has compact support or μ = Poisson), since then $\phi(B) = \infty$ and $\phi(A) = -\infty$. Even when Θ° is a proper subset of $\mathbb{R}$, (1.28) still holds for some $\varepsilon > 0$ if $\boldsymbol{\alpha}_0$ is sufficiently close to $\boldsymbol{\alpha}^*$ so that the upper and lower bounds there are in $(\phi(A), \phi(B))$.

1.7. **Applications of the general results.** Now we apply our general framework above to establish results on specific aspects of X: (1) The (mixtures of) marginal distributions of the entries in X, (2) concentration of X around the typical table $Z^{\mathbf{r},\mathbf{c}}$ in the cut norm, and (3) the empirical singular value distribution of $X - Z^{\mathbf{r},\mathbf{c}}$. All results are based on transferring properties of the comparison model Y to the conditioned model X. Accordingly, throughout this section, we assume the transference hypotheses are satisfied – Assumption 1.1 for weak transference and Assumption 1.2 for strong transference. For the latter case, we understand all statements to hold for ν-almost all margins, even when not explicitly stated.

1.7.1. *Marginal distributions.* First, we show that a mixture of the entry-wise distributions for X is very close to the corresponding mixture for Y. We denote by $d_{TV}(\cdot, \cdot)$ the total variation distance between probability measures.

Definition 1.26 (Mixture of entry-wise distribution)**.** Let $X \sim \lambda_{\mathbf{r},\mathbf{c},\rho}$. Let ξ_{ij} denote the law of X_{ij} for each i, j. Fix $I \subseteq [m-1]$, $J \subseteq [n-1]$ and define the mixture distributions

$$\widetilde{\xi}_{I,J} := \frac{1}{|I \times J|} \sum_{(i,j)\in I\times J} \xi_{ij} \quad \text{and} \quad \widetilde{\mu}_{\boldsymbol{\alpha}(I)\oplus\boldsymbol{\beta}(J)} := \frac{1}{|I \times J|} \sum_{(i,j)\in I\times J} \mu_{\boldsymbol{\alpha}(i)+\boldsymbol{\beta}(j)}.$$

Theorem 1.27 (Mixture of entry-wise distribution)**.** *Fix a δ-tame margin $(\mathbf{r},\mathbf{c})$ with an MLE $(\boldsymbol{\alpha},\boldsymbol{\beta})$ and let $X \sim \lambda_{\mathbf{r},\mathbf{c},\rho}$ and $Y \sim \mu_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}$. Then there exists a constant $C = C(\mu,\delta) > 0$ and an absolute constant $\gamma > 0$ such that each δ-tame margin $(\mathbf{r},\mathbf{c})$, the following holds: If we denote*

$$F_\mu(\mathbf{r},\mathbf{c}) := \begin{cases} C\rho & \text{if } \rho \ge C\sqrt{mn(m+n)}, \\ \gamma(m+n)\log N & \text{if the hypothesis of Thm. 1.13 holds,} \\ C(m\vee n)^{3/2}(\log mn)^2 & \text{if the hypothesis of Thm. 1.14 holds} \end{cases}$$

then we have

$$d_{TV}\left(\widetilde{\xi}_{I,J}, \widetilde{\mu}_{\boldsymbol{\alpha}(I)\oplus\boldsymbol{\beta}(J)}\right) \le \sqrt{\frac{F_\mu(\mathbf{r},\mathbf{c})}{|I\times J|}} + \exp(-F_\mu(\mathbf{r},\mathbf{c})). \tag{1.29}$$

Note that if $\mathbf{r}(i)$ and $\mathbf{c}(j)$ are constant for $i \in I$ and $j \in J$, then ξ_{ij} and $\mu_{\boldsymbol{\alpha}(i)+\boldsymbol{\beta}(j)}$ do not depend on (i,j) over $I\times J$ so the above results bound the total variation distance between the marginal distributions of the entries in X and Y in the block $I\times J$.

Remark 1.28. Under the hypothesis of Theorem 1.13, the requirement of δ-tameness in Theorem 1.27 is not necessary in the sense that δ can depend on $(\mathbf{r},\mathbf{c})$. In particular, (1.29) can be appled to a sequence of non-tame (supercritical) margins such as the one in Proposition 7.3 **(ii)**. This recovers [DLP20, Thm. 2.1] by Dittmer, Lyu, and Pak.

The following result is a continuous analog of Theorem 1.27.

Theorem 1.29. *Keep the same setting as in Theorems 1.19 and 1.27. Then there exists a constant $C = C(\mu,\delta) > 0$ such that the following hold for all $m,n \ge 1$: Fix each measurable sets $S,T \subseteq [0,1]$ of positive Lebesgue measures and let $(U,V) \sim \mathrm{Unif}(S\times T)$. Then*

$$d_{TV}\left(\mathbb{E}[\xi_{\lceil mU\rceil,\lceil nV\rceil}], \mathbb{E}[\mu_{\boldsymbol{\alpha}(U)+\boldsymbol{\beta}(V)}]\right) \le C\sqrt{\frac{F_\mu(\mathbf{r},\mathbf{c})}{\mathrm{Leb}(S\times T)}} + C\|(\mathbf{r},\mathbf{c}) - (\bar{\mathbf{r}}_m,\bar{\mathbf{c}}_n)\|_1^{1/4}.$$

In 2010, Barvinok asked whether one can obtain the asymptotic distribution of an entry in a uniformly random contingency table (for $\mu = \mathrm{Counting}(\mathbb{Z}_{\ge 0})$) for a sequence of cloned margins (see Ex. 1.9). Namely, he conjectured that the limiting marginal distribution is Bernoulli for binary ($b = 1$) contingency tables [Bar10a] and geometric for unbounded ($b = \infty$) contingency tables [Bar10b] with mean given by the corresponding entry in the typical table. The following corollary of Theorem 1.29 establishes these conjectures. Similar results hold with greater generality for possibly non-uniformly weighted contingency tables, for which the limiting marginal distribution of an entry is an exponential tilt of the base measure μ.

Corollary 1.30. *Keep the same setting in Theorem 1.29 with $\rho = 0$. Suppose $(\mathbf{r},\mathbf{c})$ is the k-cloning of some $m_0\times n_0$ margin $(\mathbf{r}_0,\mathbf{c}_0)$ with an MLE $(\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)$ (see Ex. 1.9). Then as $k\to\infty$,*

$$d_{TV}(\xi_{11}, \mu_{\boldsymbol{\alpha}_0(1)+\boldsymbol{\beta}_0(1)}) = \begin{cases} O\left(k^{-1/2}\sqrt{\log k}\right) & \text{if the hypothesis of Thm. 1.13 holds,} \\ O\left(k^{-1/4}\log k\right) & \text{if the hypothesis of Thm. 1.14 holds.} \end{cases}$$

In 2010, Chatterjee, Diaconis, and Sly showed that when X is an $n\times n$ uniformly random doubly stochastic matrix, $n\xi_{11}$ is asymptotically the exponential distribution with mean 1 [CDS14]. Our Corollary 1.30 with $\mu = \mathrm{Leb}(\mathbb{R}_{\ge 0})$ and $\mathbf{r} = \mathbf{c} = n\mathbf{1}_n$ recovers their result and extends it to general cloned margins possibly with non-uniform weights.

1.7.2. *Scaling limit in cut norm.* Next, we show that X concentrates around the typical table $Z^{\mathbf{r},\mathbf{c}}$ in the cut-norm, which is commonly used in the theory of dense graph limits. (See Sec. 6.2 for more discussion on the cut norm. See [BCL$^+$08, BCL$^+$12, Lov12] for more background on the limit theory of dense graphs utilizing cut metric.)

Definition 1.31 (Cut norm for kernels)**.** The *cut-norm* of a kernel W is defined as

$$\|W\|_\square := \sup_{S,T\subseteq[0,1]} \left| \int_{S\times T} W(x,y)\,dx\,dy \right|. \tag{1.30}$$

In 2011, Chatterjee, Diaconis, and Sly showed that uniformly random graphs with a given dense degree sequence converge almost surely to a limiting graphon in the cut metric and also identified the limit. Our Theorem 1.32 below establishes a similar scaling limit under cut norm for random matrices with given margins instead of random graphs with given degree sequences. Our result also provides a non-asymptotic bound on the convergence rate.

Theorem 1.32 (Scaling limit in cut norm)**.** *Keep the same setting as in Theorems 1.19 and 1.27. Let $W^{\mathbf{r},\mathbf{c}}$ be the limiting typical kernel in Theorem 1.19. Then there exists a constant $C = C(\mu,\delta) > 0$ such that the following hold for all $m,n \geq 1$. With probability at least $1 - 2\exp\left(-F_\mu(\mathbf{r},\mathbf{c})\right)$,*

$$\|W_X - W^{\mathbf{r},\mathbf{c}}\|_\square \leq \sqrt{\frac{F_\mu(\mathbf{r},\mathbf{c})}{mn}} + C\sqrt{\|(\mathbf{r},\mathbf{c}) - (\bar{\mathbf{r}}_m,\bar{\mathbf{c}}_n)\|_1}. \tag{1.31}$$

In particular, $W_X \to W^{\mathbf{r},\mathbf{c}}$ in cut norm almost surely as $m,n\to\infty$.

It is interesting to compare Theorem 1.32 above to the result of Chatterjee, Diaconis, and Sly in [CDS11]. Suppse μ is the counting measure on $\{0,1\}$, X is a uniformly random binary matrix with given margin and it converges a.s. in cut norm to the kernel $W^{\mathbf{r},\mathbf{c}}(x,y) = \frac{\exp(\boldsymbol{\alpha}^*(x)+\boldsymbol{\beta}^*(x))}{\exp(\boldsymbol{\alpha}^*(x)+\boldsymbol{\beta}^*(y))+1}$ (see Ex. 3.3) satisfying the continuum margin $(\mathbf{r},\mathbf{c})$. In particular, if $\boldsymbol{\alpha}^* = \boldsymbol{\beta}^*$, then our limiting kernel $W^{\mathbf{r},\mathbf{c}}$ coincides exactly to the limiting graphon in [CDS11]. This holds for symmetric margins since by shifting the MLEs can be made symmetric. Also, viewing X as the adjacency matrix of uniformly random bipartite (resp., directed) graphs with fixed degree sequence, our result gives a bipartite graph (resp., directed graph) analogue of the result in [CDS11].

1.7.3. *Empirical Singular Value Distribution.* Our last application is in analyzing the fluctuation of X around the typical table $Z^{\mathbf{r}_m,\mathbf{c}_n}$. A natural way to do so is to characterize the limiting spectral measure of the 'centered' random matrix $X - Z^{\mathbf{r}_m,\mathbf{c}_n}$ as the margins converge in L^1 to a continuum margin $(\mathbf{r},\mathbf{c})$ as $m,n\to\infty$. The study of spectral measures is a central topic in the random matrix literature (e.g., see [AGZ10] and references therein). However, there is a relatively scarce literature on limiting spectral measures of margin-conditioned random matrices, only for constant linear margins with Lebesgue and counting base measures [CDS14, Ngu14, Wu23]. We aim to address this problem for the *empirical singular value distribution* (ESD) for arbitrary margins and base measures under a mild assumption.

We establish a general result on the limit of ESD of margin-conditioned random matrices by first establishing it for the maximum likelihood tilted model and then transferring it to the conditioned model. For this, we need strong concentration of the ESD of the tilted model so that the strictly super-exponential transference costs (see Theorems 1.10 and 1.14) can be suppressed. Guionnet and Zeitouni [GZ00] provide a sub-Gaussian concentration of ESD when the entries of the random matrix have either bounded support or satisfy a uniform log-Sobolev inequality. Klochkov and Zhivotovskiy [KZ20, Lem. 1.4] show that the entries being uniformly sub-Gaussian is enough to warrant such sub-Gaussian concentration of ESD.

We now state our main result on the limit of ESD of $X - Z^{\mathbf{r}_m,\mathbf{c}_n}$ and $Y - Z^{\mathbf{r}_m,\mathbf{c}_n}$. Put succinctly, the Stieltjes transform of the limiting ESD is completely determined by the limiting 'variance profile' $\psi''(\boldsymbol{\alpha} \oplus \boldsymbol{\beta})$ through the corresponding Dyson equation.

Theorem 1.33 (Limit of ESD)**.** *Keep the same setting as in Theorem 1.19. Suppose $m/n \to \kappa \in (0,1)$.*

(i) *Let $\xi_{m,n}$ denote the empirical measure on the eigenvalues of $\tilde{Y}\tilde{Y}^*$ with $\tilde{Y} = \frac{1}{\sqrt{(m+n)s^*/2}}(Y - Z^{\mathbf{r},\mathbf{c}})$, $Y \sim \mu_{\boldsymbol{\alpha}_m \oplus \boldsymbol{\beta}_n}$, and $s^* := \sup \psi''(\boldsymbol{\alpha} \oplus \boldsymbol{\beta})$. Then there exists a probability measure ξ on $\mathbb{R}$ with support contained in $[0,2]$ such that $\xi_{m,n} \to \xi$ in probability as $m,n \to \infty$ in the weak topology. The Stieltjes transform of ξ is determined by*

$$\langle \tau_\cdot(z) \rangle := \int_{(0,1)} \tau_x(z)\, dx = \int_{\mathbb{R}} \frac{1}{t-z}\, \xi(dt) \qquad \text{for all } z \in \mathbb{H} := \{z \in \mathbb{C} \,|\, \mathrm{Im}(z) > 0\},$$

where $\tau : \mathbb{H} \to L^\infty(0,1)$ is the unique uniformly bounded solution to the Dyson equation

$$-\frac{1}{\tau_\cdot(z)} = z - S\frac{1}{1 + S^*\tau_\cdot(z)} \qquad \text{for all } z \in \mathbb{H}, \tag{1.32}$$

where S denotes the integral operator with kernel $\psi''(\boldsymbol{\alpha} \oplus \boldsymbol{\beta})$ and S^* is the adjoint operator of S.

Furthermore, the measure ξ is absolutely continuous w.r.t. the Lebesgue measure apart from a possible point mass at zero, i.e., there is a number $q^* \in [0,1]$ and a locally Hölder-continuous function $q : (0,\infty) \to [0,\infty)$ such that $\xi(dx) = q^*\delta_0(dx) + q(x)\mathbf{1}(x > 0)dx$.

(ii) *The results in* **(i)** *hold for $\tilde{X} = \frac{1}{\sqrt{(m+n)s^*/2}}(X - Z^{\mathbf{r},\mathbf{c}})$ with $X \sim \lambda_{\mathbf{r}_m,\mathbf{c}_n;\rho}$ in place of $\tilde{Y}$ under the following conditions:*

(A1) (Transference) Any one of the following holds:

(i) $\rho = o(mn)$ and $\rho \geq C_2\sqrt{mn(m+n)}$ for C_2 in Thm. 1.10;

(ii) $\rho = 0$ and the hypothesis of Thm. 1.14 is satisfied.

(A2) (Sub-Gaussian ESD concentration) μ is a sub-Gaussian probability measure.

This result is a rather direct consequence of our main results and existing results in the random matrix theory literature. The ESD of $\tilde{Y}$ is the square-root-transform of the empirical eigenvalue distribution of the inhomogeneous Wishart matrix $\tilde{Y}\tilde{Y}^*$, which is known to depend only on the variance matrix $\psi''(\boldsymbol{\alpha}_m \oplus \boldsymbol{\beta}_n)$ of $\tilde{Y}$ through the corresponding Dyson equation [AEK17]. Our Theorem 1.19 guarantees that the rescaled MLEs $(\bar{\boldsymbol{\alpha}}_m, \bar{\boldsymbol{\beta}}_n)$ converge to the limiting MLE $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ in L^2 so a suitable continuity result on the solution of the Dyson equation (e.g., [AEK19]) would suffice to deduce that the limiting ESD of $\tilde{Y}$ is given by the Dyson equation with the limiting variance profile $\psi''(\boldsymbol{\alpha} \oplus \boldsymbol{\beta})$. Then, the sub-Gaussian concentration of ESD of Wishart matrices ([GZ00] and [KZ20, Lem. 1.4]) and our transference principles allow us to conclude the $\tilde{X}$ and $\tilde{Y}$ have the same limiting ESD under the conditions stated in Theorem 1.33 **(ii)**.

There are some important special cases where we can identify the limiting ESD explicitly. If we consider symmetric constant linear margin, i.e., $\mathbf{r}_n = \mathbf{c}_n = an\mathbf{1}_n$ for some $a \in (A,B)$, then by a symmetry argument the MLEs are given by $\boldsymbol{\alpha}_n = \boldsymbol{\beta}_n = (\phi(a)/2)\mathbf{1}_n$. In particular, the limiting variance profile $\psi''(\boldsymbol{\alpha} \oplus \boldsymbol{\beta})$ is the matrix where all entries equal $\psi''(\phi(a))$. In this case, the solution to the Dyson equation (1.32) gives the Stieltjes transform of the celebrated Marchenko-Pastur quarter-circle law [MP67]. This leads to the following corollary of Theorem 1.33 on the universality of the Marchenko-Pastur law for the ESD for constant linear margins. See Fig. 2 for illustration.

Corollary 1.34 (Universality of quarter-circle law for uniform margins)**.** *Assume uniform margins $\mathbf{r}_n = \mathbf{c}_n = an\mathbf{1}_n$ for some $a \in (A,B)$ and $\rho = 0$. Suppose μ is a sub-Gaussian probability measure satisfying the hypothesis of Theorem 1.14. Then the empirical singular value distribution of $\tilde{X}_n :=$*

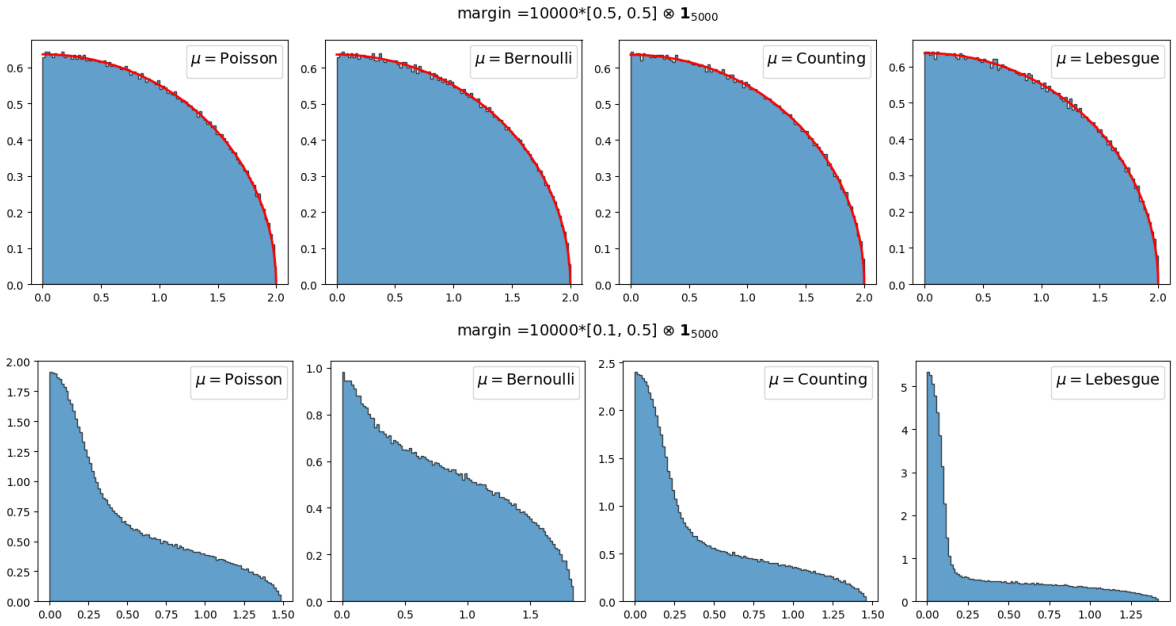

FIGURE 2. Empirical singular value distribution of $(s^*n)^{-1/2}(Y - Z^{\mathbf{r},\mathbf{r}})$ with various base measures μ and exact margin $(\mathbf{r},\mathbf{r})$ with $n^{-1}\mathbf{r} = (a,b) \otimes \mathbf{1}_{n/2}$ for $n = 10^4$. We take $(a,b) = (0.5,0.5)$ and $(0.1,0.5)$. The limiting spectral distribution for the constant margin obeys the quarter-circle law (in red) universally for all μ satisfying the hypothesis of Theorem 1.33. For the non-constant margin, the limit is not quarter-circle and depends on μ.

$(\psi''(\phi(a))n)^{-1/2}(X - a\mathbf{1}_n\mathbf{1}_n^\top)$ for $X \sim \lambda_{\mathbf{r}_n,\mathbf{c}_n}$ converges weakly to the Marchenko-Pastur quarter-circle law $\frac{1}{\pi}\sqrt{4-x^2}\mathbf{1}(x>0)\,dx$ in probability.

Another interesting corollary of Theorem 1.33 is that the limiting ESD of a margin-conditioned standard Gaussian matrix is universally Marchenko-Pastur regardless of the margin. This follows from our transference principle and the fact that the exponential tilting of a Gaussian distribution is just a translation so that the variance profile is not affected by margin conditioning.

Corollary 1.35 (Universality of MP law margin-conditioned standard Gaussian matrices)**.** *Assume μ is standard Gaussian and let $(\mathbf{r}_n,\mathbf{c}_m)$ be a sequence of $m \times n$ margins converging to a continuum margin $(\mathbf{r},\mathbf{c})$ in L^1 with $m/n \to \kappa \in (0,\infty)$. Let $X \sim \lambda_{\mathbf{r}_m,\mathbf{c}_n}$ and let $Z := Z^{\mathbf{r}_m,\mathbf{c}_n}$ be as in (3.1). Then the empirical eigenvalue distribution of $\frac{1}{n}(X-Z)(X-Z)^\top$ converges weakly in probability to the Marchenko-Pastur distribution $(1-\frac{1}{\kappa})\mathbf{1}(\kappa>1)\delta_0(dx) + \frac{\sqrt{(\lambda_+-x)(\lambda_-+x)}}{2\pi x}\mathbf{1}(\lambda_- \le x \le \lambda_+)dx$ with $\lambda_\pm = (1\pm\sqrt{\kappa})^2$.*

Note that in the square case $\kappa = 1$, the limiting ESD of $n^{-1/2}(X-Z)$ in Corollary 1.35 is the quarter-circle law stated in Corollary 1.34.

1.8. **Organization of the paper.** The rest of this paper is organized as follows. We provide background on the related literature and discussion on our results in Section 2. In Section 3, we discuss several examples by specializing the base measure μ. In Section 4, we prove the strong duality result stated in Theorem 1.7. In the following section, Section 5, we prove the transference principles stated in Section 1.3. Theorem 1.19 on scaling limit of the typical tables and the MLEs is proved in Section 6. The results on the phase diagram of tame margins stated in Section 1.5 are

proved in Section 7. In Section 8, we establish Theorem 1.25 on the linear convergence of the generalized Sinkhorn algorithm (1.26) for computing the MLEs. In Section 9, we prove the results on the marginal distribution, scaling limit in the cut norm, and empirical singular value distribution stated in Section 1.7. Lastly in Section 10, we provide concluding remarks and discuss some open problems.

2. Background, Discussions, and Conjectures

Large random matrices conditioned on margins have rich connection to diverse problems such as the static Schrödinger bridge, matrix scaling, relative entropy minimization, enumeration of contingency tables, random graphs with given degree sequences, and spectral distribution of inhomogeneous Wishart matrices. We will review some relevant background and how our results fit in different contexts.

2.1. Typical table, static Schrödinger bridge, and matrix scaling.

A central optimization problem in this work is the typical table problem described in (1.7), which can be viewed as a special instance of the discrete f-divergence static Schrödinger bridge. For a fixed convex function f, the *f-divergence* of a probability measure $\mathscr{H}$ from a measure $\mathscr{R}$ on the same probability space is

$$D_f(\mathscr{H} \,\|\, \mathscr{R}) := \int f\left(\frac{d\mathscr{H}}{d\mathscr{R}}\right) d\mathscr{R} \qquad \text{if } \mathscr{H} \ll \mathscr{R} \text{ and } \infty \text{ otherwise,}$$

where $\frac{d\mathscr{H}}{d\mathscr{R}}$ denotes the Radon-Nikodym derivative of $\mathscr{H}$ w.r.t. $\mathscr{R}$, "$\ll$" denotes absolute continuity between measures. Suppose we have a measure $\mathscr{R}$ representing our prior knowledge on the joint behavior of two random variables, say X_1 and X_2. Consider finding the joint probability distribution $\mathscr{H}$ minimizing the f-divergence from $\mathscr{R}$ with marginal distributions μ_1 and μ_2:

$$\min_{\mathscr{H}} D_f(\mathscr{H} \,\|\, \mathscr{R}) \qquad \text{subject to} \quad \mathscr{H} \in \Pi(\mu_1, \mu_2), \tag{2.1}$$

where $\Pi(\mu_1, \mu_2)$ denotes the set of all joint probability measures that have marginal distributions μ_1 and μ_2. The solution of the above problem is called the *f-divergence static Schrödinger bridge* (SSB) between densities μ_1 and μ_2 with respect to $\mathscr{R}$. The f-divergence D_f becomes the KL-divergence D_{KL} when $f(x) = x \log x$. In this case, (2.1) becomes the classical SSB (see, e.g., [For40, PTT21]). SSBs with general f-divergence have been studied more recently [CDPS17, LM22, TGS22].

Now consider the following discrete setting where the reference measure $\mathscr{R}$ lives on the integer lattice $[m] \times [n]$ and the marginal distributions μ_1 and μ_2 live on $[m]$ and $[n]$, respectively. Writing (2.1) as an optimization problem involving the Radon-Nikodym derivative $X = (x_{ij}) = \frac{d\mathscr{H}}{d\mathscr{R}} \in \mathbb{R}^{m \times n}$,

$$\min_{X \in \mathcal{T}(\mathbf{r},\mathbf{c})} \sum_{i,j} f(x_{ij}) \mathscr{R}(i,j) \quad \text{s.t.} \quad \sum_{j=1}^{n} x_{ij} \mathscr{R}(i,j) = \mu_1(i), \quad \sum_{i=1}^{m} x_{ij} \mathscr{R}(i,j) = \mu_2(j) \ \forall i, j. \tag{2.2}$$

In particular, the typical table $Z^{\mathbf{r},\mathbf{c}}$ in (6.8) is the SSB between the row and column margins w.r.t. the uniform prior measure $\mathscr{R}$ and divergence $f(\cdot) = D(\mu_{\phi(\cdot)} \| \mu)$. Also, choosing μ to be the Poisson measure reduces (2.2) to the KL-divergence SSB problem between the row and column margin w.r.t. $\mathscr{R}$ (see Ex. 3.2 for more details). In a similar manner, if one takes $\mathscr{R}$ to be the Lebesgue measure on $[0,1]^2$ in (2.1) and parameterize the measure $\mathscr{H}$ by $\frac{d\mathscr{H}}{d\mathscr{R}}$ as before, then it reduces to a continuum version of the typical table problem that we define in Section 6.2 (see (6.8)).

Another problem that is closely related to our typical table problem is the classical matrix scaling [Sin64]. Given a matrix R, can one find diagonal matrices D_1 and D_2 such that $S = D_1 R D_2$ has margin $(\mathbf{r}, \mathbf{c})$? It is well-known that such a rescaled matrix S minimizes the KL-divergence $D_{KL}(\cdot \| R)$ among all matrices in $\mathcal{T}(\mathbf{r}, \mathbf{c})$. It is not hard to see that the entrywise 'Radon-Nikodym derivative' $X = S/R$ solves (2.2) with μ Poisson and $\mathscr{R} = R$ (see [Ide16] for a recent review on matrix scaling).

The duality between the typical table and the MLE in Theorem 1.7 bears a striking resemblance to the duality between KL-divergence SSBs and Schrödinger potentials (see, e.g., [Léo13, NW22], [Nut21, Thm. 2.1, 3.2], and [Csi75, Sec. 3 (B)]). Namely, for the KL-divergence SSB, there exist functions $\boldsymbol{\alpha}, \boldsymbol{\beta}$, known as the *Schrödinger potentials*, that characterize the SSB as:

$$\frac{d\mathscr{H}}{d\mathscr{R}}(x,y) = \exp(\boldsymbol{\alpha}(x) + \boldsymbol{\beta}(y)) \qquad \mathscr{R}\text{-a.s.}, \tag{2.3}$$

which may be compared to the characterization of the typical table $Z^{\mathbf{r},\mathbf{c}} = \psi'(\boldsymbol{\alpha} \oplus \boldsymbol{\beta})$ in Theorem 1.7. The Schrödinger potentials solve the following *Kantorovich dual* problem:

$$\sup_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left(\int \boldsymbol{\alpha}(x)\, d\mu_1(x) + \int \boldsymbol{\beta}(y)\, d\mu_2(y) - \iint \exp(\boldsymbol{\alpha}(x) + \boldsymbol{\beta}(y))\, d\mathscr{R}(x,y) \right), \tag{2.4}$$

which is reminiscent of our MLE problem (1.6). Also, the MLE equations (4.2) correspond to the Schrödinger equations determining the Schrödinger potentials. The shift-invariance of the MLEs in (4.1) corresponds to that for the Schrödinger potentials (see, e.g., [Nut21, Lem. 2.11]).

2.2. **Sinkhorn algorithm and entropic optimal transport.** Notice that the log-likelihood function $g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ in (1.6) is concave since ψ is strictly convex. It is not however strictly concave since its value is invariant under adding a constant to $\boldsymbol{\alpha}$ and subtracting the same constant to $\boldsymbol{\beta}$. A natural algorithm for maximizing a concave function in two blocks of variables is alternating maximization, which is also known as block nonlinear Gauss-Seidel [Ber97, BT13]:

$$\begin{cases} \boldsymbol{\beta}_k \leftarrow \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^n} g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha}_{k-1}, \boldsymbol{\beta}), \\ \boldsymbol{\alpha}_k \leftarrow \arg\max_{\boldsymbol{\alpha} \in \mathbb{R}^m} g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha}, \boldsymbol{\beta}_k). \end{cases} \tag{2.5}$$

The log-likelihood function $g^{\mathbf{r},\mathbf{c}}(\cdot,\cdot)$ is strictly concave in each block coordinate, so each block maximization problem in (2.5) reduces to finding the unique stationary points. Hence (2.5) is equivalent to the generalized Sinkhorn algorithm (1.26) we stated in Section 1.

Our algorithm (1.26) is closely related to the celebrated Sinkhorn's algorithm for matrix scaling and computing Schrödinger potentials [FL89, Cut13]. This algorithm iteratively computes the Schrödinger potentials for the discrete SSB (2.2) with the KL-divergence as

$$\begin{cases} \text{For } 1 \le i \le n,\ \boldsymbol{\beta}_k(j) \leftarrow \log\Big(\mathbf{c}(j) \Big/ \sum_{i=1}^m \exp(\boldsymbol{\alpha}_{k-1}(i)) \mathscr{R}(i,j)\Big), \\ \text{For } 1 \le i \le m,\ \boldsymbol{\alpha}_k(i) \leftarrow \log\Big(\mathbf{r}(i) \Big/ \sum_{j=1}^n \exp(\boldsymbol{\beta}_k(j)) \mathscr{R}(i,j)\Big). \end{cases} \tag{2.6}$$

One can derive the above algorithm naturally by applying the alternating maximization procedure to the discrete instance of the Kantorovich dual (2.4). To see the connection with our algorithm (1.26), recall that the divergence $D(\mu_{\phi(\cdot)} \| \mu)$ reduces to the KL-divergence when $\mu = \text{Poisson}(1)$. In fact, in this case $\psi'(x) = e^x$ so we have $\psi'(\alpha+\beta) = \psi'(\alpha)\psi'(\beta)$. This separability simplifies our algorithm (1.26) as (2.6) with $\mathscr{R}(i,j) \equiv 1$. However, for general base measure μ, the iterates in (1.26) does not admit closed-form expressions and hence need to be computed implicitly (e.g., by zero-finding algorithms).

Sinkhorn's algorithm has been extensively studied for a particular instance of the KL-divergence Schrödinger bridge with prior $\mathscr{R}$ taken to be the Gibbs kernel $e^{-c(x,y)/\varepsilon} \mu_1(dx) \otimes d\mu_2(dy)$, where c is a cost function and $\varepsilon > 0$ is the entropic regularization parameter. In this case, it reduces to the celebrated entropic optimal transport problem [Vil21]

$$\min_{\mathscr{H} \in \Pi(\mu_1,\mu_2)} \int c(x,y)\, \mathscr{H}(dx,dy) + \varepsilon D_{KL}(\mathscr{H} \| \mu_1 \otimes \mu_2).$$

The classical case with quadratic cost is when we set $c(x,y) = (x-y)^2$. The corresponding Sinkhorn algorithm in the discrete case is given by (2.6) with both $\exp(\boldsymbol{\alpha}_{k-1}(i))$ and $\exp(\boldsymbol{\beta}_k(j))$ multiplied

by $\mathscr{R}(i,j)$. Franklin and Lorenz showed that the convergence rate of the Sinkhorn in this case is exponential (linear in the log scale) in the space of margins endowed with Hilbert's projective metric [FL89]. A similar result was obtained for the continuous setting by Chen and Pavon [CGP16]. Rüschendorf [Rus95] established asymptotic convergence of the Sinkhorn algorithm in the continuous case from the perspective of information projection. Marino and Gerolin [MG20] extended a similar result to the multi-marginal case. Carlier [Car22] established linear convergence of the multi-marginal Sinkhorn algorithm in the Euclidean metric.

All of the convergence results mentioned above concern the KL-divergence (i.e., Poisson base measure) and the analysis heavily relies on the explicit form of the Sinkhorn iterates, which is not available for general divergences. There are some recent results on asymptotic convergence of the Sinkhorn algorithm for entropic optimal transport with general divergences (e.g., [TGS22, LM22]). Our Theorem 1.25 establishes non-asymptotic linear convergence of the generalized Sinkhorn algorithm (1.26) for general base measures.

2.3. **Relative entropy minimization and typical tables.** The definition of the typical table (Def. 1.6) arises naturally from the classical 'least-action principle'. For our context, it roughly says that the conditioned random matrix in (1.2) should have the 'least-action distribution' that minimizes the relative entropy from the base measure constrained on the expected margins being $(\mathbf{r},\mathbf{c})$. More precisely, consider the following relative entropy minimization problem:

$$\min_{\mathscr{H}\in\mathscr{P}^{m\times n}} D_{KL}(\mathscr{H}\,\|\,\mathscr{R}) \quad \text{subject to} \quad \mathbb{E}_{X\sim\mathscr{H}}[(r(X),c(X))] = (\mathbf{r},\mathbf{c}), \tag{2.7}$$

where $\mathscr{P}^{m\times n}$ denotes the set of all probability measures on $\mathbb{R}^{m\times n}$. This is a matrix version of the standard relative entropy minimization with first moment constraints, which is also an instance of the information projection [Csi75].

Interestingly, it turns out that the variational problem in (2.7) is equivalent to our typical table problem (1.7) when $\mathscr{R}=\mu^{\otimes(m\times n)}$ is a probability measure. Namely, the optimal probability measure $\mathscr{H}$ solving (2.7) is the following product measure

$$\mathscr{H} = \bigotimes_{i,j} \mu_{\phi(z_{ij})}, \tag{2.8}$$

where $Z^{\mathbf{r},\mathbf{c}}=(z_{ij})$ is the typical table for margin $(\mathbf{r},\mathbf{c})$. To see this, notice that (2.7) can be thought of as an optimization problem involving the Radon-Nikodym derivative $h=\frac{d\mathscr{H}(\mathbf{x})}{d\mathscr{R}(\mathbf{x})}:\mathbb{R}^{m\times n}\to[0,\infty)$. By setting the functional derivative of the corresponding Lagrangian equal to zero, we find

$$\log h(\mathbf{x}) \propto \sum_{i,j}(\alpha_i+\beta_j)x_{ij} \qquad \mathscr{R}\text{-a.s.}$$

where α_i and β_j are the Lagrange multipliers corresponding to the constraints on the expectation of the ith row and the jth column sums. Thus, the optimal probability measure $\mathscr{H}$ solving (2.7) is a joint exponential tilting of the base distribution $\mathscr{R}$. When μ is a probability measure on $\mathbb{R}$ and $\mathscr{R}=\mu^{\otimes(m\times n)}$, then the joint tilting of $\mathscr{R}$ is the product of entrywise tilts of μ. Thus we can write $\mathscr{H}=\bigotimes_{i,j}\mu_{\theta_{ij}}$, where θ_{ij} is the unknown tilting parameter for each entry (i,j). Hence reparameterizing exponential tilting by the mean after the tilt, we deduce (2.8) via

$$(2.7) \iff \min_{X=(x_{ij})\in\mathscr{T}(\mathbf{r},\mathbf{c})} D_{KL}\left(\bigotimes_{i,j}\mu_{\phi(x_{ij})}\,\middle\|\,\mu^{\otimes(m\times n)}\right) = (1.7).$$

Information projection arises naturally as the rate function in the theory of large deviations (e.g., Sanov's theorem [San61]). Also, Csiszár's conditional limit theorem [Csi84, Thm. 1] states that i.i.d. samples under the condition that the joint empirical distribution satisfying a convex

constraint are asymptotically 'quasi-independent' (see [Csi84, Def. 2.1]) under the common distribution that solves the corresponding information projection problem.

Our main question (1.2) considers a similar problem of conditioning the joint distribution of a collection of mn i.i.d. random variables, where we first arrange them to form an $m \times n$ random matrix and then condition on its row/column sums. The variables after conditioning on the margin are not exchangeable unless $m = n$ and the row/column sums are constant. Hence, in the general case, one should not expect that the entries of the margin-constrained random matrix follow asymptotically some common law as in Csiszár's conditional limit theorem. However, our transference principles (Sec. 1.3) bear some similarity in that the conditioned joint distribution of all entries is close to the random matrix ensemble with independent entries given by the information projection (2.7). Also, our dual formulation via maximum likelihood estimation is related to the Fenchel dual approach for information projection [BD95].

2.4. **Contingency tables and phase transition.** Contingency tables are $m \times n$ matrices of nonnegative integer entries with prescribed margins with row sums $\mathbf{r} = (r_1, \cdots, r_m)$ and columns sums $\mathbf{c} = (c_1, \cdots, c_n)$, whereby $\mathrm{CT}(\mathbf{r}, \mathbf{c})$ we denote the set of all such tables. They are fundamental objects in statistics for studying dependence structure between two or more variables and also correspond to bipartite multi-graphs with given degrees and play an important role in combinatorics and graph theory, see e.g. see e.g. [Bar09, DG95, DS98]. In the combinatorics literature, counting the number $|\mathrm{CT}(\mathbf{r}, \mathbf{c})|$ of contingency tables has been extensively studied in the past decades [Bar09, BLSY10, CM07, CM10, BH12, LP22]. In the statistics literature, sampling a random contingency table has been heavily investigated, mostly by using rejection-sampling type methods [Sni91, CDHL05, Ver08, Wan20] or Markov-chain Monte Carlo methods [BC89, DG95, Wan20]. These problems are closely related to each other and have many connections and applications to other fields [BLP23] (e.g., testing hypothesis on the co-occurrence of species in ecology [CS79]).

A historic guiding principle to analyzing large contingency tables is the *independent heuristic* (IH), which was introduced by I. J. Good as far back as in 1950 [Goo50]. The heuristic suggests that for a uniformly random table with the given total sum, the row sum and the column sum constraints should act asymptotically independently as the size of the table grows to infinity. In terms of the enumeration problem, IH yields a simple analytic formula for $|\mathrm{CT}(\mathbf{r}, \mathbf{c})|$, say $|\mathrm{IH}(\mathbf{r}, \mathbf{c})|$, which has been verified rigorously to be asymptotically correct when the margins are constant or have a bounded ratio close to one [CM10]. However, in 2009, Barvinok showed that for a sequence of cloned margins (see Ex. 1.9), the row and column constraints have asymptotic positive correlation and that $|\mathrm{CT}(\mathbf{r}, \mathbf{c})|$ is exponentially larger than $|\mathrm{IH}(\mathbf{r}, \mathbf{c})|$ [Bar09].

In 2010, Barvinok and Hartigan introduced the method of maximum entropy [BH10] for estimating the volume of the transportation polytope and the number of integer matrices inside, with subsequent development of the 'typical table' method by Barvinok [Bar10a, Bar10b]. These works revolutionized the analysis of large contingency tables and have inspired this work substantially. The key idea is that a uniformly random contingency table with margins $(\mathbf{r}, \mathbf{c})$ can be approximated by the maximum entropy matrix, which maximizes a specific entropy function over the transportation polytope $\mathscr{T}(\mathbf{r}, \mathbf{c})$. Good's IH is closely related to this maximum entropy approach, except that the entropy is related to that of Poisson distributions [Goo63]. The maximum entropy method treats the entries of random contingency tables as independent Bernoulli or geometric random variables for 0-1 and unbounded tables, respectively. This corresponds precisely to the optimization problem (1.7) for our typical tables with Bernoulli and counting base measures.

In [Bar10b], Barvinok observed that there is a drastic change in the typical tables for $n \times n$ symmetric margins $\mathbf{r} = \mathbf{c} = (\lambda n, n, \ldots, n)$ between $\lambda = 2$ and $\lambda = 3$. Namely, as $n \to \infty$, the typical tables for $\lambda = 2$ stay uniformly bounded (δ-tame) but the $(1,1)$ entry blows up (non-tame) for $\lambda = 3$. Based

on this observation, he conjectured that there is a sharp phase transition in the behavior of the typical table somewhere between $\lambda = 2$ and 3 and that the uniformly random contingency table also exhibits a phase transition behavior. This conjecture was established by Dittmer, Lyu, and Pak in [DP18] except that the behavior of the random contingency table was established for Barvinok margins with a growing number of the larger value λn. They identified that the phase transition occurs at the critical ratio $1+\sqrt{2}$ and the asymptotic distribution of the uniformly random contingency table is geometric whose mean equals the corresponding entry in the typical table. Our Corollaries 1.21 and 1.23 and Theorem 1.27 generalize these results (see also Rmk. 1.28).

2.5. **Empirical spectral distribution of random matrices with given margin.** In comparison to the vast literature on the empirical spectral distribution of various random matrix ensembles, there has not been much work for random matrices with i.i.d. entries conditioned on margins and existing works only consider constant margins. In 2010 (published in 2014) Chatterjee, Diaconis, and Sly [CDS14] established the limiting ESD of large $n \times n$ uniformly random doubly stochastic matrices (nonnegative entries with rows and columns summing to one) is the Marchenko-Pastur quarter-circle law. Adopting their approach with an additional combinatorial argument, Wu [Wu23] recently obtained the quarter-circle law for the empirical singular value distribution of the uniformly random contingency tables with constant linear margins. These works utilize an elementary form of the transference principle by comparing the corresponding random matrices to the ones with i.i.d. entries from the exponential and the geometric distributions with mean 1.

Our Corollary 1.34 establishes that the same quarter-circle law is universal for random matrices with sub-Gaussian entries conditioned to have constant linear margins. However, our result does not directly imply the two earlier results since the entry-wise distributions there after tilting (exponential and geometric, respectively) have only sub-exponential tails. The authors of the aforementioned works overcome super-exponential transference costs by first showing that the maximum of the margin-conditioned random matrices behaves as that of the corresponding comparison models with i.i.d. entries, so the support of the entries can be truncated to $[0, C\log n]$ with high probability. Then one can apply the sub-Gaussian concentration of ESD for bounded entries in [GZ00] based on Talagand's inequality (see, e.g., [Tal96, Thm. 6.6]). We conjecture that the behavior of the maximum entry should remain the same for general base measures and δ-tame margins replacing constant linear margins:

Conjecture 2.1. *Let $(\mathbf{r},\mathbf{c})$ be a $(m\times n)$ δ-tame margin and let $X \sim \lambda_{\mathbf{r},\mathbf{c}}$. Then $\max_{ij} X_{ij} \le C\log(m+n)$ with high probability for some constant $C = C(\mu,\delta) > 0$.*

The authors in [CDS14, Wu23] showed Conjecture 2.1 for constant linear margins for base measures $\mathrm{Leb}(\mathbb{R}_{\ge 0})$ and $\mathrm{Counting}(\mathbb{Z}_{\ge 0})$ by appealing to a precise volume estimate for the transportation polytope in Canfield and McCay [CM07] and Barvinok [Bar09], respectively. This approach does not seem to generalize well to the broader class of base measures μ and non-uniform δ-tame margins. However, at least for the classical Lebesgue or counting base measure cases, we suspect that the above conjecture can be addressed by existing combinatorial techniques, especially when the margin has a small number of blocks.

Our Theorem 1.33 characterizes the limiting ESD for random matrices with i.i.d. sug-Gaussian entries conditioned to have general δ-tame margins. We conjecture that the sub-Gaussianity assumption in Theorem 1.33 **(ii)** is not necessary.

Conjecture 2.2. *Theorem 1.33* **(ii)** *holds without μ being sub-Gaussian.*

One way to address this conjecture is to establish Conjecture 2.1 and use a similar truncation argument as in [CDS14] with our Theorem 1.33 **(i)**.

The authors of [CDS14] conjectured that for the uniformly random doubly stochastic matrices, the limiting empirical *eigenvalue* distribution is the circular law. This conjecture was established by Ngyuen [Ngu14]. We conjecture that the same holds for constant linear margins universally for arbitrary base measure μ that satisfies the hypothesis of Corollary 1.34.

Conjecture 2.3. *Under the same setting as in Corollary 1.34, the empirical eigenvalue distribution of* $\widetilde{X}_n = (2\psi''(\phi(a))\,n)^{-1/2}(X - a\mathbf{1}\mathbf{1}^\top)$ *converges weakly to the circular law in probability.*

2.6. **Random graphs with given degree sequence.** Many scientific applications naturally prompt the investigation of graphs with specific topological constraints, such as a fixed number of edges, triangles, and so forth [YW09, DGKTB10, ODCdS+15]. As the degree sequence is one of the most fundamental observables in network science, the study of random graphs with a given degree sequence has been a major topic in the field in the last decades [MR95, MR98, CL02, CG08, BH13].

We mention two seminal works on this topic. Diaconis, Chatterjee, and Sly [CDS11] and Barvinok and Hartigan [BH13], taking statistical and combinatorial perspectives, respectively. Under a certain condition on the convergent sequence of degree sequences, the authors of [CDS11] showed that uniformly random graphs with a given degree sequence converge almost surely to a limiting graphon in the cut metric and also identified the limit. The limiting graphon is the limit of the expectation of an underlying maximum likelihood inhomogeneous statistical model for random graphs known as the '$\boldsymbol{\beta}$-model'. Similar results were obtained by Barvinok and Hartigan [BH13] from the combinatorial perspective. Their key observation was that the adjacency matrix of a large random graph with a given degree sequence concentrates around a 'maximum entropy matrix', which corresponds to the notion of typical tables for contingency tables [Bar10b]. Our present work generalizes and leverages both of these statistical and combinatorial perspectives. In particular, our Theorem 1.7 shows that these two approaches are in fact in a strong duality.

2.7. **Tame margins and the Erdős-Gallai (EG) condition.** For a sequence of nonnegative integers $d_1 \le \cdots \le d_n$, the necessary and sufficient condition for this sequence to represent the degree sequence of a simple graph with n nodes is given by the Erdős-Gallai (EG) condition [EG60]:

$$\min_{1\le k\le n}\left(k(k-1) + \sum_{i=k+1}^{n} d_i \wedge k - \sum_{i=1}^{k} d_i\right) \ge 0. \tag{2.9}$$

Chatterjee, Diaconis, and Sly [CDS11] assumed the limiting continuum degree distribution satisfies a continuous analog of the EG condition, ensuring the maximum likelihood estimate (MLE) of the underlying inhomogeneous random graph model remains uniformly bounded. Barvinok and Hartigan [BH13] instead assumed the degree sequence is δ-tame, meaning that the typical tables stay uniformly away from boundary values 0 and 1. The same notion of δ-tameness with the counting base measure is also crucial for obtaining asymptotic formulas for the number of contingency tables with unbounded nonnegative integer entries [BLSY10]. These two hypotheses of boundedness of the MLEs and the typical tables are equivalent to our notion of δ-tameness (Def. 1.8) due to the strong duality relation $Z^{\mathbf{r},\mathbf{c}} = \psi'(\boldsymbol{\alpha} \oplus \boldsymbol{\beta})$ in Theorem 1.7.

Our Theorems 1.20 and 1.24 are closely related to the EG condition (2.9) above. Note that if $(A, B) = (0, 1)$, then the inequality (1.21) in Theorem 1.20 determining the phase diagram becomes $(s+t)^2 < 4s$. Also, specializing the EG condition for n sufficiently large for the linear degree sequence with two values sn and tn, one obtains $(s+t)^2 \le 4s$. In fact, Barvinok and Hartigan [BH13, Thm. 2.1] showed that the strict inequality here implies uniform tameness of a degree sequence, which corresponds to the special case of our Theorem 1.20 with $\mu = \mathrm{Uniform}(\{0, 1\})$ restricted to symmetric margins.

In 2011, Chatterjee, Diagonis, and Sly [CDS11, Lem 4.1] showed that if a degree sequence is quadratically deep inside the EG-polytope in the sense of (1.25), then the corresponding MLE for the $\boldsymbol{\beta}$-model is uniformly bounded. They furthermore conjectured that this condition is also equivalent to the δ-tameness in Barvinok and Hartigan [BH13]. Our Theorem 1.24 on symmetric margins (together with Theorem 1.7) establishes this conjecture not just for the Bernoulli base measure, but also for any base measures with bounded support. It would be interesting to obtain an analogous characterization of δ-tameness for general asymmetric margins.

3. Examples

In this section, we discuss various examples of our framework. The first two examples with the Gaussian and the Poisson base measure are 'exactly solvable' in that the conditional law and the typical table can be expressed as an explicit function of the margins.

Example 3.1 (Gaussian base measure)**.** Fix an $m \times n$ margin $(\mathbf{r}, \mathbf{c})$. Suppose the base measure μ is the standard Gaussian distribution on $\mathbb{R}$, given by $\mu(dx) = \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx$. In this case, we have

$$\Theta = \mathbb{R}, \qquad (A, B) = \mathbb{R}, \qquad \psi(\theta) = \frac{\theta^2}{2}, \qquad \psi'(\theta) = \theta, \qquad \phi(x) = x.$$

The function $\psi''(\theta) = 1$ is increasing and log-convex. The titled measure μ_θ is just a Gaussian distribution with mean θ and variance 1. Also, the function $D(\mu_{\phi(x)} \| \mu)$ defined in (1.7) is

$$D(\mu_{\phi(x)} \| \mu) = x\phi(x) - \psi(\phi(x)) = \frac{x^2}{2}.$$

This is the KL-divergence from the standard normal $\mathcal{N}(0,1)$ to shifted normal $\mathcal{N}(x,1)$. The typical table $Z^{\mathbf{r},\mathbf{c}} = (z_{ij})$ that uniquely solves (1.7) in this case is given by

$$z_{ij} = \frac{\mathbf{r}(i)}{n} + \frac{\mathbf{c}(j)}{m} - \frac{N}{mn} \qquad \text{for all } i \in [m] \text{ and } j \in [n]. \tag{3.1}$$

In particular, the margin $(\mathbf{r}, \mathbf{c})$ is δ-tame for a prespecified constant $\delta > 0$ if and only if

$$-\delta^{-1} \le \frac{\mathbf{r}(i)}{n} + \frac{\mathbf{c}(j)}{m} - \frac{N}{mn} \le \delta^{-1} \qquad \text{for all } i \in [m] \text{ and } j \in [n].$$

Thus in this case the set of all δ-tame margins is a convex polytope.

Note that $X \sim \lambda_{\mathbf{r},\mathbf{c}}$ is distributed according to a degenerate Gaussian distribution on $\mathbb{R}^{m \times n}$ supported on the transportation polytope $\mathcal{T}(\mathbf{r}, \mathbf{c})$ satisfying

$$\mathbb{E}[X] = Z^{\mathbf{r},\mathbf{c}}, \quad \mathrm{Cov}(X_{ij}, X_{k\ell}) = \begin{cases} \frac{(m-1)(n-1)}{mn} & \text{if } i = k, j = \ell, \\ -\frac{n-1}{mn} & \text{if } i \neq k, j = \ell, \\ -\frac{m-1}{mn} & \text{if } i = k, j \neq \ell, \\ \frac{1}{mn} & \text{if } i \neq j, k \neq \ell. \end{cases}$$

One can see that the entries of X after conditioning on the margin are asymptotically independent. Also, the correlation between two distinct entries is negative if they belong to the same row/column and positive otherwise. Finally, note that the maximum likelihood tilted model Y in the Gaussian case has the law of $X + Z^{\mathbf{r},\mathbf{c}}$, which has the same conditional law as X given their margins as $(\mathbf{r}, \mathbf{c})$. ▲

Example 3.2 (Poisson base measure)**.** Suppose the base measure μ is the Poisson measure on $\mathbb{N} \cup \{0\}$ given by $\mu(k) = \frac{1}{k!}$ for $k \in \mathbb{Z}_{\ge 0}$. In this case, we have

$$\Theta = \mathbb{R}, \qquad (A, B) = (0, \infty), \qquad \psi(\theta) = e^\theta, \qquad \psi'(\theta) = e^\theta, \qquad \phi(x) = \log x.$$

Note that $\psi''(\theta) = e^\theta$ is increasing and log-convex. The function $D(\mu_{\phi(x)}\|\mu)$ defined in (1.7) is

$$D(\mu_{\phi(x)}\|\mu) = x\phi(x) - \psi(\phi(x)) = x\log x - x.$$

Noting that for any matrix $X = (x_{ij})$ with margin $(\mathbf{r},\mathbf{c})$ the total sum $\sum_{i,j} x_{ij}$ is constant and equals to $N = \sum_{i=1}^m \mathbf{r}(i) = \sum_{j=1}^n \mathbf{c}(j)$, the typical table problem in (1.7) reduces to

$$\operatorname*{argmin}_{X\in\mathscr{T}(\mathbf{r},\mathbf{c})\cap(0,\infty)^{m\times n}} \sum_{i,j} x_{ij}\log x_{ij}. \tag{3.2}$$

Notice that (3.2) is an instance of discrete Schrödinger bridge problem (2.2) with uniform prior measure $\mathscr{R}(i,j) \equiv 1/mn$.

Since $\psi'(\theta) = e^\theta$, the typical table $Z^{\mathbf{r},\mathbf{c}} = (z_{ij})$ that uniquely solves (3.2) takes the following form

$$z_{ij} = e^{\boldsymbol{\alpha}(i)+\boldsymbol{\beta}(j)} \qquad \text{for all } i\in[m] \text{ and } j\in[n]$$

for some dual variables (Schrödinger potentials, see (2.3)) $\boldsymbol{\alpha},\boldsymbol{\beta}$. Invoking the margin condition $Z^{\mathbf{r},\mathbf{c}} \in \mathscr{T}(\mathbf{r},\mathbf{c})$, we find that the typical table is precisely the Fisher-Yates table

$$z_{ij} = \frac{\mathbf{r}(i)\mathbf{c}(j)}{N} \qquad \text{for all } i\in[m] \text{ and } j\in[n]. \tag{3.3}$$

In [Goo63], Good referred the typical table above as the 'independence table' and observed that it maximizes the entropy function $\sum_{i,j}\frac{x_{ij}}{N}\log\frac{N}{x_{ij}}$ subject to the margin constraint, which is equivalent to (3.2) since $\sum_{i,j} x_{ij} = N$ for matrices in the polytope $\mathscr{T}(\mathbf{r},\mathbf{c})$.

The Poisson case is exactly solvable in the sense the margin-conditioned Poisson matrix $X \sim \lambda_{\mathbf{r},\mathbf{c}}$ follows the multivariate hypergeometric (Fisher-Yates) distribution

$$\mathbb{P}(X=\mathbf{x}) = \frac{\prod_{i=1}^m \mathbf{r}(i)!\prod_{j=1}^n \mathbf{c}(j)!}{N!\prod_{i=1}^m\prod_{j=1}^n \mathbf{x}_{ij}!} \tag{3.4}$$

for $\mathbf{x}\in\mathscr{T}(\mathbf{r},\mathbf{c})\cap\mathbb{Z}_{\geq 0}^{m\times n}$. This fact holds for a more general model of independence in the statistic literature [DS98], where X has independent Poisson entries X_{ij} with mean $p_i q_j$ for some $p_i, q_j \geq 0$. Together with (3.3), it follows that the conditional expectation of X is exactly the typical table:

$$\mathbb{E}[X] = \big(\mathbf{r}(i)\mathbf{c}(j)/N\big)_{i,j} = Z^{\mathbf{r},\mathbf{c}}.$$

In [Bar10b, p.5], Barbinok wrote that "*It looks plausible that the independence table in* (3.3) *is close with high probability to the random contingency table* $X\in\mathscr{T}(\mathbf{r},\mathbf{c})$ *if it was sampled from the Fisher-Yates distribution* (3.4) *instead of the uniform distribution.*" Our Theorem 1.32 specialized to $\mu = \text{Poisson}(1)$ confirms this speculation. ▲

Example 3.3 (Binomial base measure)**.** Suppose the base measure μ is the Binomial distribution with parameters $(B,\frac{1}{2})$, given by $\mu(i) = \binom{B}{i}2^{-B}$ for $i\in\{0,1,\cdots,B\}$. For any $\theta\in\mathbb{R}$, the measure μ_θ is the Binomial distribution with parameters $\left(B,\frac{e^\theta}{1+e^\theta}\right)$. We also have

$$\Theta = \mathbb{R}, \qquad (A,B) = (0,B), \qquad \psi(\theta) = B\log\frac{1+e^\theta}{2}, \qquad \psi'(\theta) = \frac{Be^\theta}{1+e^\theta}, \qquad \phi(x) = \log\frac{x}{B-x}.$$

In this case the function $f(x) = D(\mu_{\phi(x)}\|\mu)$ defined in (1.7) is

$$f(x) = x\phi(x) - \psi(\phi(x)) = x\log x + (B-x)\log(B-x) + B\log\frac{2}{B}.$$

The typical table $Z^{\mathbf{r},\mathbf{c}} = (z_{ij})$ that uniquely solves (1.7) is given by

$$z_{ij} = \frac{B}{e^{-\boldsymbol{\alpha}(i)-\boldsymbol{\beta}(j)}+1}, \qquad i\in[m], j\in[n],$$

where $\boldsymbol{\alpha},\boldsymbol{\beta}$ depend on $(\mathbf{r},\mathbf{c})$ implicitly so as to satisfy $Z^{\mathbf{r},\mathbf{c}}\in\mathscr{T}(\mathbf{r},\mathbf{c})$.

In particular, taking $B = 1$ we get the Bernoulli base measure, which was studied using the maximum entropy principle in Barvinok and Hartigan [BH10]. Note that the function f above (up to a constant) is the negative entropy of the Binomial distribution with mean x. Hence the typical table in this case is the mean of the Binomal random matrix with independent entries maximizing the entropy within the transportation polytope $\mathscr{T}(\mathbf{r},\mathbf{c})$.

Unlike the Gaussian and the Poisson case, with the above information alone it is not clear how $X \sim \lambda_{\mathbf{r},\mathbf{c}}$ would look like. Our Theorem 1.14 states that X can be approximated by a random matrix Y with independent Binomial entries with mean $\mathbb{E}[Y] = Z^{\mathbf{r},\mathbf{c}}$. ▲

Example 3.4 (Counting base measure)**.** Suppose the base measure μ is counting measure on $\mathbb{Z}_{\geq 0}$, given by $\mu(i) = 1$ for $i \in \mathbb{Z}_{\geq 0}$. In this case, for any $\theta \in \Theta^\circ$, the measure μ_θ is a Geometric distribution with parameter $p = 1 - e^\theta$. Also we have

$$\Theta = (-\infty, 0), \quad (A, B) = (0, \infty), \quad \psi(\theta) = -\log(1 - e^\theta), \quad \psi'(\theta) = \frac{e^\theta}{1 - e^\theta}, \quad \phi(x) = -\log(1 + x^{-1}).$$

Note that $\psi''(\theta) = e^\theta(1 - e^\theta)^2$ is increasing and log-convex on Θ. The function $f(x) = D(\mu_{\phi(x)} \| \mu)$ defined in (1.7) is

$$f(x) = x\phi(x) - \psi(\phi(x)) = x \log x - (1 + x) \log(1 + x)$$

for $x \in (0, \infty)$. The typical table $Z^{\mathbf{r},\mathbf{c}} = (z_{ij})$ that uniquely solves (1.7) with this f is given by

$$z_{ij} = \frac{1}{e^{-\boldsymbol{\alpha}(i) - \boldsymbol{\beta}(j)} - 1}, \qquad i \in [m], j \in [n],$$

where $\boldsymbol{\alpha}, \boldsymbol{\beta}$ depend on $(\mathbf{r},\mathbf{c})$ implicitly so as to satisfy $Z^{\mathbf{r},\mathbf{c}} \in \mathscr{T}(\mathbf{r},\mathbf{c})$. Recall that $f(x)$ is defined to be the relative entropy from the base measure μ to the geometric distribution $\mu_{\phi(x)}$ on $\{0, 1, \dots\}$ with mean x. Since the base measure μ is the counting measure, it coincides with the negative entropy of $\mu_{\phi(x)}$. Hence our definition of the typical table in this case coincides with that of Barvinok and Hartigan [BH10] formulated through the maximum entropy principle.

Clearly, $X \sim \lambda_{\mathbf{r},\mathbf{c}}$ is uniformly distributed over the set $\mathscr{T}(\mathbf{r},\mathbf{c}) \cap \mathbb{Z}_{\geq 0}^{m \times n}$ of all nonnegative integer-valued contingency tables with margin $(\mathbf{r},\mathbf{c})$. Our Theorem 1.13 states that X can be approximated by a random matrix Y with independent geometric entries with mean $\mathbb{E}[Y] = Z^{\mathbf{r},\mathbf{c}}$. Also, our Theorem 1.32 shows that X is very close to the deterministic matrix $Z^{\mathbf{r},\mathbf{c}}$ in cut norm with high probability as long as the margin is δ-tame for some $\delta > 0$. This is warranted if the ratios between the extreme values in the row/column sums are strictly less than the critical threshold $\lambda_c = 1 + \sqrt{1 + s^{-1}}$, where $s > 0$ is a constant lower bounding the linearly rescaled row and column sums (see Cor. 1.23). ▲

Example 3.5 (Negative binomial base measure)**.** Suppose the base measure μ puts the mass $\mu(i) = \binom{r+i-1}{i}$, for $i \in \mathbb{Z}_{\geq 0}$. Here r is a positive integer, which, following standard negative binomial terminology, can be thought of as the number of heads/successes r, and $\mu(i)$ is the number of ways one can get i tails/failures before getting r heads/successes. We note that this μ is the r-fold convolution of the counting measure on $\mathbb{Z}_{\geq 0}$. In this case, we have

$$\Theta = (-\infty, 0), \quad (A, B) = (0, \infty), \quad \psi(\theta) = -r \log(1 - e^\theta), \quad \psi'(\theta) = \frac{re^\theta}{1 - e^\theta}, \quad \phi(x) = -\log\left(1 + \frac{r}{x}\right).$$

Also $\psi''(\theta) = re^\theta(1 - e^\theta)^2$ is increasing and log-convex on Θ. The tilted measure μ_θ is a negative binomial distribution with parameters $(r, p = 1 - e^\theta)$. The function $f(x) = D(\mu_{\phi(x)} \| \mu)$ defined in (1.7) is just $f(x) = -(x + r)\log(x + r) + x \log x + r \log r$, which is the entropy of the negative binomial

distribution with mean x. The the typical table $Z^{\mathbf{r},\mathbf{c}} = (z_{ij})$ that uniquely solves (1.7) with this f is given by

$$z_{ij} = \frac{r}{e^{-\boldsymbol{\alpha}(i)-\boldsymbol{\beta}(j)} - 1}, \qquad i \in [m], j \in [n],$$

where $\boldsymbol{\alpha}, \boldsymbol{\beta}$ depend on $(\mathbf{r}, \mathbf{c})$ implicitly so as to satisfy $Z^{\mathbf{r},\mathbf{c}} \in \mathcal{T}(\mathbf{r}, \mathbf{c})$.

Our Theorem 1.14 states that $X \sim \lambda_{\mathbf{r},\mathbf{c}}$ can be approximated by a random matrix Y with negative binomial entries with mean $\mathbb{E}[Y] = Z^{\mathbf{r},\mathbf{c}}$. Our Theorem 1.32 yields that X is very close to the typical table $Z^{\mathbf{r},\mathbf{c}}$ in cut norm with high probability as long as the margin is δ-tame for some $\delta > 0$. Note that the threshold for δ-tameness now becomes $\lambda_c = 1 + \sqrt{1 + rs^{-1}}$ (see Cor. 1.23). ▲

Example 3.6 (Lebesgue base measure on positive reals)**.** Suppose the base measure μ is Lebesgue measure on $(0, \infty)$, given by $\mu(dx) = dx$. In this case, for any $\theta < 0$, the measure μ_θ is the Exponential distribution with mean $-\frac{1}{\theta}$. We also have

$$\Theta = (-\infty, 0), \qquad (A, B) = (0, \infty), \qquad \psi(\theta) = -\log(-\theta), \qquad \psi'(\theta) = -\frac{1}{\theta}, \qquad \phi(x) = -\frac{1}{x}.$$

Note that $\psi''(\theta) = \theta^{-2}$ is increasing and log-convex on Θ. The function $f(x) = D(\mu_{\phi(x)} \| \mu)$ defined in (1.7) is just $f(x) = -1 - \log x$, which is the negative entropy of the exponential distribution with mean x. The the typical table $Z^{\mathbf{r},\mathbf{c}} = (z_{ij})$ that uniquely solves (1.7) with this f is given by

$$z_{ij} = \frac{-1}{\boldsymbol{\alpha}(i) + \boldsymbol{\beta}(j)}, \qquad i \in [m], j \in [n], \tag{3.5}$$

where $\boldsymbol{\alpha}, \boldsymbol{\beta}$ depend on $(\mathbf{r}, \mathbf{c})$ implicitly so as to sayisfy $Z^{\mathbf{r},\mathbf{c}} \in \mathcal{T}(\mathbf{r}, \mathbf{c})$. The typical table $Z^{\mathbf{r},\mathbf{c}}$ is also known as the *analytic center* of $\mathcal{T}(\mathbf{r}, \mathbf{c}) \cap \mathbb{R}_{\geq 0}^{m \times n}$, which uniquely maximizes the product of the coordinates over the polytope. It has served a central role in the development of interior point methods, see [Ren88].

Our Theorem 1.13 states that $X \sim \lambda_{\mathbf{r},\mathbf{c}}$ can be approximated by a random matrix Y with independent exponential entries with mean $\mathbb{E}[Y] = Z^{\mathbf{r},\mathbf{c}}$. Chatterjee, Diaconis, and Sly [CDS14] analyzed large uniformly random doubly stochastic matrices, which is the special case of the Lebesgue base measure with constant margin $\mathbf{r} = \mathbf{c} = n\mathbf{1}_n$ and exact margin condition $\rho = 0$. In this case, the maximum likelihood tilted model is simply the $n \times n$ random matrix of i.i.d. exponential entries with mean one and the typical table is $\mathbf{1}_n \mathbf{1}_n^\top$. ▲

Example 3.7 (Gamma distribution)**.** Suppose the base measure μ has Gamma density $\mu(dx) = e^{-ax} x^{\gamma-1} \mathbf{1}(x > 0)\, dx$ with rate parameter $a \geq 0$ and shape parameter $\gamma > 0$. This measure entails several special cases. Taking $a = 1/2$ and $\gamma = k/2$ for $k \geq 1$ an integer, μ becomes χ^2-distribution with k degrees of freedom. Also, taking $a = 0$ and $\gamma = k$, μ becomes the k-fold convolution of the Lebesgue measure on $\mathbb{R}_{\geq 0}$. In particular, it becomes the Lebesgue measure on $\mathbb{R}_{\geq 0}$ when $a = 0$ and $\gamma = 1$. For any $\theta < a$, the tilted measure μ_θ is the Gamma distribution with parameters $(a - \theta, \gamma)$. Also, we have

$$\Theta = (-\infty, a), \quad (A, B) = (0, \infty), \quad \psi(\theta) = \log\Gamma(\gamma) - \gamma \log(a - \theta), \quad \psi'(\theta) = \frac{\gamma}{a - \theta}, \quad \phi(x) = a - \frac{\gamma}{x}.$$

Note that $\psi''(\theta) = \gamma(a - \theta)^{-2}$ is increasing and log-convex on Θ. The function $f(x) = D(\mu_{\phi(x)} \| \mu)$ defined in (1.7) is $f(x) = ax - \gamma + \log\Gamma(\gamma) - \gamma \log(x/\gamma)$, which is the negative entropy of the Gamma distribution with mean x and shape parameter γ. The typical table $Z^{\mathbf{r},\mathbf{c}} = (z_{ij})$ that uniquely solves (1.7) with this f is given by

$$z_{ij} = \frac{\gamma}{a - (\boldsymbol{\alpha}(i) + \boldsymbol{\beta}(j))}, \qquad i \in [m], j \in [n],$$

where $\boldsymbol{\alpha}, \boldsymbol{\beta}$ depend on $(\mathbf{r}, \mathbf{c})$ implicitly so as to satisfy $Z^{\mathbf{r},\mathbf{c}} \in \mathcal{T}(\mathbf{r}, \mathbf{c})$.

Our Theorem 1.14 states that $X \sim \lambda_{\mathbf{r},\mathbf{c}}$ can be approximated by a random matrix Y with independent Gamma entries with mean $\mathbb{E}[Y] = Z^{\mathbf{r},\mathbf{c}}$. Our Theorem 1.32 shows that it is very close to the deterministic matrix $Z^{\mathbf{r},\mathbf{c}}$ in cut norm with high probability as long as the margin is δ-tame for some $\delta > 0$. This is warranted if the ratios between the extreme values in the row/column sums are strictly less than the critical threshold $\lambda_c = 2$ regardless of the shape parameter γ (see Cor. 1.23). ▲

Example 3.8 (Laplace base measure)**.** Suppose the base measure is the Laplace measure with mean 0 and scale parameter $b = 1$, which has a density on $\mathbb{R}$ with respect to the Lebesgue measure given by $\mu(dx) = \frac{1}{2}\exp(-|x|)\,dx$. In this case, we have

$$\Theta = (-1,1), \qquad (A,B) = \mathbb{R}, \qquad \psi(\theta) = -\log(1-\theta^2), \qquad \psi'(\theta) = \frac{2\theta}{1-\theta^2}, \qquad \phi(x) = \frac{x}{1+\sqrt{1+x^2}}$$

In this case the function $f(x) = D(\mu_{\phi(x)} \| \mu)$ defined in (1.7) is just $f(x) = (x+k)/2 - \frac{k}{2}\log x + \frac{k}{2}\log k$. The typical table $Z^{\mathbf{r},\mathbf{c}} = (z_{ij})$ that uniquely solves (1.7) with this f is given by

$$z_{ij} = \frac{2(\boldsymbol{\alpha}(i) + \boldsymbol{\beta}(j))}{1 - (\boldsymbol{\alpha}(i) + \boldsymbol{\beta}(j))^2}, \qquad i \in [m], j \in [n],$$

where $\boldsymbol{\alpha}, \boldsymbol{\beta}$ depend on $(\mathbf{r},\mathbf{c})$ implicitly so as to sayisfy $Z^{\mathbf{r},\mathbf{c}} \in \mathcal{T}(\mathbf{r},\mathbf{c})$. Our Theorem 1.32 shows that it is very close to the deterministic matrix $Z^{\mathbf{r},\mathbf{c}}$ in cut norm with high probability as long as the margin is δ-tame for some $\delta > 0$. However, our Theorem 1.22 on the phase diagram for δ-tame margins does not apply for the Laplace case since ψ'' is not increasing. ▲

Example 3.9 (A counterexample to strong transference)**.** Suppose $\mu = \text{Uniform}(\{0, 1, \sqrt{2}\})$ and $(\mathbf{r},\mathbf{r})$ is an $n \times n$ margin with $\mathbf{r} = n\mathbf{1}_n$. On the one hand, note that $U \sim \mu^{\otimes(n\times n)}$ satisfies $U \in \mathcal{T}(\mathbf{r},\mathbf{r})$ if and only if $U_{ij} \equiv 1$ for all i, j. Thus, the conditioned random matrix $X \sim \lambda_{\mathbf{r},\mathbf{r}}$ is almost surely the all ones matrix $\mathbf{1}_n\mathbf{1}_n^\top$. On the other hand, note that the symmetric MLE is $(\boldsymbol{\alpha},\boldsymbol{\alpha})$ for $(\mathbf{r},\mathbf{r})$ is given by $\boldsymbol{\alpha} = \alpha^*\mathbf{1}_n$ with $\alpha^* = \phi(1)$, where $\phi = (\psi')^{-1}$ with $\psi(x) = \log((1 + \exp(x) + \exp(x\sqrt{2}))/3)$. A direct computation shows $\alpha^* = \log(1 + 2^{-1/2}) \approx 0.6232$. Thus the deterministic matrix $X = \mathbf{1}_n\mathbf{1}_n^\top$ cannot be approximated by the random matrix Y with i.i.d. entries from the non-degenerate probability distribution μ_{α^*}. For instance, the event $X = \mathbf{1}_n\mathbf{1}_n^\top$ occurs almost surely, but the transferred event $Y = \mathbf{1}_n\mathbf{1}_n^\top$ occurs with probability $\mu_{\alpha^*}(\{1\})^{n^2} = \exp(-O(n^2))$, where $\mu_{\alpha^*}(\{1\}) = \exp(\alpha^* - \psi'(\alpha^*))/3 \approx 0.3532$. Consequently, subsequent results about the marginal distribution of an entry and limiting ESD based on strong transference must also not hold for X. For instance, the law of X_{11} is the point mass δ_1 but the law of Y_{11} is μ_{α^*} with support $\{0, 1, \sqrt{2}\}$. Also, $X = \mathbf{1}_n\mathbf{1}_n^\top$ is rank one so its limiting ESD is the point mass δ_0, whereas $Y \sim \mu_{\boldsymbol{\alpha}\oplus\boldsymbol{\alpha}}$ has i.i.d. entries from μ_{α^*}, which has limiting ESD the quarter-circle law. ▲

4. Proof of the strong duality

In this section, we prove Theorem 1.7 on the strong duality between the MLEs and the typical tables.

First we remark that the parameters $\boldsymbol{\alpha} \in \mathbb{R}^m$ and $\boldsymbol{\beta} \in \mathbb{R}^n$ in the $(\boldsymbol{\alpha},\boldsymbol{\beta})$-model are not identifiable, since $(\boldsymbol{\alpha},\boldsymbol{\beta})$ and $(\boldsymbol{\alpha} + \delta\mathbf{1}_m, \boldsymbol{\beta} - \delta\mathbf{1}_n)$ generates the same distribution on $\mathbb{R}^{m\times n}$. More precisely,

$$(\boldsymbol{\alpha},\boldsymbol{\beta})\text{-model} \overset{d}{=} (\boldsymbol{\alpha}',\boldsymbol{\beta}')\text{-model} \iff \exists \lambda \in \mathbb{R} \text{ s.t. } \boldsymbol{\alpha}' - \boldsymbol{\alpha} \equiv \lambda \equiv \boldsymbol{\beta} - \boldsymbol{\beta}', \tag{4.1}$$

where the condition on the right states that the coordinate-wise difference vectors $\boldsymbol{\alpha}' - \boldsymbol{\alpha}$ and $\boldsymbol{\beta} - \boldsymbol{\beta}'$ have a constant value of λ. The above claim is easy to verify. Since we have the total sum condition $\sum_{i=1}^m \mathbf{r}(i) = \sum_{j=1}^n \mathbf{c}(j)$, one of the $m+n$ linear equations that define the polytope $\mathcal{T}(\mathbf{r},\mathbf{c})$ is redundant. Consequently, an MLE, even if it exists, is not unique. Thanks to this shift equivalence

property, the standard MLE uniquely exists if an MLE exists. Also, because of the non-compactness of the parameter space, the MLE may not even exist.

In the lemma below, we deduce a set of equations characterizing MLEs for a given margin $(\mathbf{r},\mathbf{c})$.

Lemma 4.1 (The MLE equation)**.** *$(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})$ is a solution of (1.6) if and only if*

$$\sum_{j=1}^{n}\psi'(\hat{\boldsymbol{\alpha}}(i)+\hat{\boldsymbol{\beta}}(j))=\mathbf{r}(i)\quad \text{for } i\in[m],\qquad \sum_{i=1}^{m}\psi'(\hat{\boldsymbol{\alpha}}(i))+\hat{\boldsymbol{\beta}}(j))=\mathbf{c}(j)\quad \text{for } j\in[n], \tag{4.2}$$

that is, the expected matrix $\psi'(\boldsymbol{\alpha}\oplus\boldsymbol{\beta})=\mathbb{E}_{\mu_{\boldsymbol{\alpha}+\boldsymbol{\beta}}}[Y]$ satisfies the margin $(\mathbf{r},\mathbf{c})$ when $(\boldsymbol{\alpha},\boldsymbol{\beta})=(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})$. We call (4.2) the MLE equation for the $(\boldsymbol{\alpha},\boldsymbol{\beta})$-model.

Proof. The log-likelihood function $g^{\mathbf{r},\mathbf{c}}$ in (1.6) is smooth and concave and the set of admissible dual variables $(\boldsymbol{\alpha},\boldsymbol{\beta})$ for the MLE problem (1.6) is open. It follows that $(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})$ is a global maximizer of $g^{\mathbf{r},\mathbf{c}}$ if and only if it is a critical point of $g^{\mathbf{r},\mathbf{c}}$, i.e., $\nabla g^{\mathbf{r},\mathbf{c}}(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})=\mathbf{0}$. This holds precisely if the system of equations (4.2) hold. □

Instead of arguing the existence of MLEs directly by solving the MLE equations, we will first show that the typical table is uniquely defined under a mild condition and then establish equivalence between typical tables and MLEs. The key advantage in analyzing the typical table first is that the objective function H defining the usual table in (1.7) is strictly convex. To see this, denote $f(x)=D(\mu_{\phi(x)}\|\mu)$ defined in (1.8). Note that f is strictly convex since

$$f'(x)=\phi(x)+x\phi'(x)-\psi'(\phi(x))\phi'(x)=\phi(x),\quad f''(x)=\phi'(x)=\frac{1}{\psi''(\phi(x))}=\frac{1}{\mathrm{Var}(\mu_{\phi(x)})}>0. \tag{4.3}$$

Thus the objective function H in (1.7) is strictly convex on the set $\mathscr{T}(\mathbf{r},\mathbf{c})\cap(A,B)^{m\times n}$. As long as this set it is nonempty, there is a unique minimizer of H, which is exactly the typical table $Z^{\mathbf{r},\mathbf{c}}$.

Lemma 4.2 (Existence and uniqueness of typical table)**.** *For an $m\times n$ margin $(\mathbf{r},\mathbf{c})$, the typical table $Z^{\mathbf{r},\mathbf{c}}$ in (1.7) exists if and only if the set $\mathscr{T}(\mathbf{r},\mathbf{c})\cap(A,B)^{m\times n}$ is non-empty. Furthermore, the typical table is unique if it exists.*

Proof. If a typical table $Z^{\mathbf{r},\mathbf{c}}$ exists, it belongs to the intersection $\mathscr{T}(\mathbf{r},\mathbf{c})\cap(A,B)^{m\times n}$. To show the other direction, suppose there exists some $W=(w_{ij})\in\mathscr{T}(\mathbf{r},\mathbf{c})\cap(A,B)^{m\times n}$. Then there exists $\delta>0$ such that $W\in[A_\delta,B_\delta]^{m\times n}$ (see Def. 1.8). Also $H(W)<\infty$ so $\inf_{Z\in\mathscr{T}(\mathbf{r},\mathbf{c})}H(Z)<\infty$. Let $Z^{(k)}$ be a sequence of matrices in $\mathscr{T}(\mathbf{r},\mathbf{c})\cap(A,B)^{m\times n}$ satisfying

$$H(Z^{(k)})\le\inf_{Z\in\mathscr{T}(\mathbf{r},\mathbf{c})}H(Z)+\frac{1}{k}$$

We begin by showing that for any i,j,

$$A<\liminf_{k\to\infty}z_{ij}^{(k)}\le\limsup_{k\to\infty}z_{ij}^{(k)}<B. \tag{4.4}$$

By passing to a subsequence, assume that $z_{ij}^{(k)}\to z_{ij}^{(\infty)}\in[A,B]$ for all ij, where $Z^{(\infty)}=(z_{ij}^{(\infty)})_{ij}$ is a (possibly) extended real-valued matrix.

For any $\lambda\in[0,1]$ let $Z^{(k,\lambda)}:=(1-\lambda)Z^{(k)}+\lambda W$ and note that $Z^{(k,\lambda)}\in\mathscr{T}(\mathbf{r},\mathbf{c})\cap[A_{\lambda\delta},B_{\lambda\delta}]^{m\times n}$. Hence for all sufficiently large $k\ge1$,

$$H(Z^{(k)})\le\inf_{Z\in\mathscr{T}(\mathbf{r},\mathbf{c})\cap[A_{\lambda\delta},B_{\lambda\delta}]^{m\times n}}H(Z)+\frac{1}{k}\le H(Z^{(k,\lambda)})+\frac{1}{k}.$$

By convexity of H,

$$\frac{1}{k\lambda}\ge\frac{H(Z^{(k)})-H(Z^{(k,\lambda)})}{\lambda}\ge\langle\nabla H(Z^{(k,\lambda)}),Z^{(k)}-W\rangle=\sum_{i,j}\phi\Big(z_{ij}^{(k,\lambda)}\Big)(z_{ij}^{(k)}-w_{ij}), \tag{4.5}$$

where we used (4.3). Letting $k \to \infty$ followed by $\lambda \searrow 0$ in (4.5) we get

$$\begin{aligned} 0 &\geq -\sum_{ij} \phi(z_{ij}^{(\infty)})(z_{ij}^{(\infty)} - w_{ij}) \\ &= \sum_{(i,j)\in \mathscr{I}_A} \phi(A)(A - w_{ij}) + \sum_{(i,j)\in \mathscr{I}_B} \phi(B)(B - w_{ij}) + \sum_{(i,j)\in \mathscr{I}_{A,B}} \phi(z_{ij}^{(\infty)})(z_{ij}^{(\infty)} - w_{ij}), \end{aligned}$$

where

$$\mathscr{I}_A := \{(i,j) : z_{ij}^{(\infty)} = A\}, \quad \mathscr{I}_B = \{(i,j) : z_{ij}^{(\infty)} = B\}, \quad \mathscr{I}_{A,B} := \{(i,j) : z_{ij}^{(\infty)} \in (A,B)\}.$$

Note that $\mathscr{I}_A = \emptyset$ if $A = -\infty$ and $\phi(A) = -\infty$ if A is finite. Similarly, $\mathscr{I}_B = \emptyset$ if $B = \infty$ and $\phi(B) = \infty$ if B is finite. Since the third term above is finite, the above equality holds only if $\mathscr{I}_A = \mathscr{I}_B = \emptyset$, which gives (4.4).

Given (4.4), we have the existence of $\delta > 0$ such that

$$\inf_{Z\in\mathscr{T}(\mathbf{r},\mathbf{c})} H(Z) = \inf_{Z\in\mathscr{T}(\mathbf{r},\mathbf{c})\cap[A_\delta,B_\delta]^{m\times n}} H(Z).$$

The right-hand side has a minimizer since it is for minimizing a strictly convex function $H(\cdot)$ over a compact set. □

Now we establish Theorem 1.7.

Proof of Theorem 1.7. We have already shown in Lemma 4.2 that $Z^{\mathbf{r},\mathbf{c}}$ exists if and only if $\mathscr{T}(\mathbf{r},\mathbf{c}) \cap (A,B)^{m\times n}$ is non-empty. Next, we show this set is non-empty if and only if an MLE exists. If an MLE $(\boldsymbol{\alpha},\boldsymbol{\beta})$ exists, then $\psi'(\boldsymbol{\alpha}\oplus\boldsymbol{\beta}) \in \mathscr{T}(\mathbf{r},\mathbf{c})\cap(A,B)^{m\times n}$ by Lemma 4.1. Conversely, if this set is non-empty, then $Z^{\mathbf{r},\mathbf{c}}$ exists. From (4.3), we have $\nabla H(Z) = \big(\phi(z_{ij})\big)_{ij}$. Since ϕ is differentiable, we can apply the multivariate Lagrange multiplier method, to conclude the existence of dual variables $\boldsymbol{\alpha} \in \mathbb{R}^m$ and $\boldsymbol{\beta} \in \mathbb{R}^n$ such that $\phi(Z) = \boldsymbol{\alpha}\oplus\boldsymbol{\beta}$, or equivalently, $Z = \psi'(\boldsymbol{\alpha}\oplus\boldsymbol{\beta})$. Since Z satisfies the margins $(\mathbf{r},\mathbf{c})$, Lemma 4.1 implies that $(\boldsymbol{\alpha},\boldsymbol{\beta})$ is an MLE.

Next, we show the first two equivalences in (1.9). The first equivalence is shown in Lemma 4.1. For the second equivalence, in the paragraph above we have shown $Z^{\mathbf{r},\mathbf{c}} = \psi'(\boldsymbol{\alpha}\oplus\boldsymbol{\beta})$ for some $(\boldsymbol{\alpha},\boldsymbol{\beta})$, and it has to be an MLE by Lemma 4.1 since $Z^{\mathbf{r},\mathbf{c}}$ has margin $(\mathbf{r},\mathbf{c})$. Conversely, suppose there exists $\boldsymbol{\alpha},\boldsymbol{\beta}$ such that $Z := \psi'(\boldsymbol{\alpha}\oplus\boldsymbol{\beta}) \in \mathscr{T}(\mathbf{r},\mathbf{c})$. We claim that $Z = Z^{\mathbf{r},\mathbf{c}}$. Since Z has margin $(\mathbf{r},\mathbf{c})$, it follows that $(\boldsymbol{\alpha},\boldsymbol{\beta})$ is an MLE for $(\mathbf{r},\mathbf{c})$ due to Lemma 4.2. Therefore, the claim follows once we prove the strong duality relation (1.10).

Now we show (1.10). For this, fix an MLE $(\boldsymbol{\alpha},\boldsymbol{\beta})$ and denote $Z = (z_{ij})_{ij} := \psi'(\boldsymbol{\alpha}\oplus\boldsymbol{\beta})$. Then since $Z \in \mathscr{T}(\mathbf{r},\mathbf{c})$ by Lemma 4.1,

$$\begin{aligned} H(Z) &= \sum_{i=1}^m \sum_{j=1}^n (\boldsymbol{\alpha}(i) + \boldsymbol{\beta}(j)) z_{ij} - \sum_{i,j} \psi(\phi(z_{ij})) \\ &= \sum_{i=1}^m \boldsymbol{\alpha}(i) \sum_{j=1}^m z_{ij} + \sum_{j=1}^n \boldsymbol{\beta}(j) \sum_{i=1}^n z_{ij} - \sum_{i,j} \psi(\boldsymbol{\alpha}(i) + \boldsymbol{\beta}(j)) = g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha},\boldsymbol{\beta}), \end{aligned}$$

where $g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha},\boldsymbol{\beta})$ is defined in (1.6). Thus the above yields

$$H(Z) = g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \sup g^{\mathbf{r},\mathbf{c}}(\cdot,\cdot).$$

Also, from the previous part, we can write $Z^{\mathbf{r},\mathbf{c}} = \psi'(\boldsymbol{\alpha}^*\oplus\boldsymbol{\beta}^*)$ for some MLE $(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)$. It follows that

$$\sup g^{\mathbf{r},\mathbf{c}}(\cdot,\cdot) = H(\psi'(\boldsymbol{\alpha}^*\oplus\boldsymbol{\beta}^*)) = H(Z^{\mathbf{r},\mathbf{c}}) \leq H(Z) = g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \sup g^{\mathbf{r},\mathbf{c}}(\cdot,\cdot).$$

Thus all terms that appear above must equal, verifying (1.10). □

We conclude this section with a simple remark on the boundedness of the standard MLE.

Remark 4.3. If $(\boldsymbol{\alpha},\boldsymbol{\beta})$ is the standard MLE for a δ-tame margin $(\mathbf{r},\mathbf{c})$, we have

$$\|(\boldsymbol{\alpha},\boldsymbol{\beta})\|_\infty \le 2\max\{|\phi(A_\delta)|, |\phi(B_\delta)|\}. \tag{4.6}$$

Note that we have $\phi(Z^{\mathbf{r},\mathbf{c}}) = \boldsymbol{\alpha}\oplus\boldsymbol{\beta}$ and $\boldsymbol{\alpha}(1)=0$. Since $(\mathbf{r},\mathbf{c})$ is δ-tame, we get $\phi(A_\delta)\le \boldsymbol{\alpha}(i)+\boldsymbol{\beta}(j)\le \phi(B_\delta)$ for all i,j. Setting $i=1$ and recalling that $\boldsymbol{\alpha}(1)=0$, the above gives $\phi(A_\delta)\le\boldsymbol{\beta}(j)\le\phi(B_\delta)$ for all j. In turn, it follows that

$$\phi(A_\delta)-\phi(B_\delta)\le \boldsymbol{\alpha}(i)\le \phi(B_\delta)-\phi(A_\delta) \quad \text{for all } i.$$

This implies (4.6).

5. Proof of the transference principles

5.1. **The weak transference principle.** In this section, we prove the weak transference principle stated in Theorem 1.10. To prove the second part, we need the following lemma.

Lemma 5.1 (Concentration of quadratic forms for the $(\boldsymbol{\alpha},\boldsymbol{\beta})$-model)**.** *Let $(\mathbf{r},\mathbf{c})$ be an $m\times n$ δ-tame margin and $Y\sim\mu_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}$, where $(\boldsymbol{\alpha},\boldsymbol{\beta})$ is an MLE for the margin $(\mathbf{r},\mathbf{c})$. Define positive constants*

$$L_1 := \min\{\phi(A_\delta)-\phi(A_{\delta/2}), \phi(B_{\delta/2})-\phi(B_\delta)\}, \quad L^+ := \sup_{|s|\le L_1}\psi''(s).$$

Denote $\tilde Y = Y-\mathbb{E}[Y]$. For each $t>0$, $\mathbf{x}\in[-1,1]^m$, and $\mathbf{y}\in[-1,1]^n$,

$$\mathbb{P}\left(\mathbf{x}^\top\tilde Y\mathbf{y}\ge t\|\mathbf{x}\|^2\|\mathbf{y}\|^2\right)\le 2\exp\left(-\|\mathbf{x}\|^2\|\mathbf{y}\|^2\frac{t(t\wedge L_1L_2)}{2L_2}\right). \tag{5.1}$$

Furthermore, for $s\in[0,L_1L_2]$,

$$\mathbb{P}(Y\in\mathscr{T}_{smn}(\mathbf{r},\mathbf{c}))\ge 1-2(3^m+3^n)\exp\left(-\frac{s^2mn}{2L_2}\right). \tag{5.2}$$

In particular, for $\rho=\sqrt{16L_2mn(m+n)}$ and if $m,n\ge1$ are large enough so that $\frac1m+\frac1n\le L_1^2L_2/16$,

$$\mathbb{P}(Y\in\mathscr{T}_\rho(\mathbf{r},\mathbf{c}))\ge 1-2(1/3)^{m+n}. \tag{5.3}$$

Proof. We will first show (5.1). Note that if $T\sim\mu_\theta$, then

$$\mathbb{E}[\exp(s(T-\mathbb{E}[T]))] = \exp(\psi(s+\theta)-\psi(\theta)-s\psi'(\theta)).$$

Write $t' := t\|\mathbf{x}\|^2\|\mathbf{y}\|^2$ and $s'_{ij}=s\mathbf{x}(i)\mathbf{y}(j)$. For each $s\ge0$, we have

$$\mathbb{P}\left(\mathbf{x}^\top\tilde Y\mathbf{y}\ge t'\right)\le\exp\left(-st'+\sum_{i,j}\psi(s'_{ij}+\theta_{ij})-\psi(\theta_{ij})-s'_{ij}\psi'(\theta_{ij})\right).$$

Denote $\varphi(s):=\psi(s+\theta)-\psi(\theta)-s\psi'(\theta)$. Then

$$\varphi'(s)=\psi'(s+\theta)-\psi'(\theta),\quad \varphi'(0)=0,\quad \varphi''(s)=\psi''(s+\theta)\ge0.$$

Hence φ is concave and is minimized at 0. Let $Z=(z_{ij})$ denote the typical table for $(\mathbf{r},\mathbf{c})$. Since $(\mathbf{r},\mathbf{c})$ is δ-tame, by Theorem 1.7, $\mathbb{E}[Y]=Z\in[A_\delta,B_\delta]^{m\times n}$. So we have

$$\phi(A_{\delta/2})\le s+\phi(z_{ij})\le\phi(B_{\delta/2})$$

whenever $|s|\le L_1$. Hence $\sup_{|s|\le L_1}|\varphi''(s)|\le L_2$. Then by Taylor's theorem,

$$\varphi(s)\le\varphi(0)+\varphi'(0)s+\frac{L_2}{2}s^2=\frac{L_2}{2}s^2 \qquad \text{for all } s\in[-L_1,L_1].$$

Applying the above bound for $\theta=\phi(\mathbb{E}[Y_{ij}])$ and $s=s'_{ij}$ all i,j, from the previous inequality we get

$$\mathbb{P}\left(\mathbf{x}^\top\tilde Y\mathbf{y}\ge t\|\mathbf{x}\|^2\|\mathbf{y}\|^2\right)\le\exp\left(-\|\mathbf{x}\|^2\|\mathbf{y}\|^2\left(st-\frac{L_2}{2}s^2\right)\right) \quad \text{for all } s\in[-L_1,L_1], \tag{5.4}$$

where we have also used $\sum_{i,j}\mathbf{x}(i)^2\mathbf{y}(j)^2 = \|\mathbf{x}\|^2\|\mathbf{y}\|^2$. In order to optimize the above bound, denoting $\bar{t} = t \wedge L_1L_2$, write

$$st - \frac{L_2}{2}s^2 = s(t-\bar{t}) + \left(s\bar{t} - \frac{L_2}{2}s^2\right).$$

The quadratic function in s in the second term of the right-hand side above is minimized at $s = \bar{t}/L_2 \in [-L_1, L_1]$ with minimum value $\bar{t}^2/2L_2$. For this choice of s, and noting $t \geq \bar{t}$, the above is at least $\frac{t\bar{t}}{2L_2}$. Hence we deduce (5.1) from (5.4).

Next, we deduce (5.2). Fix $\mathbf{x} \in [-1,1]^m$ and $\mathbf{y} \in [-1,1]^n$. Substituting $t = s\frac{mn}{\|\mathbf{x}\|^2\|\mathbf{y}\|^2}$ for $s \in [0, L_1L_2]$ in (5.1), we get

$$\mathbb{P}\left(\mathbf{x}^\top \tilde{Y}\mathbf{y} \geq smn\right) \leq 2\exp\left(-\frac{smn}{2L_2}\left(\frac{smn}{\|\mathbf{x}\|^2\|\mathbf{y}\|^2} \wedge L_1L_2\right)\right) \leq 2\exp\left(-\frac{s^2mn}{2L_2}\right), \tag{5.5}$$

where the last inequality uses $\|\mathbf{x}\|^2 \leq m$ and $\|\mathbf{y}\|^2 \leq n$. For a vector $\mathbf{w}$, let $\mathbf{w}^+ := \mathbf{w} \wedge \mathbf{0}$ and $\mathbf{w}^- := (-\mathbf{w}) \wedge \mathbf{0}$, where minimum and maximum are applied coordinatewise. Now observe that

$$\|r(\tilde{Y})\|_1 = \|r(\tilde{Y})^+\|_1 + \|r(\tilde{Y})^-\|_1 \leq 2\max_{\mathbf{x}\in\{0,1\}^m} |\mathbf{x}^\top \tilde{Y}\mathbf{1}_n| \leq 2\max_{\mathbf{x}\in\{-1,0,1\}^m} \mathbf{x}^\top \tilde{Y}\mathbf{1}_n. \tag{5.6}$$

Hence by using (5.5),

$$\mathbb{P}\left(\|r(Y) - \mathbf{r}\|_1 \geq smn\right) \leq \left(2\max_{\mathbf{x}\in\{-1,0,1\}^m} \mathbf{x}^\top \tilde{Y}\mathbf{1}_n \geq smn\right) \leq 3^m 2\exp\left(-\frac{s^2mn}{8L_2}\right).$$

A similar upper bound holds for $\|c(\tilde{Y})\|_1$. Thus by a union bound,

$$\mathbb{P}(Y \notin \mathcal{T}_{smn}(\mathbf{r},\mathbf{c})) = \mathbb{P}\left(\max\{\|r(Y)-\mathbf{r}\|_1, \|c(Y)-\mathbf{c}\|_1\} > smn\right) \leq 2(3^m + 3^n)\exp\left(-\frac{s^2mn}{8L_2}\right).$$

Lastly, choose $s = \sqrt{16L_2(m^{-1}+n^{-1})}$. Then $s \leq L_1L_2$ by the hypothesis, so we can apply (5.2). The bound (5.3) then follows immediately. □

Proof of Theorem 1.10. Let $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ denote the standard MLE for $(\mathbf{r},\mathbf{c})$. by (4.6), their L^∞-norm is at most $C = 2\max\{\phi(A_\delta), \phi(B_\delta)\}$. Then by Hölder's inequality,

$$|g^{r(\mathbf{x}),c(\mathbf{x})}(\boldsymbol{\alpha},\boldsymbol{\beta}) - g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha},\boldsymbol{\beta})| \leq C\|(r(\mathbf{x}),c(\mathbf{x})) - (\mathbf{r},\mathbf{c})\|_1 \leq C\rho =: D.$$

Hence

$$\sup_{\mathbf{x}\in\mathcal{T}_\rho(\mathbf{r},\mathbf{c})} \exp\left(g^{r(\mathbf{x}),c(\mathbf{x})}(\boldsymbol{\alpha},\boldsymbol{\beta})\right) \leq \exp\left(g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha},\boldsymbol{\beta}) + D\right),$$
$$\inf_{\mathbf{x}\in\mathcal{T}_\rho(\mathbf{r},\mathbf{c})} \exp\left(g^{r(\mathbf{x}),c(\mathbf{x})}(\boldsymbol{\alpha},\boldsymbol{\beta})\right) \geq \exp\left(g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha},\boldsymbol{\beta}) - D\right).$$

By Assumption 1.1, $\mu^{\otimes(m\times n)}\left(\mathcal{T}_\rho(\mathbf{r},\mathbf{c})\right) \in (0,\infty)$. Hence for each measurable set $\mathcal{E} \subseteq \mathbb{R}^{m\times n}$,

$$\begin{aligned}\mathbb{P}(Y \in \mathcal{E} \mid Y \in \mathcal{T}_\rho(\mathbf{r},\mathbf{c})) &= \frac{\int_{\mathcal{T}_\rho(\mathbf{r},\mathbf{c})} \mathbf{1}(\mathbf{x}\in\mathcal{E})\exp(g^{r(\mathbf{x}),c(\mathbf{x})}(\boldsymbol{\alpha}',\boldsymbol{\beta}'))\,\mu^{\otimes(m\times n)}(d\mathbf{x})}{\int_{\mathcal{T}_\rho(\mathbf{r},\mathbf{c})}\exp(g^{r(\mathbf{x}),c(\mathbf{x})}(\boldsymbol{\alpha},\boldsymbol{\beta}))\,\mu^{\otimes(m\times n)}(d\mathbf{x})} \\ &\geq \frac{\exp\left(g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha}',\boldsymbol{\beta}') - D\right)}{\exp\left(g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha}',\boldsymbol{\beta}') + D\right)} \frac{\int_{\mathcal{T}_\rho(\mathbf{r},\mathbf{c})}\mathbf{1}(\mathbf{x}\in\mathcal{E})\,\mu^{\otimes(m\times n)}(d\mathbf{x})}{\int_{\mathcal{T}_\rho(\mathbf{r},\mathbf{c})}\mu^{\otimes(m\times n)}(d\mathbf{x})} \geq \exp(-2D)\,\mathbb{P}(X\in\mathcal{E}).\end{aligned}$$

This is enough to conclude (1.12). The second part in (1.13) follows immediately from (1.12) and Lemma 5.1. □

5.2. **The strong transference principles.** Next, we prove the strong transference principle stated in Theorem 1.11 under Assumption 1.2 and prove the subsequent transference results in Corollary 1.12 and Theorems 1.13 and 1.14.

We first give some further details on the disintegration construction of $\lambda_{\mathbf{r},\mathbf{c}}$ under Assumption 1.2. Since μ is assumed to be σ-finite, according to [CP97, Thm. 1 and 2], there exists a family of σ-finite Borel measures $\{\lambda_{\mathbf{t}}\}$ on $\mathbb{R}^{m+n}$ that disintegrate $\mu^{\otimes(m\times n)}$ w.r.t. π (i.e., *π-disintegration*):

(i) For η-almost all $\mathbf{t}\in\mathbb{R}^{m+n}$, $\lambda_{\mathbf{t}}$ lives on $\{\pi=\mathbf{t}\}$, that is, $\lambda_{\mathbf{t}}\{\pi\neq\mathbf{t}\}=0$. Also, each $\lambda_{\mathbf{t}}$ is a probability measure.

(ii) For each nonnegative measurable function $h:\mathbb{R}^{m\times n}\to\mathbb{R}$, $\mathbf{t}\mapsto\int h\,d\lambda_{\mathbf{t}}$ is measurable and

$$\int h(\mathbf{x})\,\mu^{\otimes(m\times n)}(d\mathbf{x})=\iint h(\mathbf{x})\,\lambda_{\mathbf{t}}(d\mathbf{x})\,\nu(d\mathbf{t}).$$

Furthermore, such a π-disintegration of $\mu^{\otimes(m\times n)}$ is unique up to an almost-sure equivalence: if $\{\lambda^*_{\mathbf{t}}\}$ is another π-disintegration of $\mu^{\otimes(m\times n)}$, then $\nu\{\mathbf{t}:\lambda_{\mathbf{t}}\neq\lambda^*_{\mathbf{t}}\}=0$.

Since ν is assumed to be σ-finite under Assumption 1.2, the above properties ensure that for ν-almost all $(\mathbf{r},\mathbf{c})\in\mathbb{R}^{m\times n}$, and for each $\mu^{\otimes(m\times n)}$-integrable random variable T, $\int T\,d\lambda_{\mathbf{t}}$ is a version of the conditional expectation $\mathbb{E}[T\,|\,\pi=(\mathbf{r},\mathbf{c})]$. Hence $\lambda_{\mathbf{r},\mathbf{c}}$ gives the law $\mathbb{P}(\cdot\,|\,\pi=(\mathbf{r},\mathbf{c}))$ of the margin-conditioned random matrix X for ν-almost all margins.

Proof of Theorem 1.11. Recall that the the law of Y is absolutely continuous w.r.t. $\mu^{\otimes(m\times n)}$ with probability density $\exp(g^{r(\cdot),c(\cdot)}(\boldsymbol{\alpha},\boldsymbol{\beta}))$, which is a positive constant over each fiber $\pi^{-1}(\mathbf{r},\mathbf{c})$. Let $\nu_Y:=\pi_\#(\mu_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}})$ denote the pushforward of the law $\mu_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}$ of Y on $\mathbb{R}^{m+n}$. Then ν_Y is absolutely continuous w.r.t. $\nu=\pi_\#(\mu^{\otimes(m\times n)})$ with positive density $\exp(g^{\mathbf{r}(\cdot),\mathbf{c}(\cdot)}(\boldsymbol{\alpha},\boldsymbol{\beta}))$. Since ν is assumed to be σ-finite, it follows that ν_Y also σ-finite. Hence [CP97, Thm. 3] implies that the π-disintegration $\{\lambda_{\mathbf{t}}\}$ for $\mu^{\otimes(m\times n)}$ is the π-disintegration for $\mu_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}$ as well. It follows that the law of Y conditional on $Y\in\mathscr{T}(\mathbf{r},\mathbf{c})$ is given by $\lambda_{\mathbf{r},\mathbf{c}}$, which is the law of X conditional on $X\in\mathscr{T}(\mathbf{r},\mathbf{c})$ for ν-almost all margins. This shows **(i)**.

Next, we show **(ii)**. Let $\overline{\lambda}_{\mathbf{r},\mathbf{c}}$ denote the law of $\overline{Y}\in\mathbb{R}^{(m-1)\times(n-1)}$ conditional on $\pi(Y)=(\mathbf{r},\mathbf{c})$. Let $\overline{\mu}_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}$ denote the law of unconditional $\overline{Y}$. Note that $\overline{\lambda}_{\mathbf{r},\mathbf{c}}$ is the pullback of $\lambda_{\mathbf{r},\mathbf{c}}$ via the completion map $\Gamma_{\mathbf{r},\mathbf{c}}$ defined in (1.14). To show $\overline{\lambda}_{\mathbf{r},\mathbf{c}}\ll\overline{\mu}_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}$ for ν-almost all margins, fix $S\subseteq\mathbb{R}^{(m-1)\times(n-1)}$ write $\mathbf{y}=(\overline{\mathbf{y}},\check{\mathbf{y}})$. Using the disintegration property,

$$\mathbb{P}(\overline{Y}\in S)=\mathbb{E}\left[\mathbb{P}\left(\overline{Y}\in S\,\middle|\,r(Y),c(Y)\right)\right]=\int\overline{\lambda}_{\mathbf{r}',\mathbf{c}'}(S)\exp(g^{\mathbf{r}',\mathbf{c}'}(\boldsymbol{\alpha},\boldsymbol{\beta}))\,\nu(d(\mathbf{r}',\mathbf{c}')).$$

Thus if $\overline{\mu}_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}(S)=0$, then $\overline{\lambda}_{\mathbf{r}',\mathbf{c}'}(S)=0$ for ν-almost all margins $(\mathbf{r}',\mathbf{c}')$. This yields the Radon-Nikodym derivative $p_{\mathbf{r},\mathbf{c}}$ for which $\overline{\lambda}_{\mathbf{r},\mathbf{c}}(d\overline{\mathbf{y}})=p_{\mathbf{r},\mathbf{c}}(\overline{\mathbf{y}})\overline{\mu}_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}(d\overline{\mathbf{y}})$ for ν-almost all margins. By using **(i)**, for ν-almost all margins and bounded measurable functions h,

$$\mathbb{E}[h(\overline{X})]=\int h(\overline{\mathbf{x}})\,\overline{\lambda}_{\mathbf{r},\mathbf{c}}(d\overline{\mathbf{x}})=\int h(\overline{\mathbf{y}})\,p_{\mathbf{r},\mathbf{c}}(\overline{\mathbf{y}})\overline{\mu}_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}(d\overline{\mathbf{y}})=\mathbb{E}[p_{\mathbf{r},\mathbf{c}}(\overline{Y})h(\overline{Y})].$$

Now in order to conclude (1.15) from the above, it is enough to justify Bayes' theorem to deduce $p_{\mathbf{r},\mathbf{c}}(\overline{\mathbf{y}})=p_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c})$ for $\overline{\mu}_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}\otimes\nu$-almost all $(\overline{\mathbf{y}},(\mathbf{r},\mathbf{c}))$.

Given $\overline{\mathbf{y}}\in\mathbb{R}^{(m-1)\times(n-1)}$, $\check{\mathbf{y}}\in\mathbb{R}^{m+n-1}$ is in one-to-one correspondence with the margin $(\mathbf{r},\mathbf{c})$ by the relationship $\mathbf{y}_{ij}=\Gamma_{\mathbf{r},\mathbf{c}}(\overline{\mathbf{y}})$ for (i,j) with $i=m$ or $j=n$. Accordingly, we define a new measure ζ on $\mathbb{R}^{m\times n}$ as the pushforward of the law of Y via the one-to-one map $S:\mathbf{y}\mapsto(\overline{\mathbf{y}},(r(\mathbf{y}),c(\mathbf{y})))$. Let π_i for $i=1,2$ denote the projection of $(\overline{\mathbf{y}},(r(\mathbf{y}),c(\mathbf{y})))$ onto the ith coordinate. Clearly $(\pi_1)_\#(\zeta)=\overline{\mu}_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}$ and $(\pi_2)_\#(\zeta)=\nu_Y$. Then $\zeta(d\overline{\mathbf{y}},d(\mathbf{r},\mathbf{c})):=\overline{\lambda}_{\mathbf{r},\mathbf{c}}(d\overline{\mathbf{y}})\,\nu_Y(d(\mathbf{r},\mathbf{c}))$, meaning that the π_2-disintegration of ζ is

given by $\overline{\lambda}_{\mathbf{r},\mathbf{c}}$. Indeed, for nonnegative functions h, using the fact that the law of $\overline{Y}$ given its margin $(\mathbf{r},\mathbf{c})$ is $\overline{\lambda}_{\mathbf{r},\mathbf{c}}$,

$$\mathbb{E}[h(\overline{Y}, r(Y), c(Y))] = \mathbb{E}\left[(h\circ S)(Y)\right] = \mathbb{E}[\mathbb{E}[(h\circ S)(Y) \,|\, r(Y), c(Y)]] \\ = \iint h(\overline{\mathbf{y}}, (\mathbf{r},\mathbf{c}))\, \overline{\lambda}_{\mathbf{r},\mathbf{c}}(d\overline{\mathbf{y}})\, \nu_Y(d(\mathbf{r},\mathbf{c})).$$

Next, we disintegrate the measure ζ using the other projection $\pi_1 : (\overline{\mathbf{y}}, (\mathbf{r},\mathbf{c})) \mapsto \overline{\mathbf{y}}$. Let $\nu_{\overline{\mathbf{y}}}(d(\mathbf{r},\mathbf{c}))$ give the π_1-disintegration of ζ, which are the laws of the margin of Y given $\overline{Y} = \overline{\mathbf{y}}$. Since the push-forward of ζ via this projection map is $\overline{\mu}_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}$, the two different ways of disintegrating ζ give the following Bayes' theorem:

$$\overline{\lambda}_{\mathbf{r},\mathbf{c}}(d\overline{\mathbf{y}})\, \nu_Y(d(\mathbf{r},\mathbf{c})) = \nu_{\overline{\mathbf{y}}}(d(\mathbf{r},\mathbf{c}))\, \overline{\mu}_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}(d\overline{\mathbf{y}}).$$

Using $\mathbb{P}((r(Y),c(Y)) \in E) = \mathbb{E}[\mathbb{P}((r(Y),c(Y)) \in E) \,|\, \overline{Y})]$, we see that $\nu_{\overline{\mathbf{y}}} \ll \nu_Y$ for $\overline{\mu}_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}$-almost all $\overline{\mathbf{y}}$'s. Hence the Radon-Nikodym derivative $p_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c})$ exists such that $\nu_{\overline{\mathbf{y}}}(d(\mathbf{r},\mathbf{c})) = p_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c})\nu_Y(d(\mathbf{r},\mathbf{c}))$. It follows that

$$p_{\mathbf{r},\mathbf{c}}(\overline{\mathbf{y}})\overline{\mu}_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}(d\overline{\mathbf{y}})\nu_Y(d(\mathbf{r},\mathbf{c})) = p_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c})\nu_Y(d(\mathbf{r},\mathbf{c}))\,\overline{\mu}_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}(d\overline{\mathbf{y}}).$$

From this we conclude $p_{\mathbf{r},\mathbf{c}}(\overline{\mathbf{y}}) = p_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c})$ for $\overline{\mu}_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}} \otimes \nu_Y$-almost all $(\overline{\mathbf{y}}, (\mathbf{r},\mathbf{c}))$. Since ν_Y has a positive density w.r.t. ν, this finishes the proof. □

Next, we prove Corollary 1.12. We extract the key mechanism behind this result in the following corollary of Theorem 1.11.

Corollary 5.2 (Sufficient condition for strong transference)**.** *Keep the same setting as in Theorem 1.11. Further assume that there exists a Borel measure ζ on $\mathbb{R}$ such that*

(a) For almost all $\overline{\mathbf{y}}$ under the law of $\overline{Y}$, $\nu_{\overline{\mathbf{y}}}(\cdot) \ll \zeta^{\otimes(m+n-1)}$ with bounded Radon-Nikodym derivative $q_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c})$; and

(b) $\zeta^{\otimes(m+n-1)} \ll \nu$ locally ν-a.s. That is, there exists a σ-finite Borel set V such that $\nu_Y(V) = 1$ and for each $(\mathbf{r}',\mathbf{c}') \in V$, there exists an open neighborhood U of $(\mathbf{r}',\mathbf{c}')$ such that $\zeta^{\otimes(m+n-1)}|_U \ll \nu|_U$.

Then (1.15) in Theorem 1.11 holds with

$$\sup_{\overline{\mathbf{y}}} p_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c}) = \left(\sup_{\overline{\mathbf{y}}} q_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c})\right)\mathbb{E}[q_{\overline{Y}}(\mathbf{r},\mathbf{c})]^{-1}.$$

Proof. We claim that for each $(\mathbf{r},\mathbf{c}) \in V$, the map $\overline{\mathbf{y}} \mapsto p_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c})$ is proportional to $q_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c})$. This will be enough to conclude. Indeed, taking $h \equiv 1$ in (1.15) in Theorem 1.11, we have $\mathbb{E}[p_{\overline{Y}}(\mathbf{r},\mathbf{c})] = 1$. Hence from the claim and since $\nu_Y(V) = 1$, the normalizing constant for $p_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c})$ must be $\mathbb{E}[q_{\overline{Y}}(\mathbf{r},\mathbf{c})]^{-1}$. This and (1.15) yield

$$\mathbb{E}[h(\overline{X})] = \mathbb{E}\left[\frac{q_{\overline{Y}}(\mathbf{r},\mathbf{c})}{\mathbb{E}[q_{\overline{Y}}(\mathbf{r},\mathbf{c})]}h(\overline{Y})\right] = \left(\sup_{\overline{\mathbf{y}}} p_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c})\right)\mathbb{E}[q_{\overline{Y}}(\mathbf{r},\mathbf{c})]^{-1}\mathbb{E}[h(\overline{Y})],$$

as desired.

It remains to justify the claim. For simplicity denote $\tilde{\zeta} := \zeta^{\otimes(m+n-1)}$. First note that the hypothesis, for $\overline{\mu}_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}} \otimes \nu_Y$-almost all $(\overline{\mathbf{y}}, (\mathbf{r},\mathbf{c}))$,

$$p_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c})\, \nu_Y(d(\mathbf{r},\mathbf{c})) = \nu_{\overline{\mathbf{y}}}(d(\mathbf{r},\mathbf{c})) = q_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c})\, \tilde{\zeta}(d(\mathbf{r},\mathbf{c})). \tag{5.7}$$

Fix $(\mathbf{r},\mathbf{c}) \in V$. By the hypothesis, there exists an open neighborhood $U \in \mathbb{R}^{m+n-1}$ of $(\mathbf{r},\mathbf{c})$ such that $\tilde{\zeta}|_U \ll \nu|_U$. Hence the local Radon-Nikodym derivative $\frac{d\tilde{\zeta}|_U}{d\nu_Y|_U}$ exists. From (5.7). it follows that, for

$\overline{\mu}_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}$-almost all $\overline{\mathbf{y}}$,

$$p_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c}) = q_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c})\frac{d\tilde{\zeta}|_U}{d\nu_Y|_U}(\mathbf{r},\mathbf{c}).$$

This shows the claim, as desired. □

Proposition 5.3. *The map $\pi : \mathbb{R}^{m\times n} \to \mathbb{R}^{m+n-1}$ defined by $\pi(\mathbf{x}) = (r_1(\mathbf{x}),\dots,r_{m-1}(\mathbf{x}),c_1(\mathbf{x}),\dots,c_n(\mathbf{x}))$ is an open map.*

Proof. The proof is straightforward and we omit the details. □

Proof of Corollary 1.12. We will verify the hypothesis of Corollary 5.2 holds. Given that, let p^{ij} denote the Radon-Nikodym derivative of $\mu_{\boldsymbol{\alpha}(i)+\boldsymbol{\beta}(j)}$ w.r.t. ζ. Then using in (1.17) for $\nu_{\overline{\mathbf{y}}}$,

$$q_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c}) = \frac{d\nu_{\overline{\mathbf{y}}}}{d\tilde{\zeta}} = \prod_{i=m \text{ or } j=n} p^{ij}(\Gamma_{\mathbf{r},\mathbf{c}}(\overline{\mathbf{y}})_{ij}). \tag{5.8}$$

Specifically, under the hypothesis,

$$p^{ij}(x) = \exp(x(\boldsymbol{\alpha}(i)+\boldsymbol{\beta}(j)) - \psi(\boldsymbol{\alpha}(i)+\boldsymbol{\beta}(j)))p(x). \tag{5.9}$$

It follows that

$$\mathbb{E}[q_{\overline{Y}}(\mathbf{r},\mathbf{c})] = \exp(g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha},\boldsymbol{\beta}))\int\prod_{i,j} p(\Gamma_{\mathbf{r},\mathbf{c}}(\overline{\mathbf{x}})_{ij})\,\zeta^{\otimes(m-1)\times(n-1)}(d\overline{\mathbf{x}}). \tag{5.10}$$

Now we give the remaining details for the discrete and the continuous cases. First, when ζ is the counting measure on $\mathbb{Z}$, then for each $(\mathbf{r},\mathbf{c})$ in the support of ν, $\nu_Y\{(\mathbf{r},\mathbf{c})\} = \mathbb{P}(r(Y)=\mathbf{r}, c(Y)=\mathbf{c}) > 0$ so using the expression in (1.17) for $\nu_{\overline{\mathbf{y}}}$,

$$p_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c}) = \frac{d\nu_{\overline{\mathbf{y}}}}{d\nu_Y}(\mathbf{r},\mathbf{c}) = \frac{\mathbb{P}(\check{Y}=\check{\Gamma}_{\mathbf{r},\mathbf{c}}(\overline{\mathbf{y}}))}{\mathbb{P}(r(Y)=\mathbf{r},c(Y)=\mathbf{c})}.$$

Then we can choose $q_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c}) = \mathbb{P}(\check{Y}=\check{\Gamma}_{\mathbf{r},\mathbf{c}}(\overline{\mathbf{y}}))$ and the counting measure $\zeta^{\otimes(m+n-1)}$ is trivially locally absolutely continuous w.r.t. ν_Y on the support of ν_Y. This justifies the hypothesis of Corollary 5.2. (In fact, the above identity is trivially true in the discrete case without appealing to Corollary 5.2. Also, noting that the supremum of $\mathbb{P}(\check{Y}=\check{\Gamma}_{\mathbf{r},\mathbf{c}}(\overline{\mathbf{y}}))$ over $\overline{\mathbf{y}}$ is one, we deduce that $\sup_{\overline{\mathbf{y}}} p_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c})$ in this case is the reciprocal of the right-hand side of (5.10).)

Next, assume ζ is the Lebesgue measure on $\mathbb{R}$. Note that the support of the law $\overline{\mu}_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}$ of $\overline{Y}$ is $\mathrm{supp}(p)^{m\times n}$, where $\mathrm{supp}(p)$ is the closure of $\mathscr{C} := \{p>0\}$. Since $\mathscr{C}$ is an open subset in $\mathbb{R}$, it follows that $\mathrm{supp}(p)\setminus\mathscr{C}$ has Lebesgue measure zero. Hence letting $V = \pi(\{p>0\}^{m\times n})$ where π is the margin map, we have $\nu_Y(V)=1$. Also, by Proposition 5.3, V is an open subset of $\mathbb{R}^{m+n-1}$. We will verify that Corollary 5.2(b) holds. Indeed, fix $(\mathbf{r},\mathbf{c})\in V$. Since V is open in $\mathbb{R}^{m+n-1}$, we can choose an open neighborhood $U_{\mathbf{r},\mathbf{c}}$ of this margin that is contained in V. Now it must be that $\zeta^{\otimes(m+n-1)} \ll \nu_Y$ on $U_{\mathbf{r},\mathbf{c}}$. If this is not the case, then there must be some open ball $\mathscr{B}\subseteq U_{\mathbf{r},\mathbf{c}}$ such that $\nu_Y(\mathscr{B})=0$, which must have positive Lebesgue measure $\zeta^{\otimes(m+n-1)}(\mathscr{B})>0$. However, $\mathscr{B}\subseteq V$ and by construction $V\subseteq\mathrm{supp}(\nu_Y)$. Hence the open ball $\mathscr{B}$ contains some margin $(\mathbf{r}',\mathbf{c}')$ in V, so $\nu_Y(\mathscr{B})>0$, which is a contradiction. □

We will also use the lower bound on the number of integer-valued contingency tables due to Brändén, Leake, and Pak [BLP23], which were obtained by using Lorentzian polynomials by Brändén and Huh [BH20].

Lemma 5.4. *Let $(\mathbf{r},\mathbf{c})$ be a $(m\times n)$ integer-valued margin such that $0<\mathbf{r}(i)/n<\lfloor D\rfloor$ and $0<\mathbf{c}(j)/m<\lfloor D\rfloor$ for all i,j for some $D>1$. Then the following hold:*

(i) $\mathcal{T}(\mathbf{r},\mathbf{c}) \cap (\{0,1,\dots,\lceil D \rceil\})^{m\times n}$ *is non-empty.*

(ii) *Let* μ *be the counting measure on* $[0,B]\cap \mathbb{Z}_{\geq 0}$ *for any* $B \in \{\lceil D\rceil, \lceil D\rceil+1,\dots\}\cup\{\infty\}$. *Let* $Y \sim \mu_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}$ *where* $(\boldsymbol{\alpha},\boldsymbol{\beta})$ *is an MLE for margin* $(\mathbf{r},\mathbf{c})$ *and let* $N := \sum_i \mathbf{r}(i) = \sum_j \mathbf{c}(j)$. *Then*

$$\mathbb{P}\big(Y \in \mathcal{T}(\mathbf{r},\mathbf{c})\big) \geq N^{-(11/2)(m+n)}.$$

Proof. Note that **(i)** follows immediately from **(ii)** with $B = \lceil D\rceil$, so it suffices to show **(ii)**. Existence of MLE $(\boldsymbol{\alpha},\boldsymbol{\beta})$ for the margin $(\mathbf{r},\mathbf{c})$ follows from Theorem 1.7 since the Fisher-Yates table $\big(\mathbf{r}(i)\mathbf{c}(j)/N\big)_{i,j}$ takes entries from the open interval $(0,B)$. Recall that the log-likelihood of Y at any matrix in $\mathcal{T}(\mathbf{r},\mathbf{c})$ with entries from $[0,B]\cap\mathbb{Z}_{\geq 0}$ w.r.t. the counting measure $\mu^{\otimes(m\times n)}$ is $g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha},\boldsymbol{\beta})$. So

$$\mathbb{P}\big(Y \in \mathcal{T}(\mathbf{r},\mathbf{c})\big) = \exp\big(g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha},\boldsymbol{\beta})\big)\,\mathrm{CT}(\mathbf{r},\mathbf{c}),$$

where $\mathrm{CT}(\mathbf{r},\mathbf{c}) = |\mathcal{T}(\mathbf{r},\mathbf{c})\cap\{0,1,\dots,B\}^{m\times n}|$ denotes the number of contingency tables with margin $(\mathbf{r},\mathbf{c})$ and entries from $\{0,1,\dots,B\}$. Then the result will follow once the following inequality is verified:

$$\mathrm{CT}(\mathbf{r},\mathbf{c}) \geq N^{-(11/2)(m+n)} \exp\big(-g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha},\boldsymbol{\beta})\big). \tag{5.11}$$

To deduce (5.11), we use the lower bound on $\mathrm{CT}(\mathbf{r},\mathbf{c})$ in [BLP23, Thm. 2.1] in conjunction with the discussion in [BLP23, Sec. 7.1] to get

$$\mathrm{CT}(\mathbf{r},\mathbf{c}) \geq N^{-(11/2)(m+n)}\mathrm{Cap}(\mathbf{r},\mathbf{c}),$$

where

$$\mathrm{Cap}(\mathbf{r},\mathbf{c}) := \inf_{\mathbf{x}\in(0,\infty)^m,\mathbf{y}\in(0,\infty)^n}\left(\prod_{i=1}^m x_i^{-r_i}\prod_{j=1}^n y_j^{-c_j}\prod_{i,j}\frac{1-(x_iy_j)^{B+1}}{1-x_iy_j}\right).$$

By making change of variables $x_i \mapsto e^{\boldsymbol{\alpha}(i)}$ and $y_j \mapsto e^{\boldsymbol{\beta}(j)}$,

$$\begin{aligned}\log \mathrm{Cap}(\mathbf{r},\mathbf{c}) &= \inf_{\mathbf{x},\mathbf{y}}\left(-\sum_{i=1}^m \mathbf{r}(i)\log x_i - \sum_{j=1}^n \mathbf{c}(y)\log y_j + \sum_{i,j}\log\frac{1-(x_iy_j)^{B+1}}{1-x_iy_j}\right)\\ &= \inf_{\boldsymbol{\alpha},\boldsymbol{\beta}}\left(-\sum_{i=1}^m \mathbf{r}(i)\boldsymbol{\alpha}(i) - \sum_{j=1}^n \mathbf{c}(j)\boldsymbol{\beta}(j) + \sum_{i,j}\psi(\boldsymbol{\alpha}(i)+\boldsymbol{\beta}(j))\right) = -g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha},\boldsymbol{\beta}).\end{aligned}$$

This is enough to conclude (5.11). □

Proof of Theorem 1.13. First assume μ is the counting measure on $[0,b]\cap Z$ for some $b\in[0,\infty]$. By Lemma 5.4, $\mathbb{P}(Y\in\mathcal{T}(\mathbf{r},\mathbf{c})) \geq N^{-(11/2)(m+n)}$ with $N = \sum_i \mathbf{r}(i)$. Now notice that the upper bound on the transference cost in Corollary 1.12 is simply $\mathbb{P}(Y\in\mathcal{T}(\mathbf{r},\mathbf{c}))^{-1}$ (see the proof of Corollary 1.12; one can also use Theorem 1.10 with $\rho = 0$ instead of Corollary 1.12.)

Next, assume ζ is the Lebesgue measure on $\mathbb{R}$ and let $p(x) = \mathbf{1}(x\geq 0)$. Here we use the lower bound on the volume of the transportation polytope due to Barvinok [Bar09, Thm. 1.2] to get

$$\int\prod_{i,j} h(\Gamma_{\mathbf{r},\mathbf{c}}(\overline{\mathbf{x}})_{ij})\,\zeta^{\otimes(m-1)\times(n-1)}(d\overline{\mathbf{x}}) \geq a\,e^{mn}N^{-\gamma(m+n)},$$

where $\gamma>0$ is an absolute constant and $a = \sup_{\mathbf{x}\in\mathcal{T}(\mathbf{r},\mathbf{c})\cap\mathbb{R}_{\geq 0}^{m\times n}}\prod_{ij}\mathbf{x}_{ij}$. Since $Z^{\mathbf{r},\mathbf{c}}$ minimizes the function $f(x) = -1-\log x$ over $\mathcal{T}(\mathbf{r},\mathbf{c})\cap\mathbb{R}_{\geq 0}^{m\times n}$ (see Ex. 3.6), we have

$$ae^{mn} = e^{-H(Z^{\mathbf{r},\mathbf{c}})} = e^{-g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha},\boldsymbol{\beta})},$$

where the second equality is due to Theorem 1.7.

Lastly, note that the tilted law $\mu_{\boldsymbol{\alpha}(i)+\boldsymbol{\beta}(j)}$ for Y_{ij} is the exponential distribution with mean $Z_{ij}^{\mathbf{r},\mathbf{c}}$, so its density w.r.t. ζ has maximum value $1/Z_{ij}^{\mathbf{r},\mathbf{c}}$. In the proof of [Bar09, Thm. 1.2], Barvinok shows

$$Z_{ij}^{\mathbf{r},\mathbf{c}} \geq N^{-1}\left(\frac{n}{\mathbf{r}(i)}+\frac{m}{\mathbf{c}(j)}\right)^{-1}. \tag{5.12}$$

This lower bound can be justified easily as follows. Recall that $Z_{ij}^{\mathbf{r},\mathbf{c}} = \frac{-1}{\boldsymbol{\alpha}(i)+\boldsymbol{\beta}(j)}$ from (3.5). By shifting the MLE, without loss of generality assume $\boldsymbol{\alpha}(i), \boldsymbol{\beta}(j) < 0$ for all i, j. If $-\boldsymbol{\alpha}(i) > Nn/\mathbf{r}(i)$ for some i, then $Z_{ij}^{\mathbf{r},\mathbf{c}} \leq -1/\boldsymbol{\alpha}(i) < \mathbf{r}(i)/Nn$ for all j, so summing it over j we get $\mathbf{r}(i) < \mathbf{r}(i)/N$, which is a contradiction since $N \geq 1$. Hence $-\boldsymbol{\alpha}(i) \leq Nn/\mathbf{r}(i)$ for all i, and similarly $-\boldsymbol{\beta}(j) \leq Nm/\mathbf{c}(j)$. Then (5.12) follows at once. Hence for $q_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c})$ in Corollary 1.12, using the hypothesis $\mathbf{r}(i) \geq 1, \mathbf{c}(j) \geq 1$,

$$\sup_{\overline{\mathbf{y}}} q_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c}) = \prod_{i=m \text{ or } j=n} \frac{1}{Z_{ij}^{\mathbf{r},\mathbf{c}}} \leq N^{m+n}(n+m)^{m+n} \leq 2N^{2(m+n)}.$$

Thus by Corollary 1.12 we obtain

$$\sup_{\overline{\mathbf{y}}} p_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c}) \leq 2N^{2(m+n)} \exp(-g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha},\boldsymbol{\beta})) \exp(g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha},\boldsymbol{\beta})) N^{\gamma(m+n)} = 2N^{(\gamma+2)(m+n)}.$$

□

Remark 5.5. Although the volume estimate [Bar09, Thm. 1.2] we used in the proof above is stated only for the integer-valued margins, its proof does not require the margins to be integer-valued. A close inspection of the proof reveals that for the upper bound, one needs a uniform lower bound on the row and column sums, which is guaranteed by the assumption that they are positive integer-valued. The lower bound, on the other hand, does not require such an assumption because it is a direct consequence of the general result [Bar09, Thm. 3.1 (1)] on the volume of the affine subspace intersecting the standard symplex.

Remark 5.6. Chatterjee, Diaconis, and Sly obtained a similar result in [CDS14, Lem. 2.1] for constant margins and Lebesgue base measure by relying on (the lower bound of) the volume estimate of the Birkoff polytope (i.e., $\mathscr{T}(\mathbf{1},\mathbf{1})$) due to Canfield and McKay [CM07].

In the rest of this section, we prove Theorem 1.14. Recall that $\Gamma_{\mathbf{r},\mathbf{c}}(\overline{Y})$ is the $m \times n$ matrix obtained from the interior matrix $\overline{Y}$ uniquely completing it so that the margin is $(\mathbf{r},\mathbf{c})$ (see (1.14)). In the following lemma, we show that it is enough to lower bound the probability that the entries in the last row and the last column of $\Gamma_{\mathbf{r},\mathbf{c}}(\overline{Y})$ take values from a compact interval.

Lemma 5.7. *Keep the same setting as in Corollary 1.12. Suppose $(\mathbf{r},\mathbf{c})$ is δ-tame and the Radon-Nikodym derivatives of μ_θ w.r.t. ζ for $\theta \in [\phi(A_\delta), \phi(B_\delta)]$ is uniformly upper bounded by some constant $M = M(\mu,\delta) > 0$. Fix a Borel set $I \subseteq \mathbb{R}$ and assume that $c_I := \inf_{x\in I} p(x) > 0$. Then (1.15) in Theorem 1.11 holds with*

$$\sup_{\overline{\mathbf{y}}} p_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c}) \leq (M/d_I c_I)^{m+n}\, \mathbb{P}\left(\Gamma_{\mathbf{r},\mathbf{c}}(\overline{Y})_{ij} \in I \ \forall (i,j) \text{ s.t. } i=m \text{ or } j=n\right)^{-1} \tag{5.13}$$

for some constant $d_I = d_I(\mu,\delta) > 0$.

Proof. Denote $T = \prod_{i=m \text{ or } j=n} p^{ij}(\Gamma_{\mathbf{r},\mathbf{c}}(\overline{Y})_{ij})$, where p^{ij} denotes the Radon-Nikodym derivatives of μ_θ w.r.t. ζ (see (5.9)). Note that T equals $q_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c}) = dv_{\overline{\mathbf{y}}}/d\zeta^{\otimes(m+n-1)}$ in Corollary 1.12 evaluated at $\overline{Y}$ (see (1.17)). Let $\mathscr{E}$ denote the event in the probability in (5.13). Denote

$$d_I := \inf_{\phi(A_\delta)\leq\theta\leq\phi(B_\delta)} \inf_{x\in I} e^{x\theta-\psi(\theta)} \in (0,\infty).$$

Then on the event $\mathscr{E}$,

$$T \geq \prod_{i=m \text{ or } j=n} \inf_{x \in I} p^{ij}(x) \geq (d_I c_I)^{m+n}.$$

It follows that by Corollary 1.12,

$$\sup_{\bar{\mathbf{y}}} p_{\bar{\mathbf{y}}}(\mathbf{r},\mathbf{c}) \leq M^{m+n}\mathbb{E}[T]^{-1} \leq (M/d_I c_I)^{m+n}\mathbb{P}(\mathscr{E})^{-1}.$$

□

Proof of Theorem 1.14. Let $\mathscr{E}(I)$ denote the event in (5.13). By Lemma 5.7, it suffices to lower bound the probability of $\mathscr{E}(I)$ for some compact interval I in the support of μ. Our argument is based on partitioning Y into a 2×2 block matrix where the size of the $(2,2)$ block is about $\sqrt{m}\times\sqrt{n}$. By concentration inequalities and union bounds, the largest block Y^{11} will have its row and column sums close to its expectations with probability bounded away from zero. On this event, we will show that there is a sufficient probability to assign values of the entries in the remaining three blocks *except* the ones in the last row and the law column in a way that the required values to complete $\overline{Y}$ to $\Gamma_{\mathbf{r},\mathbf{c}}(\overline{Y})$ all lie in a compact interval in the support of μ.

In the rest of the proof, we will use $C = C(\mu,\delta) > 0$ to denote a positive constant depending only on μ and δ whose value can change from line to line.

Denote $m_0 := \lfloor \sqrt{m}\log m \rfloor$, $n_0 := \lfloor \sqrt{n}\log n \rfloor$, $m' := m - m_0$ and $n' := n - m_0$. Partition the row indices into sets $\{1,\dots,m'\}$ and $\{m'+1,\dots,m\}$. Make a similar partitioning for the column indices. Also, we have the following 2×2 block partitioning of the $m\times n$ random matrix Y:

$$Y = \begin{bmatrix} Y^{11} & Y^{12} \\ Y^{21} & Y^{22} \end{bmatrix}$$

where Y^{22} is $m_0 \times n_0$. The expected margin of Y is exactly $(\mathbf{r},\mathbf{c})$. Partition the typical table Z similarly. Denote $Z^{k\ell}$ for $1 \leq k,\ell \leq 2$ for the block entries of the corresponding partitioning of the Z. Note that $\mathbb{E}[Y] = Z$ so $\mathbb{E}[Y^{k\ell}] = Z^{k\ell}$ for all $k,\ell \in \{1,2\}$. Note that each $Y_{ij} \sim \mu_{\boldsymbol{\alpha}(i)+\boldsymbol{\beta}(j)}$ with $\phi(A_\delta) \leq \boldsymbol{\alpha}(i)+\boldsymbol{\beta}(j) \leq \phi(B_\delta)$ by δ-tameness of $(\mathbf{r},\mathbf{c})$. Hence $\{Y_{ij}\}'$s have a uniformly bounded sub-exponential norm, so by Bernstein's inequality for sums of independent sub-exponential random variables (see, e.g., [Ver18, Thm. 2.8.1]),

$$\begin{aligned}
\mathbb{P}\left(\left|Y^{11}_{\bullet\bullet} - Z^{11}_{\bullet\bullet}\right| \leq c\sqrt{mn\log mn}\right) &\geq 1 - 2\exp\left(-2a\log mn\right), \\
\mathbb{P}\left(\left|Y^{11}_{i\bullet} - Z^{11}_{i\bullet}\right| \leq c\sqrt{n\log n}\right) &\geq 1 - 2\exp\left(-2a\log n\right) \quad \text{for } 1 \leq i \leq m', \\
\mathbb{P}\left(\left|Y^{11}_{\bullet j} - Z^{11}_{\bullet j}\right| \leq c\sqrt{m\log m}\right) &\geq 1 - 2\exp\left(-2a\log m\right) \quad \text{for } 1 \leq j \leq n'
\end{aligned}$$

for some constant $c = c(\mu,\delta) > 0$. Let $\mathscr{D}^{11}$ denote the intersection of all events that appear on the left-hand side of the above inequalities. By a union bound,

$$\mathbb{P}(\mathscr{D}^{11}) \geq 1 - 6(m\vee n)(m\wedge n)^{-2a} \geq 1 - 6(m\wedge n)^{-a}.$$

By the hypothesis $m,n \geq 1$ are sufficiently large so that the last expression above is at least $1/2$. Then $\mathbb{P}(\mathscr{D}^{11}) \geq 1/2$. Hence by Lemma 5.7, it suffices to show

$$\mathbb{P}(\mathscr{E}(I) \,|\, \mathscr{D}^{11}) \geq \exp(-C(m\sqrt{n}\log n + n\sqrt{m}\log m)\log mn), \tag{5.14}$$

where $I \subseteq \mathbb{R}$ is a Borel set such that $c_I = \inf_{x\in I} p(x) > 0$.

We first argue for the discrete case when ζ is the counting measure on $\mathbb{Z}$ and μ has probability density p w.r.t. ζ. We will take $I = [0, \lceil B_{\delta/2} \rceil] \cap \mathbb{Z}$. By the hypothesis, $\{p > 0\} = [A,B] \cap \mathbb{Z}$. By translating μ, without loss of generality we assume $\lfloor A_{\delta/2} \rfloor = 0$. Then $p(0) > 0$. We also have

$\{1, \lceil B_{\delta/2} \rceil\} \subset \{p > 0\}$ since p cannot be concentrated at a single value (otherwise ν is not σ-finite). Write $\mathbf{r} = (\mathbf{r}^1, \mathbf{r}^2)$ and $\mathbf{c} = (\mathbf{c}^1, \mathbf{c}^2)$ where $\mathbf{r}^2 \in \mathbb{R}^{m_0}$ and $\mathbf{c}^2 \in \mathbb{R}^{n_0}$. Define events

$$\begin{aligned}
\mathscr{D}^{22} &:= \left\{ Y^{22}_{\bullet\bullet} = Z^{22}_{\bullet\bullet} + Y^{11}_{\bullet\bullet} - Z^{11}_{\bullet\bullet},\, Y^{22} \in I^{m_0 \times n_0} \right\}, \\
\mathscr{D}^{12} &:= \left\{ \|\pi(Y^{12}) - (\mathbf{r}^1 - r(Y^{11}), \mathbf{c}^2 - c(Y^{22}))\|_\infty = 0,\, Y^{12} \in I^{m' \times n_0} \right\}, \\
\mathscr{D}^{21} &:= \left\{ \|\pi(Y^{21}) - (\mathbf{r}^2 - r(Y^{22}), \mathbf{c}^1 - c(Y^{11}))\|_\infty = 0,\, Y^{21} \in I^{m_0 \times n'} \right\},
\end{aligned}$$

where π is the map that sends a matrix $\mathbf{x}$ to its margin $(r(\mathbf{x}), c(\mathbf{x}))$. We claim

$$\begin{aligned}
\mathbb{P}(\mathscr{D}^{22} \,|\, \mathscr{D}^{11}) &\geq \exp(-C m_0 n_0), \\
\mathbb{P}(\mathscr{D}^{12} \,|\, \mathscr{D}^{22} \cap \mathscr{D}^{11}) &\geq \exp(-C m n_0), \\
\mathbb{P}(\mathscr{D}^{21} \,|\, \mathscr{D}^{22} \cap \mathscr{D}^{11}) &\geq \exp(-C m_0 n).
\end{aligned} \tag{5.15}$$

Since $\mathscr{D}^{22} \cap \mathscr{D}^{12} \cap \mathscr{D}^{21}$ implies $\mathscr{E}(I)$ and since $\mathscr{D}^{12}, \mathscr{D}^{21}$ are conditionally independent on $\mathscr{D}^{11} \cap \mathscr{D}^{22}$, this would be enough to conclude (5.14).

To justifty the first inequality in (5.15), recall that since $(\mathbf{r}, \mathbf{c})$ is δ-tame, $Z^{22}_{\bullet\bullet} \in [A_\delta m_0 n_0, B_\delta m_0 n_0]$. Since $|Y^{11}_{\bullet\bullet} - Z^{11}_{\bullet\bullet}| \leq c\sqrt{mn \log mn} \ll m_0 n_0$ on $\mathscr{D}^{11}$,

$$A_{\delta/2} m_0 n_0 \leq Z^{22}_{\bullet\bullet} + Y^{11}_{\bullet\bullet} - Z^{11}_{\bullet\bullet} < B_{\delta/2} m_0 n_0 \quad \text{on } \mathscr{D}^{11}$$

for m, n large enough depending only on δ. The term in the middle is an integer since $Z^{22}_{\bullet\bullet} - Z^{11}_{\bullet\bullet} = \sum_{n' < j \leq n} \mathbf{c}(j) - \sum_{1 \leq i \leq m'} \mathbf{r}(i)$. Note that since μ assigns a positive probability for all values in the support, so does its exponential tilt with finite tilting parameter. Hence, there exists a matrix $\mathbf{y}^{22} \in I^{m_0 \times n_0}$ such that $\mathbf{y}^{22}_{\bullet\bullet} = Z^{22}_{\bullet\bullet} + Y^{11}_{\bullet\bullet} - Z^{11}_{\bullet\bullet}$ and $\mathbb{P}(Y^{22} = \mathbf{y}^{22}) > 0$. Since the $m_0 n_0$ entries of Y^{22} are independent, it follows that

$$\mathbb{P}(\mathscr{D}^{22} \,|\, \mathscr{D}^{11}) \geq \mathbb{P}(Y^{22} = \mathbf{y}^{22} \,|\, \mathscr{D}^{11}) = \mathbb{P}(Y^{22} = \mathbf{y}^{22}) \geq \exp(-C m_0 n_0).$$

Next, consider the second inequality in (5.15). The event $\mathscr{D}^{12}$ requires the row and column sums of Y^{12} to take the prescribed values so that the first m' row sums and the last n_0 columns sums of Y match the corresponding values in the target margin $(\mathbf{r}, \mathbf{c})$. Denote the required row and column sums for Y^{12} as $\mathbf{r}^{12} := \mathbf{r}^1 - r(Y^{11})$ and $\mathbf{c}^{12} := \mathbf{c}^2 - c(Y^{22})$, respectively. On $\mathscr{D}^{22}$, they have the same total sum (hence $(\mathbf{r}^{12}, \mathbf{c}^{12})$ is a margin for Y^{12}) since

$$\begin{aligned}
\sum_i \mathbf{r}^{12}(i) &= \sum_{1 \leq i \leq m'} \mathbf{r}(i) - Y^{11}_{\bullet\bullet} = Z^{11}_{\bullet\bullet} + Z^{12}_{\bullet\bullet} - Y^{11}_{\bullet\bullet}, \\
\sum_j \mathbf{c}^{12}(j) &= \sum_{n' < j \leq n} \mathbf{c}(j) - Y^{22}_{\bullet\bullet} = Z^{12}_{\bullet\bullet} + Z^{22}_{\bullet\bullet} - Y^{22}_{\bullet\bullet}
\end{aligned}$$

and the last expressions in each line above coincide on $\mathscr{D}^{22}$. Proceeding as before, it remains to show that there is at least one realization of Y^{12} in $I^{m' \times n_0}$ with margin $(\mathbf{r}^{12}, \mathbf{c}^{12})$. According to Lemma 5.4 **(i)**, we only need to verify that

$$\begin{aligned}
A_{\delta/2} n_0 < \mathbf{r}(i) - Y^{11}_{i\bullet} < B_{\delta/2} n_0 \quad &\text{for all } 1 \leq i \leq m' \quad \text{and} \\
A_{\delta/2} m_0 < \mathbf{c}(j) - Y^{22}_{\bullet j} < B_{\delta/2} m_0 \quad &\text{for all } n' < j \leq n.
\end{aligned}$$

Writing $\mathbf{r}(i) = Z^{11}_{i\bullet} + Z^{12}_{i\bullet}$ for $1 \leq i \leq m'$ and $\mathbf{c}(j) = Z^{12}_{\bullet j} + Z^{22}_{\bullet j}$ for $n' < j \leq n$, and noting that $Z \in [A_\delta, B_\delta]^{m \times n}$, one can see that the above inequalities hold almost surely for all sufficiently large m, n (depending only on c, δ) on the event $\mathscr{D}^{11}$. A symmetric argument applies for Y^{21}. This shows (5.15), completing the proof for the discrete case.

Lastly, we argue for the continuous case with $I = [A_{\delta/4}, B_{\delta/4}]$. By a discretization and perturbation argument, we will show that (5.15) still holds with some negligible error. Assume $\zeta = \text{Leb}(\mathbb{R})$

and $\mathrm{supp}(p) = [A,B] \cap \mathbb{R}$. Consider the rescaled integer grid $\varepsilon\mathbb{Z}$ in $\mathbb{R}$ for $\varepsilon = O(\delta/m^2n^2)$. For each $i,j = 1,2$, let $[Y^{ij}]$ denote the random matrix whose entries are the nearest point in $\varepsilon\mathbb{Z}$ to the corresponding entries in Y^{ij}. Since the support of p is an interval, the support of each entry of $[Y^{ij}]$ is a set of consecutive points in $\varepsilon\mathbb{Z}$. By the same argument for Y^{22} in the discrete case, we have

$$\mathbb{P}\Big(\big|[Y^{22}]_{\bullet\bullet} - \big(Z^{22} + Y^{11}_{\bullet\bullet} - Z^{11}_{\bullet\bullet}\big)\big| \le \varepsilon,\, [Y^{22}] \in J_\delta \cap \varepsilon\mathbb{Z} \,\Big|\, \mathscr{D}^{11}\Big) \ge \exp(-Cm_0n_0), \tag{5.16}$$

where $J_\delta = [\varepsilon\lfloor \varepsilon^{-1}A_{\delta/2}\rfloor, \varepsilon\lceil \varepsilon^{-1}B_{\delta/2}\rceil]$. Note that for each matrix $\mathbf{y}^{22} \in I^{m_0\times n_0}$, $[Y^{22}] = \mathbf{y}^{22}$ with probability at least $\Omega(\varepsilon^{m_0n_0}) \ge \exp(-Cm_0n_0\log mn)$. Hence denoting $J_\delta \pm \varepsilon = \{x \pm \varepsilon : x \in J_\delta\}$,

$$\mathbb{P}\Big(\big|Y^{22}_{\bullet\bullet} - \big(Z^{22} + Y^{11}_{\bullet\bullet} - Z^{11}_{\bullet\bullet}\big)\big| \le (m_0n_0+1)\varepsilon,\, Y^{22} \in (J_\delta \pm \varepsilon)^{m_0\times n_0} \,\Big|\, \mathscr{D}^{11}\Big) \ge \exp(-Cm_0n_0\log mn).$$

Let $\mathscr{D}^{22}_\varepsilon$ denote the event in (5.16) and define

$$\mathscr{D}^{12}_\varepsilon := \Big\{\|\pi(Y^{12}) - (\mathbf{r}^1 - r(Y^{11}), \mathbf{c}^2 - c(Y^{22}))\|_\infty \le (m+n)\varepsilon,\, Y^{12} \in (J_\delta \pm \varepsilon)^{m'\times n_0}\Big\},$$
$$\mathscr{D}^{21}_\varepsilon := \Big\{\|\pi(Y^{21}) - (\mathbf{r}^2 - r(Y^{22}), \mathbf{c}^1 - c(Y^{11}))\|_\infty \le (m+n)\varepsilon,\, Y^{21} \in (J_\delta \pm \varepsilon)^{m_0\times n'}\Big\}.$$

Similarly, we can show

$$\mathbb{P}(\mathscr{D}^{12}_\varepsilon \,|\, \mathscr{D}^{11} \cap \mathscr{D}^{22}_\varepsilon) \ge \exp(-Cmn_0\log mn),$$
$$\mathbb{P}(\mathscr{D}^{21}_\varepsilon \,|\, \mathscr{D}^{11} \cap \mathscr{D}^{22}_\varepsilon) \ge \exp(-Cm_0n\log mn).$$

Combining the above, we deduce

$$\mathbb{P}\Big(\Gamma_{\mathbf{r},\mathbf{c}}(\overline{Y})_{ij} \in J_\delta \pm (m+n+1)\varepsilon \;\forall (i,j) \text{ s.t. } i = m \text{ or } j = n \,\Big|\, \mathscr{D}^{11}\Big) \ge \exp(-C(mn_0 + m_0n)\log mn),$$

Hence for m,n large enough dependong only on δ so that $J_\delta \pm (m+n+1)\varepsilon \subseteq [A_{\delta/4}, B_{\delta/4}] = I$, we verify (5.14). □

Remark 5.8 (Combinatorial applications)**.** Our probabilistic argument above gives some interesting combinatorial results. In the proof of Theorem 1.14 we have bounded above the transference cost (1.16) with $q_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c})$ given by (5.8). Then using (5.10), we obtain

$$\int \prod_{i,j} p(\Gamma_{\mathbf{r},\mathbf{c}}(\overline{\mathbf{x}})_{ij})\, \zeta^{\otimes(m-1)\times(n-1)}(d\overline{\mathbf{x}})$$
$$\ge \frac{\exp(-g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha},\boldsymbol{\beta}))}{\sup_{\mathbf{y}} q_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c})} \exp(-C(m\sqrt{n}\log n + n\sqrt{m}\log m)\log mn).$$

The integral on the left-hand side above is a p-weighted volume of the transportation polytope $\mathscr{T}(\mathbf{r},\mathbf{c})$ in units of basic cells in $\mathbb{R}^{(m-1)\times(n-1)}$. When $p(x) = \mathbf{1}(x \ge 0)$ so that μ is the Lebesgue measure on $\mathbb{R}_{\ge 0}$, it gives a lower bound on the volume of the nonnegative transportation polytope (in this case $\sup_{\mathbf{y}} q_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c}) = 1$), which may be compared to the result of Barvinok [BH12] and Barvinok and Rudelson [BR24]. Our lower bound is weaker than their results but ours applies for more general non-constant densities p.

In the discrete case, the integral above becomes the weighted count $\sum_{\mathbf{x}\in\mathscr{T}(\mathbf{r},\mathbf{c})\cap\mathbb{Z}^{m\times n}} \prod_{ij} p(\mathbf{x}_{ij})$ of the inter points in the transportation polytope $\mathscr{T}(\mathbf{r},\mathbf{c})$ and $\sup_{\mathbf{y}} q_{\overline{\mathbf{y}}}(\mathbf{r},\mathbf{c}) = 1$. Setting $p(x) = \mathbf{1}(x \ge 0)$, it gives a lower bound on the number of contingency tables with margin $(\mathbf{r},\mathbf{c})$. This gives a weaker lower bound than the known result (5.11) of Brändén, Leake, and Pak [BLP23] but ours also applies for non-constant densities p. For the discrete case, we essentially lower bounded $\mathbb{P}\big(Y \in \mathscr{T}(\mathbf{r},\mathbf{c})\big)$, which equals $\exp\big(g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha},\boldsymbol{\beta})\big)\,\mu^{\otimes(m\times n)}(\mathscr{T}(\mathbf{r},\mathbf{c}))$ noting that $g^{\mathbf{r},\mathbf{c}}(\boldsymbol{\alpha},\boldsymbol{\beta})$ is the log-likelihood of Y taking any matrix in $\mathscr{T}(\mathbf{r},\mathbf{c})$ w.r.t. $\mu^{\otimes(m\times n)}$. This is why our upper bound on the transference cost implies

a lower bound on the corresponding weighted volume of the polytope. It would be interesting to know if the above results can be obtained by a purely combinatorial argument.

6. Proof of scaling limit of the typical tables and MLEs

Our goal in this section is to prove Theorem 1.19. This will take several steps.

6.1. Lipschitz continuity of typical tables and standard MLEs.

In this section, we establish a stability result (Theorem 6.1 below) that bounds the perturbation on the typical tables in terms of the perturbation of the margin. This result is central to our limit theory and it will be extended to the continuum setting (Theorem 6.7), which will be used critically in proving Theorem 1.19.

Theorem 6.1 (Lipschitz continuity of typical table w.r.t. margin)**.** *For each $m \times n$ δ-tame margins $(\mathbf{r},\mathbf{c})$ and $(\mathbf{r}',\mathbf{c}')$, denoting $C_\delta = \max\{|\phi(A_\delta)|, |\phi(B_\delta)|\}$, we have*

$$\|Z^{\mathbf{r},\mathbf{c}} - Z^{\mathbf{r}',\mathbf{c}'}\|_F^2 \le 4C_\delta \left(\sup_{\phi(A_\delta)\le w\le \phi(B_\delta)} \psi''(w) \right) \|(\mathbf{r},\mathbf{c}) - (\mathbf{r}',\mathbf{c}')\|_1. \tag{6.1}$$

Our key insight behind the proof of the above result is to *use the dual variable (MLE) to bridge the margin and the corresponding typical table.* Namely, while it is difficult to directly construct a map from a margin to the corresponding typical table, being the solution of a constrained optimization problem (1.7), it is easy to write down the margin and the typical table corresponding to a given MLE, by using Theorem 1.7. A schematic for this approach is given in the diagram below:

$$\begin{array}{ccccc} \textit{margin} & & \textit{MLE} & & \textit{typical table} \\ (\mathbf{r}(1),\dots,\mathbf{r}(m),\mathbf{c}(1),\dots,\mathbf{c}(n)) & \xleftarrow{\ \Lambda\ } & (\boldsymbol{\alpha}(1),\dots,\boldsymbol{\alpha}(m),\boldsymbol{\beta}(1),\dots,\boldsymbol{\beta}(n)) & \xrightarrow{\ \Phi\ } & \mathtt{VEC}(Z) \end{array}$$

Here for each matrix $X \in \mathbb{R}^{m\times n}$, let $\mathtt{VEC}(X) \in \mathbb{R}^{mn}$ denotes is vectorization, which is obtained by stacking the jth column underneath its $(j-1)$st column for $j = 2,\dots,n$.

Denote $q := m+n$ and $p := mn$. Define a map $\Lambda : \mathbb{R}^q \to \mathbb{R}^q$, $(\boldsymbol{\alpha},\boldsymbol{\beta}) \mapsto (\mathbf{r},\mathbf{c})$ where

$$\begin{cases} \mathbf{r}(i) = \sum_{j=1}^n \psi'(\boldsymbol{\alpha}(i)+\boldsymbol{\beta}(j)) & \text{for all } i = 1,\dots,m, \\ \mathbf{c}(j) = \sum_{i=1}^m \psi'(\boldsymbol{\alpha}(i)+\boldsymbol{\beta}(j)) & \text{for all } j = 1,\dots,n. \end{cases}$$

Also define a map $\Phi : \mathbb{R}^{m+n} \to \mathbb{R}^{m\times n}$, $(\boldsymbol{\alpha},\boldsymbol{\beta}) \mapsto Z = (z_{ij})$ by $Z = \psi'(\boldsymbol{\alpha}\oplus\boldsymbol{\beta})$. Both maps are well-defined and are differentiable. It is important to understand the structure of the Jacobian matrices of these maps:

$$J_\Lambda(\boldsymbol{\alpha},\boldsymbol{\beta}) := \left[\frac{\partial(\mathbf{r}(1),\dots,\mathbf{r}(m),\mathbf{c}(1),\dots,\mathbf{c}(n))}{\partial(\boldsymbol{\alpha}(1),\dots,\boldsymbol{\alpha}(m),\boldsymbol{\beta}(1),\dots,\boldsymbol{\beta}(n))} \right]_{q\times q},$$

$$J_\Phi(\boldsymbol{\alpha},\boldsymbol{\beta}) := \left[\frac{\partial\, \mathtt{VEC}(\Phi(\boldsymbol{\alpha},\boldsymbol{\beta}))^\top}{\partial(\boldsymbol{\alpha}(1),\dots,\boldsymbol{\alpha}(m),\boldsymbol{\beta}(1),\dots,\boldsymbol{\beta}(n))} \right]_{p\times q}.$$

These matrices turn out to admit precise factorization into some important matrices, which we introduce below. For each dual variables $(\boldsymbol{\alpha},\boldsymbol{\beta}),(\boldsymbol{\alpha}',\boldsymbol{\beta}') \in \mathbb{R}^m \times \mathbb{R}^n$, define matrices

$$P^{\boldsymbol{\alpha},\boldsymbol{\beta}} := \operatorname{diag}\big(\psi''(\boldsymbol{\alpha}(1)+\boldsymbol{\beta}(1)),\dots,\psi''(\boldsymbol{\alpha}(1)+\boldsymbol{\beta}(n)),\dots,\psi''(\boldsymbol{\alpha}(m)+\boldsymbol{\beta}(n))\big) \in \mathbb{R}^{p\times p}$$

$$P^{\boldsymbol{\alpha},\boldsymbol{\beta};\boldsymbol{\alpha}',\boldsymbol{\beta}'} := \int_0^1 P^{(1-t)(\boldsymbol{\alpha},\boldsymbol{\beta})+t(\boldsymbol{\alpha}',\boldsymbol{\beta}')}\, dt.$$

Let $h_i(X)$ and $h^j(X)$ denote the ith row sum and jth columns sum of X, respectively. Then $\nabla h_i(X)$ is the $m \times n$ matrix where the ith row is filled with 1s and 0s elsewhere. Similarly, $\nabla h^j(X)$ is the $m \times n$ matrix with 1s on the jth column and 0s elsewhere. Then define a matrix

$$Q := \begin{bmatrix} \texttt{VEC}(\nabla h_1) & \dots & \texttt{VEC}(\nabla h_m) & \texttt{VEC}(\nabla h^1) & \dots \texttt{VEC}(\nabla h^n) \end{bmatrix} \in \mathbb{R}^{p \times q}.$$

Proposition 6.2. $J_{\Lambda_0}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = Q^\top P^{\boldsymbol{\alpha},\boldsymbol{\beta}} Q, \qquad J_{\Phi_0}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = P^{\boldsymbol{\alpha},\boldsymbol{\beta}} Q.$

Proof. Fix a matrix $E = (E_{ij})_{i,j} \in \mathbb{R}^{m \times n}$. We claim that

$$Q^\top \operatorname{diag}(\texttt{VEC}(E)) Q = \left[\begin{array}{ccc|ccc} E_{1\bullet} & & & E_{11} & \dots & E_{1n} \\ & \ddots & & & \ddots & \\ & & E_{m\bullet} & E_{m1} & \dots & E_{mn} \\ \hline E_{11} & \dots & E_{m1} & E_{\bullet 1} & & \\ & \ddots & & & \ddots & \\ E_{1n} & \dots & E_{mn} & & & E_{\bullet n} \end{array}\right]. \tag{6.2}$$

To see this, note that the columns of Q are the indicators of entries in the corresponding row/column sum for a $m \times n$ matrix. Hence, the $(1,1)$ entry in the left-hand side of (6.2) is the sum of all entries in E that appear in the first row sum, which is $E_{1\bullet}$; Its $(1,2)$ entry is zero since there is no entry of E shared in the first and the second row sums; Its $(1,m)$ entry is the sum of all entries in E appearing in the first row sum and the first column sum, which is E_{11}.

Now one can directly compute the Jacobian $J_{\Lambda;\boldsymbol{\alpha},\boldsymbol{\beta}}$ and it has the 2×2 block matrix form in the right-hand side above with $E_{ij} = \psi''(\boldsymbol{\alpha}(i) + \boldsymbol{\beta}(j))$. Hence the first formula in the assertion follows from the identity (6.2). The second formula can be verified by a straightforward computation. □

Suppose we have two δ-tame margins $(\mathbf{r}, \mathbf{c})$ and $(\mathbf{r}', \mathbf{c}')$. Can we find a canonical path within $\mathscr{M}^\delta$ that interpolates between these two δ-tame margins? A natural candidate would be the linear interpolation between the two margins. However, it is not necessarily true that a convex combination of two δ-tame margins is again δ-tame. It turns out that the right way is to linearly interpolate between the corresponding MLEs in the dual space and map it back to the space of tables and margins. More precisely, let $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $(\boldsymbol{\alpha}', \boldsymbol{\beta}')$ be any MLEs for margins $(\mathbf{r}, \mathbf{c})$ and $(\mathbf{r}', \mathbf{c}')$, respectively, which exist by Theorem 1.7. For each $\lambda \in [0,1]$, let $(\boldsymbol{\alpha}_\lambda, \boldsymbol{\beta}_\lambda) := (1-\lambda)(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \lambda(\boldsymbol{\alpha}', \boldsymbol{\beta}')$ be the convex combination of the MLEs. Define a margin

$$\gamma(\lambda) := \text{the margin satisfied by } Z_\lambda, \tag{6.3}$$

where the $m \times n$ table $Z_\lambda = ((z_\lambda)_{ij})$ is defined by $Z_\lambda = \psi'(\boldsymbol{\alpha}_\lambda \oplus \boldsymbol{\beta}_\lambda)$. Notice that for any $\lambda \in [0,1]$, the interpolated dual variable $(\boldsymbol{\alpha}_\lambda, \boldsymbol{\beta}_\lambda)$ is an MLE for the margin $\gamma(\lambda)$ due to Lemma 4.1.

Proposition 6.3 (C^1-connectedness of $\mathscr{M}^\delta$)**.** *Keep the same setting as before. Then γ in (6.3) defines a C^1-path γ from $[0,1]$ into the set of all δ-tame $m \times n$ margins such that $\gamma(0) = (\mathbf{r}, \mathbf{c})$, $\gamma(1) = (\mathbf{r}', \mathbf{c}')$.*

Proof. The argument is straightforward and we omit the details. □

We can write the change in margin and the typical table as we move along the secant line between the two MLEs by a line integral. Using Proposition 6.2 along the way, we derive the following key lemma for the proof of Theorem 6.1.

Lemma 6.4. *Keep the same setting as before. Then denoting $v := \texttt{VEC}(\boldsymbol{\alpha}' - \boldsymbol{\alpha}, \boldsymbol{\beta}' - \boldsymbol{\beta}) \in \mathbb{R}^q$, we have*

$$(\mathbf{r}', \mathbf{c}') - (\mathbf{r}, \mathbf{c}) = Q^\top P^{\boldsymbol{\alpha},\boldsymbol{\beta};\boldsymbol{\alpha}',\boldsymbol{\beta}'} Q v \quad \text{and} \quad \texttt{VEC}(Z') - \texttt{VEC}(Z) = P^{\boldsymbol{\alpha},\boldsymbol{\beta};\boldsymbol{\alpha}',\boldsymbol{\beta}'} Q v.$$

Proof. We first use the chain rule to write

$$\frac{d}{dt}\gamma(t) = J_\Lambda(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t)\, v.$$

By Proposition 6.2, we can write

$$\gamma(1) - \gamma(0) = \int_0^1 J_\Lambda(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t)\, v\, dt = Q^\top \left[\int_0^1 P^{\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t}\, dt\right] Qv = Q^\top P^{\boldsymbol{\alpha}, \boldsymbol{\beta}; \boldsymbol{\alpha}', \boldsymbol{\beta}'} Qv,$$

where the integral in the rightmost expression above is done entry-wise. Similarly, using Proposition 6.2, we get

$$\Phi(\boldsymbol{\alpha}', \boldsymbol{\beta}') - \Phi(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \int_0^1 J_\Phi(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t)\, v\, dt = \left[\int_0^1 P^{\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t}\, dt\right] Qv.$$

This shows the assertion. □

Now we are ready to show Theorem 6.1.

***Proof of Theorem** 6.1.* Let $(\boldsymbol{\alpha}, \boldsymbol{\beta}) := (\boldsymbol{\alpha}^{\mathbf{r},\mathbf{c}}, \boldsymbol{\beta}^{\mathbf{r},\mathbf{c}})$ and $(\boldsymbol{\alpha}', \boldsymbol{\beta}') := (\boldsymbol{\alpha}^{\mathbf{r}',\mathbf{c}'}, \boldsymbol{\beta}^{\mathbf{r}',\mathbf{c}'})$ denote the standard MLEs for margins $(\mathbf{r}, \mathbf{c})$ and $(\mathbf{r}', \mathbf{c}')$, respectively. Write $\overline{P} := P^{\boldsymbol{\alpha}, \boldsymbol{\beta}; \boldsymbol{\alpha}', \boldsymbol{\beta}'}$ and $R := \overline{P}^{1/2} Q \in \mathbb{R}^{p \times q}$. Then by Lemma 6.4, we have the following relations

$$\mathtt{VEC}(Z') - \mathtt{VEC}(Z) = \overline{P}^{1/2} Rv, \quad \gamma(1) - \gamma(0) = R^\top Rv,$$

where v is as in Lemma 6.4. On the one hand, by Proposition 6.3, $A_\delta \le \psi'(\boldsymbol{\alpha}_t(i) + \boldsymbol{\beta}_t(j)) \le B_\delta$ for all $t \in [0,1]$ and i, j. Hence $\|\overline{P}\|_2 \le \sup_{\phi(A_\delta) \le w \le \phi(B_\delta)} \psi''(w)$, so we get

$$\|Z - Z'\|_F = \|\mathtt{VEC}(Z') - \mathtt{VEC}(Z)\|_2 \le \|\overline{P}^{1/2}\|_2 \|Rv\|_2 \le \left(\sup_{\phi(A_\delta) \le w \le \phi(B_\delta)} \psi''(w)\right)^{1/2} \|Rv\|_2. \tag{6.4}$$

On the other hand, by Hölder's inequality,

$$\|Rv\|_2^2 = v^\top R^\top Rv \le \|v^\top\|_\infty \cdot \|R^\top Rv\|_1 \le 4C\, \|(\mathbf{r}, \mathbf{c}) - (\mathbf{r}', \mathbf{c}')\|_1, \tag{6.5}$$

where $C = \max\{|\phi(A_\delta)|, |\phi(B_\delta)|\}$ as in (4.6). Then (6.1) follows from (6.4) and (6.5). □

6.2. **Stability of typical kernels and continuous MLEs.** In this section, we prove Theorem 1.19.

Let μ be a measure as in the introduction and let A, B be endpoints of $\mathrm{supp}(\mu)$ as in (1.1). Let $\mathscr{W} := L^1([0,1]^2)$ denote the set of all measurable functions from $[0,1]^2 \mapsto \mathbb{R}$ which are integrable. We will equip $\mathscr{W}$ with the cut-metric topology defined as follows. The *strong cut metric* on $\mathscr{W}$ is defined by $d_\square(U, W) = \|U - W\|_\square$ using the cut norm $\|\cdot\|_\square$ in (1.30). The *weak cut metric* $\delta_\square$ on $\mathscr{W}$ is defined as

$$\delta_\square(U, W) := \inf_{\xi, \eta : [0,1] \to [0,1]} \|U(p(\cdot), q(\cdot)) - W\|_\square,$$

where the infimum above is over all Lebesgue-measure-preserving maps ξ, η on the unit interval $[0,1]$. Note that the following gives an equivalent definition of the cut-norm in (1.30):

$$\|W\|_\square = \sup_{a,b : [0,1] \to [0,1]} \left| \int_{[0,1]^2} a(x) W(x,y) b(y)\, dx dy \right|, \tag{6.6}$$

where the supremum above is over all continuous functions a, b on the unit interval $[0,1]$. The two definitions (1.30) and (6.6) differ upto multiplicative universal constants ([Lov12, Lem. 8.10]).

Let $\mathscr{W}^{(A,B)}$ denote the set of all kernels W in $\mathscr{W}$ taking calues from (A,B). Similarly, define $\mathscr{W}^{[A,B]}$ to be the set of all kernels taking values from $[A,B]$. With $f(x) = D(\mu_{\phi(x)} \| \mu)$ being as in (1.7), define a function $G : \mathscr{W}^{(A,B)} \mapsto \mathbb{R}$ by

$$G(W) := \int_{[0,1]^2} f(W(x,y))\, dxdy. \tag{6.7}$$

Let $W(\cdot,\bullet) := \int_{[0,1]} W(\cdot,y)dy$ and $W(\bullet,\cdot) := \int_{[0,1]} W(x,\cdot)dx$ be univariate functions which denote the row and column marginals of W. Let $(\mathbf{r},\mathbf{c})$ be a continuum margin (defined above (1.19)). Define

$$\mathscr{W}_{\mathbf{r},\mathbf{c}} := \Big\{ W \in \mathscr{W} : W(\cdot,\bullet) = \mathbf{r}(\cdot), W(\bullet,\cdot) = \mathbf{c}(\cdot) \Big\}, \quad \mathscr{W}_{\mathbf{r},\mathbf{c}}^{(A,B)} := \mathscr{W}_{\mathbf{r},\mathbf{c}} \cap \mathscr{W}^{(A,B)}, \quad \mathscr{W}_{\mathbf{r},\mathbf{c}}^{[A,B]} := \mathscr{W}_{\mathbf{r},\mathbf{c}} \cap \mathscr{W}^{[A,B]}.$$

Here $\mathscr{W}_{\mathbf{r},\mathbf{c}}$ is the set of all kernels with continuum margin $(\mathbf{r},\mathbf{c})$.

Now for each continuum margin $(\mathbf{r},\mathbf{c})$, consider the minimization problem

$$\inf_{W \in \mathscr{W}_{\mathbf{r},\mathbf{c}}^{(A,B)}} G(W). \tag{6.8}$$

If $\mathscr{W}_{\mathbf{r},\mathbf{c}}$ is non-empty, then it is a convex subset of $\mathscr{W}$. Consequently, invoking strict convexity of G w.r.t. the cut-metric topology (see Proposition 6.10), there can exist at most one global minimizer of the above optimization problem. We denote it by $W^{\mathbf{r},\mathbf{c}}$, if it exists. In this case, we call $W^{\mathbf{r},\mathbf{c}}$ the *typical kernel* for the margin $(\mathbf{r},\mathbf{c})$. We say a continuum margin $(\mathbf{r},\mathbf{c})$ δ*-tame* if a typical kernel $W^{\mathbf{r},\mathbf{c}}$ exists and satisfies $A_\delta \le W \le B_\delta$.

Our proof of Theorem 1.19 will follow by combining the following three results. First, we characterize the typical kernels for δ-tame margins using continuous dual variables in the same way as we did in the discrete case in Theorem 1.7.

Lemma 6.5 (Characterization of typical kernels and tame margins)**.**

(i) *Fix a δ-tame continuum margin $(\mathbf{r},\mathbf{c})$ in $L^1[0,1]$. Let $C = C_\delta$ as before. Then there exists bounded measurable functions $\boldsymbol{\alpha} : [0,1] \to [-2C, 2C]$ and $\boldsymbol{\beta} : [0,1] \to [-C, C]$ such that*

$$W^{\mathbf{r},\mathbf{c}}(x,y) \overset{a.s.}{=} \psi'(\boldsymbol{\alpha}(x) + \boldsymbol{\beta}(y)). \tag{6.9}$$

(ii) *Conversely, suppose there exists bounded measurable functions $\boldsymbol{\alpha}, \boldsymbol{\beta} : [0,1] \to \mathbb{R}$ such that the function $W^*(x,y) = \psi'(\boldsymbol{\alpha}(x) + \boldsymbol{\beta}(y))$ satisfies the margin $(\mathbf{r},\mathbf{c}) \in L^1[0,1]$. Then W^* is the unique typical kernel for the margin $(\mathbf{r},\mathbf{c})$.*

We record an immediate consequence of Lemma 6.5 **(ii)** below.

Proposition 6.6. *Let $(\mathbf{r},\mathbf{c})$ be an $m \times n$ margin with typical table $Z^{\mathbf{r},\mathbf{c}} \in (A,B)^{m \times n}$. Then $W_{Z^{\mathbf{r},\mathbf{c}}}$ is the unique typical kernel for the continuum margin $(\bar{\mathbf{r}},\bar{\mathbf{c}})$ (see (1.19)).*

Second, we have the following continuous extension of Theorem 6.1, which establishes Lipschitz continuity of typical kernels w.r.t. tame margins. In addition, we also establish Lipschitz continuity of the dual variables w.r.t. tame margins.

Theorem 6.7 (Lipschitz continuity of typical kernels and dual variables)**.** *Let $(\mathbf{r},\mathbf{c}), (\mathbf{r}',\mathbf{c}')$ be δ-tame continuum margins on $[0,1]^2$. Furthermore, let $C_\delta := \max\{|\phi(A_\delta)|, |\phi(B_\delta)|\}$. Then*

$$\| W^{\mathbf{r},\mathbf{c}} - W^{\mathbf{r}',\mathbf{c}'} \|_2^2 \le 4C_\delta \left(\sup_{\phi(A_\delta) \le w \le \phi(B_\delta)} \psi''(w) \right) \| (\mathbf{r},\mathbf{c}) - (\mathbf{r}',\mathbf{c}') \|_1. \tag{6.10}$$

Furthermore, let $(\boldsymbol{\alpha},\boldsymbol{\beta})$ *and* $(\boldsymbol{\alpha}',\boldsymbol{\beta}')$ *denote the dual variables characterizing* $W^{\mathbf{r},\mathbf{c}}$ *and* $W^{\mathbf{r}',\mathbf{c}'}$ *via* (6.9)*, respectively. Without loss of generality, assume* $\int_0^1 \boldsymbol{\alpha}(x)\,dx = \int_0^1 \boldsymbol{\alpha}'(x)\,dx = 0$*. Then*

$$\|\boldsymbol{\alpha}-\boldsymbol{\alpha}'\|_2^2 + \|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_2^2 \le 4C_\delta \left(\frac{\sup_{\phi(A_\delta)\le w\le \phi(B_\delta)} \psi''(w)}{\inf_{\phi(A_\delta)\le w\le \phi(B_\delta)} \psi''(w)}\right) \|(\mathbf{r},\mathbf{c})-(\mathbf{r}',\mathbf{c}')\|_1. \tag{6.11}$$

Lastly, we show that the L^1-limit of discrete δ-tame margins is a δ-tame continuum margin.

Proposition 6.8. *Let* $(\mathbf{r}_k,\mathbf{c}_k)_{k\ge 1}$ *be a sequence of* $(m_k\times n_k)$ δ*-tame discrete margins converging to some continuum margin* $(\mathbf{r},\mathbf{c})$ *in* L^1*. Then* $(\mathbf{r},\mathbf{c})$ *is* δ*-tame and*

$$W^{(\bar{\mathbf{r}}_k,\bar{\mathbf{c}}_k)} \to W^{\mathbf{r},\mathbf{c}} \quad \text{in } L^2 \text{ as } k\to\infty.$$

Assuming these three results, we deduce Theorem 1.19 below.

***Proof of Theorem** 1.19.* By the hypothesis and Proposition 6.8, the limiting continuum margin $(\mathbf{r},\mathbf{c})$ is δ-tame. Hence the existence of continuum dual variable $(\boldsymbol{\alpha},\boldsymbol{\beta})$ and the claimed characterization of the typical kernel $W^{\mathbf{r},\mathbf{c}}$ in part **(i)** follows from Lemma 6.5. Part **(ii)** follows directly from Theorem 6.7 and Proposition 6.6. □

The rest of this section is devoted to showing the three results stated above. We first prove Lemma 6.5.

***Proof of Lemma** 6.5.* We first show **(i)**. By the hypothesis of part **(i)**, the typical table $W^{\mathbf{r},\mathbf{c}}$ exists and it is δ-tame. Let $U(\cdot,\cdot)$ be a measurable function from $[0,1]^2$ to $[-1,1]$ which satisfies $U(\cdot,\bullet) = U(\bullet,\cdot) \overset{a.s.}{=} 0$. Then the function $\widetilde{W}^{(\lambda)} := W^{\mathbf{r},\mathbf{c}} + \lambda U \in \mathscr{W}_{\mathbf{r},\mathbf{c}}$ for all $\lambda\in(-\delta,\delta)$. Since $W^{\mathbf{r},\mathbf{c}}$ is the typical table, the function $\lambda \mapsto G(\widetilde{W}^{(\lambda)})$ is uniquely maximized on $(-\delta,\delta)$ at $\lambda=0$, which gives

$$0 = \frac{\partial}{\partial\lambda} G(\widetilde{W}^{(\lambda)})\Big|_{\lambda=0} = \int_{[0,1]^2} \phi\big(W^{\mathbf{r},\mathbf{c}}(x,y)\big)\,U(x,y)\,dx\,dy.$$

Since this holds for all bounded measurable U which integrates to 0 along both marginals, it follows that there exists functions $\boldsymbol{\alpha},\boldsymbol{\beta}:[0,1]\mapsto\mathbb{R}$ such that

$$\phi\big(W^{\mathbf{r},\mathbf{c}}(x,y)\big) \overset{a.s.}{=} \boldsymbol{\alpha}(x)+\boldsymbol{\beta}(y), \text{ which implies } W^{\mathbf{r},\mathbf{c}}(x,y) \overset{a.s.}{=} \psi'(\boldsymbol{\alpha}(x)+\boldsymbol{\beta}(y)).$$

Finally, the fact that $W^{\mathbf{r},\mathbf{c}}$ is δ-tame implies $\phi(A_\delta) \overset{a.s.}{\le} \boldsymbol{\alpha}(x)+\boldsymbol{\beta}(y) \overset{a.s.}{\le} \phi(B_\delta)$. Since changing $(\boldsymbol{\alpha},\boldsymbol{\beta})$ to $(\boldsymbol{\alpha}+\eta,\boldsymbol{\beta}-\eta)$ has no impact on $W^{\mathbf{r},\mathbf{c}}$ for any $\eta\in\mathbb{R}$, we can assume without loss of generality that $\int_{[0,1]}\boldsymbol{\alpha}(x)dx = 0$. This, along with the previous display implies

$$\phi(A_\delta) \overset{a.s.}{\le} \boldsymbol{\beta}(y) \overset{a.s.}{\le} \phi(B_\delta),$$

which in turn implies

$$\phi(A_\delta)-\phi(B_\delta) \overset{a.s}{\le} \boldsymbol{\alpha}(x) \overset{a.s}{\le} \phi(B_\delta)-\phi(A_\delta).$$

The desired conclusion of part (a) follows.

Next, we show **(ii)**. Let $\widetilde{W}\in\mathscr{W}_{\mathbf{r},\mathbf{c}}$ be arbitrary. It suffices to show that $G(W^*)\ge G(\widetilde{W})$. Since G is strictly convex, it suffices to show that the function $\lambda\mapsto G\big((1-\lambda)W^*+\lambda\widetilde{W}\big)$ on $[0,1]$ has a derivative which vanishes at $\lambda=0$. This is equivalent to checking

$$\int_{[0,1]^2} \phi\big(W^*(x,y)\big)\,U(x,y)\,dx\,dy = 0, \text{ where } U = W^*-\widetilde{W}.$$

But this follows on noting that $\phi\big(W^*(x,y)\big) = \boldsymbol{\alpha}(x)+\boldsymbol{\beta}(y)$, and $U(\cdot,\bullet) = U(\bullet,\cdot) \overset{a.s.}{=} 0$. □

Our next aim is to prove Theorem 6.7. We will need some preparation. We first recall the following standard fact about step function approximation.

Lemma 6.9 (Approximation by stepfunctions in L^1)**.** *Suppose $|A|, |B| < \infty$ and let $h : [A,B]^d \to \mathbb{R}$ be a measurable function for some integer $d \geq 1$. Then there exists a sequence of stepfunctions $(h_n)_{n\geq 1}$ such that $\|h_n - h\|_1 \to 0$ as $n \to \infty$.*

Typical construction of such stepfunctions in Lemma 6.9 is by block-averaging over diadic partitions. The L^1 convergence can be shown by applying Lévy's upward convergence theorem. We omit the details.

Next, we establish the basic properties of the function G in (6.7).

Proposition 6.10. *The function G in (6.7) is well-defined and strictly convex. Furthermore, for any $\delta > 0$, G restricted on $\mathscr{W}^{[A_\delta, B_\delta]}$ is lower semi-continuous with respect to the cut distance $\delta_\square$.*

Proof. By the assumption on μ, there exists some tilting parameter $\theta_0 \in \Theta$, for which μ_{θ_0} is a probability measure. Fix $\theta \in \Theta$. Noting that the KL divergence between two probability distributions is nonnegative,

$$D(\mu_\theta \| \mu) = D(\mu_\theta \| \mu_{\theta_0}) + \theta_0 \psi'(\theta) - \psi(\theta_0) \geq \theta_0 \psi'(\theta) - \psi(\theta_0).$$

It follows that for the function $f(x) = D(\mu_{\phi(x)} \| \mu)$ in (1.7), $f(x) \geq \theta_0 x - \psi(\theta_0)$. This yields that $G(W) \in (-\infty, \infty]$ for all $W \in \mathscr{W}^{(A,B)}$ since

$$G(W) = \int_{[0,1]^2} f(W(x,y))\, dx dy \geq \theta_0 \int_{[0,1]^2} W(x,y)\, dx dy - \psi(\theta_0) > -\infty.$$

For strict convexity, fix $\lambda \in (0,1)$. Since f is strictly convex on (A,B) (see (4.3)), we have $f(\lambda y + (1-\lambda)x) < \lambda f(y) + (1-\lambda) f(x)$ for $x, y \in (A,B)$ and $x \neq y$. It follows that $G(\lambda W' + (1-\lambda)W) < \lambda G(W') + (1-\lambda) G(W)$ for $W, W' \in \mathscr{W}^{(A,B)}$ with $W \neq W'$ almost surely.

Next, we consider G restricted on $\mathscr{W}^{[A_\delta, B_\delta]}$. To show lower semi-continuity, let W_k be a sequence of functions in $\mathscr{W}^{[A_\delta, B_\delta]}$ converging to $W \in \mathscr{W}^{[A_\delta, B_\delta]}$ in $\delta_\square$-metric. We wish to show that

$$\liminf_{k\to\infty} G(W_k) \geq G(W).$$

Noting that $G(W) = G(W(\xi(\cdot), \eta(\cdot)))$ for every measure-preserving transformations ξ, η on $[0,1]$, without loss of generality we can assume $\|W_k - W\|_\square \to 0$ as $k \to \infty$.

Define W^L to be the $L \times L$ block-ageraging of W for every $W \in \mathscr{W}$ and $L > 0$. By convexity of G and Jensen's inequality, $G(W_k) \geq G(W_k^L)$. For fixed L, $G(W_k^L) \to G(W^L)$ using the continuity of G on $L \times L$ stepfunctions. It follows that

$$\liminf_{k\to\infty} G(W_k) \geq \liminf_{L\to\infty} \liminf_{k\to\infty} G(W_k^L) = \liminf_{L\to\infty} G(W^L).$$

Hence it suffices to show that

$$\liminf_{L\to\infty} G(W^L) \geq G(W).$$

To this end, let U, V be independent uniform $[0,1]$ variables. Then $W^L(U,V)$ converges to $W(U,V)$ in probability. For every integrable function $W' : [0,1] \to \mathbb{R}$, we have $G(W') = \mathbb{E}[f(W'(U,V))]$. Then noting that f is continuous and the range of $f \circ W^L$ is in $[A_\delta, B_\delta]$ for all $L \geq 1$, we get

$$\lim_{L\to\infty} G(W^L) = \lim_{L\to\infty} \mathbb{E}[f(W^L(U,V))] = \mathbb{E}[f(W(U,V))] = G(W),$$

where the equality in the middle follows from the bounded convergence theorem noting that $f \circ W$ is bounded for W taking values from $[A_\delta, B_\delta]$. □

We now prove Lipschitz continuity of typical kernels and dual variables stated in Theorem 6.7. Our approach is to extend the discrete analog (Theorem 6.1) to the continuous case by discretizing the dual variables and passing to the limit. The key ingredient for these continuous extensions is Lemma 6.5, which we have established above.

Proof of Theorem 6.7. We will first show (6.10) by pushing the discrete result in Theorem 6.1 to the continuum limit by block averaging the MLEs. Fix an integer $L \geq 1$. Then by Lemma 6.5, there exists bounded and measurable functions $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\alpha}', \boldsymbol{\beta}' : [0,1] \to [-2C_\delta, 2C_\delta]$ such that

$$W^{\mathbf{r},\mathbf{c}}(x,y) = \psi'(\boldsymbol{\alpha}(x) + \boldsymbol{\beta}(y)) \quad \text{and} \quad W^{\mathbf{r}',\mathbf{c}'}(x,y) = \psi'(\boldsymbol{\alpha}'(x) + \boldsymbol{\beta}'(y)), \tag{6.12}$$

where $W^{\mathbf{r},\mathbf{c}}, W^{\mathbf{r}',\mathbf{c}'} \in \mathscr{W}^{[A_\delta, B_\delta]}$. In particular,

$$\mathbf{r}(x) = \int_0^1 \psi'(\boldsymbol{\alpha}(x) + \boldsymbol{\beta}(y))\, dy, \quad \mathbf{c}(y) = \int_0^1 \psi'(\boldsymbol{\alpha}(x) + \boldsymbol{\beta}(y))\, dx,$$

and similarly for $(\mathbf{r}', \mathbf{c}')$.

For each function $h : [0,1] \to \mathbb{R}$ and an integer $L \geq 1$, let h_L denote the block average of h over intervals $[(i-1)2^{-L}, i2^{-L})$ for $i = 1, \ldots, 2^L$. By Lemma 6.9, there exists $\varepsilon = \varepsilon(L) > 0$ such that

$$\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_L\|_1 + \|\boldsymbol{\beta} - \boldsymbol{\beta}_L\|_1 + \|\boldsymbol{\alpha}' - \boldsymbol{\alpha}'_L\|_1 + \|\boldsymbol{\beta}' - \boldsymbol{\beta}'_L\|_1 \leq \varepsilon(L) \to 0 \quad \text{as } L \to \infty.$$

Now define block margin $(\mathbf{r}_L, \mathbf{c}_L)$ by

$$\mathbf{r}_L(x) = \int_0^1 \psi'(\boldsymbol{\alpha}_L(x) + \boldsymbol{\beta}_L(y))\, dy, \quad \mathbf{c}_L(y) = \int_0^1 \psi'(\boldsymbol{\alpha}_L(x) + \boldsymbol{\beta}_L(y))\, dx.$$

Namely, the block margin $(\mathbf{r}_L, \mathbf{c}_L)$ is obtained by taking block-average of the dual variable $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Further, note that from (6.12),

$$\phi(A_\delta) \leq \boldsymbol{\alpha}_L(x) + \boldsymbol{\beta}_L(y) \leq \phi(B_\delta) \quad \text{and} \quad \phi(A_\delta) \leq \boldsymbol{\alpha}'_L(x) + \boldsymbol{\beta}'_L(y) \leq \phi(B_\delta).$$

Hence $(\mathbf{r}_L, \mathbf{c}_L)$ and $(\mathbf{r}'_L, \mathbf{c}'_L)$ are both δ-tame margins for all $L \geq 1$.

Denote $D_\delta := \sup_{\phi(A_\delta) \leq w \leq \phi(B_\delta)} \psi''(w) > 0$. Note that

$$\|\mathbf{r} - \mathbf{r}_L\|_1 + \|\mathbf{c} - \mathbf{c}_L\|_1 \leq D_\delta \int_{[0,1]^2} |\boldsymbol{\alpha}_L(x) - \boldsymbol{\alpha}(x)| + |\boldsymbol{\beta}_L(y) - \boldsymbol{\beta}(y)|\, dy\, dx \leq 2D_\delta \varepsilon(L). \tag{6.13}$$

By a similar argument,

$$\|\mathbf{r}' - \mathbf{r}'_L\|_1 + \|\mathbf{c}' - \mathbf{c}'_L\|_1 \leq 2D_\delta \varepsilon(L). \tag{6.14}$$

Then by a triangle inequality,

$$\|W^{\mathbf{r},\mathbf{c}} - W^{\mathbf{r}',\mathbf{c}'}\|_2 \leq \|W^{\mathbf{r},\mathbf{c}} - W^{\mathbf{r}_L,\mathbf{c}_L}\|_2 + \|W^{\mathbf{r}',\mathbf{c}'} - W^{\mathbf{r}'_L,\mathbf{c}'_L}\|_2 + \|W^{\mathbf{r}_L,\mathbf{c}_L} - W^{\mathbf{r}'_L,\mathbf{c}'_L}\|_2. \tag{6.15}$$

In order to bound the last term, we can apply Theorem 6.1 since both the margins and the typical kernels are stepfunctions on the intervals $[(i-1)2^{-L}, i2^{-L})$, $i = 1, \ldots, 2^L$ and rectangles form by them, respectively. Thus by Theorem 6.1 and inequalities (6.13) and (6.14),

$$\|W^{\mathbf{r}_L,\mathbf{c}_L} - W^{\mathbf{r}'_L,\mathbf{c}'_L}\|_2^2 \leq 2C_\delta D_\delta \|(\mathbf{r}_L, \mathbf{c}_L) - (\mathbf{r}'_L, \mathbf{c}'_L)\|_1 \leq 2C_\delta D_\delta \left(4D_\delta \varepsilon + \|(\mathbf{r}, \mathbf{c}) - (\mathbf{r}', \mathbf{c}')\|_1\right).$$

The first two terms on the right-hand side of (6.15) vanishes as $L, L' \to \infty$ by Proposition 6.8. This shows (6.10).

Next, we show the Lipschitz continuity of MLEs as stated in (6.11). By mean value theorem and (4.3), ϕ restricted on $[A_\delta, B_\delta]$ is L-Lipschitz continuous for

$$L = \sup_{A_\delta \leq t \leq B_\delta} \frac{1}{\psi''(\phi(t))} = \frac{1}{\inf_{\phi(A_\delta) \leq w \leq \phi(B_\delta)} \psi''(w)}.$$

Define a kernel $V(x,y) := \boldsymbol{\alpha}(x) + \boldsymbol{\beta}(y) = \phi(W^{\mathbf{r},\mathbf{c}}(x,y))$ for $x, y \in [0,1]$ and similarly define V' using $(\boldsymbol{\alpha}', \boldsymbol{\beta}')$. Then by a simple computation using the fact that $\int_0^1 \boldsymbol{\alpha}(x)\, dx = \int_0^1 \boldsymbol{\alpha}'(x)\, dx = 0$, we have

$$\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_2^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_2^2 = \|V - V'\|_2^2 \leq L \|W^{\mathbf{r},\mathbf{c}} - W^{\mathbf{r}',\mathbf{c}'}\|_2^2.$$

Then (6.11) follows from the above and (6.10). □

Lastly in this section, we prove Proposition 6.8.

Proof of Proposition 6.8. Let Z_k denote the typical table for the $(m_k \times n_k)$ δ-tame margin $(\mathbf{r}_{m_k}, \mathbf{c}_{n_k})$. By Proposition 6.6, $W_k := W_{Z_k}$ is the typical kernel for the corresponding continuum step margin $(\bar{\mathbf{r}}_{m_k}, \bar{\mathbf{c}}_{n_k})$ (see (1.19)). By Theorem 6.7, W_k converges to some kernel W^* in L^2 as $k \to \infty$. Since $W_k \in \mathscr{W}^{[A_\delta, B_\delta]}$, it follows that $W^* \in \mathscr{W}^{[A_\delta, B_\delta]}$. It is easy to see that W^* has continuum margin $(\mathbf{r}, \mathbf{c})$.

By Lemma 6.5, there exists a constant $C = C_\delta > 0$ such that for each $k \geq 1$, there exists bounded measurable functions $\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k : [0,1] \to [-C, C]$ for which

$$W_k(x,y) \overset{a.s.}{=} \psi'(\boldsymbol{\alpha}_k(x) + \boldsymbol{\beta}_k(y)).$$

Without loss of generality, we may assume $\int_0^1 \boldsymbol{\alpha}_k(x)\, dx = 0$ for all $k \geq 1$. Then by (6.11) in Theorem 6.7, we have that $\boldsymbol{\alpha}_k \to \boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}_k \to \boldsymbol{\beta}^*$ in L^2 for some bounded measurable functions $\boldsymbol{\alpha}^*, \boldsymbol{\beta}^* : [0,1] \to [-C, C]$. Note that $\int_0^1 \boldsymbol{\alpha}^*(x)\, dx = 0$.

Now define a kernel $W^*(x,y) = \psi'(\boldsymbol{\alpha}^*(x) + \boldsymbol{\beta}^*(x))$. By mean value theorem, ψ' restricted on $[\phi(A_\delta), \phi(B_\delta)]$ is L-Lipschitz continuous for some constant $L = L(\delta) > 0$. Hence

$$\|W_k - W^*\|_2^2 \leq L(\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}^*\|_2^2 + \|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*\|_2^2).$$

It follows that $W_k \to W^*$ in L^2. Since $W_k \in \mathscr{W}^{[A_\delta, B_\delta]}_{\bar{\mathbf{r}}_k, \bar{\mathbf{c}}_k}$ and $(\bar{\mathbf{r}}_k, \bar{\mathbf{c}}_k) \to (\mathbf{r}, \mathbf{c})$ in L^1, this yields $W^* \in \mathscr{W}^{[A_\delta, B_\delta]}_{\mathbf{r},\mathbf{c}}$. By definition of W^* and Lemma 6.5, we deduce that W^* is the unique typical kernel for margin $(\mathbf{r}, \mathbf{c})$. Since W^* takes values from $[A_\delta, B_\delta]$, we conclude that $(\mathbf{r}, \mathbf{c})$ is δ-tame. □

7. Proof of phase diagrams for tame margins

In this section, we prove various sufficient conditions for tame margins stated in Section 1.5. Our first goal is to prove Theorem 1.22, which requires some preparation. We first reduce the problem to symmetric margins.

Lemma 7.1 (Reduction to symmetric margins)**.** *Suppose μ is such that Θ° is unbounded. Fix an $m \times n$ margin $(\mathbf{r}, \mathbf{c})$ with $\mathscr{T}(\mathbf{r}, \mathbf{c}) \cap (A, B)^{m \times n}$ non-empty and assume $s < \mathbf{r}(i)/n < t$ and $s < \mathbf{c}(j)/n < t$ for all i, j for some $s, t \in (A, B)$. Then for all k_0 sufficiently large, there exists a $k_0 \times k_0$ symmetric margin $(\tilde{\mathbf{r}}, \tilde{\mathbf{r}})$ such that $(\mathbf{r}, \mathbf{c})$ is δ-tame if $(\tilde{\mathbf{r}}, \tilde{\mathbf{r}})$ is so, and $s < \tilde{\mathbf{r}}(i)/k_0 < t$ for all i.*

Proof. Without loss of generality we assume $(-\infty, a) \subseteq \Theta^\circ$ for some $a \in \mathbb{R}$. By Lemma 4.2, there exists an MLE $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and a unique typical table Z for $(\mathbf{r}, \mathbf{c})$. By permuting the rows and columns, without loss of generality assume that the coordinates in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are ascending. We will consider to symmeri cases: (1) $\boldsymbol{\alpha}(m) - \boldsymbol{\alpha}(1) \leq \boldsymbol{\beta}(n) - \boldsymbol{\beta}(1)$ and (2) $\boldsymbol{\alpha}(m) - \boldsymbol{\alpha}(1) > \boldsymbol{\beta}(n) - \boldsymbol{\beta}(1)$. We will first argue for case (1).

Assume $\boldsymbol{\alpha}(m) - \boldsymbol{\alpha}(1) \leq \boldsymbol{\beta}(n) - \boldsymbol{\beta}(1)$. By shifting the MLE, without loss of generality also assume $\boldsymbol{\alpha}(m) = \boldsymbol{\beta}(n)$, so $\boldsymbol{\beta}(1) \leq \boldsymbol{\alpha}(1)$. Since Θ° is an open interval and it is assumed to be unbounded, This implies that $2\boldsymbol{\alpha}(m) = \boldsymbol{\alpha}(m) + \boldsymbol{\beta}(n) = 2\boldsymbol{\beta}(n) < a$. Fix an integer $k \geq 1$ and define $\tilde{\boldsymbol{\alpha}}_k := ((\mathbf{1}_k \otimes \boldsymbol{\alpha})^\top, \boldsymbol{\beta}^\top)^\top$, which is obtained by concatenating $\boldsymbol{\alpha}$ k times followed by $\boldsymbol{\beta}$. Let $k_0 := km + n$. Define the following symmetric $k_0 \times k_0$ matrix with $(k+1) \times (k+1)$ block structure:

$$\tilde{Z} := \psi'(\tilde{\boldsymbol{\alpha}} \oplus \tilde{\boldsymbol{\alpha}}^\top) = \begin{bmatrix} \mathbf{1}_k \mathbf{1}_k^\top \otimes \psi'(\boldsymbol{\alpha} \oplus \boldsymbol{\alpha}^\top) & \mathbf{1}_k \otimes \psi'(\boldsymbol{\alpha} \oplus \boldsymbol{\beta}^\top) \\ \mathbf{1}_k^\top \otimes \psi'(\boldsymbol{\alpha}^\top \oplus \boldsymbol{\beta}) & \psi'(\boldsymbol{\beta} \oplus \boldsymbol{\beta}^\top) \end{bmatrix}.$$

Let $\tilde{\mathbf{r}}$ denote the row sums of $\tilde{Z}$. Then by Theorem 1.7, $(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\alpha}})$ is an MLE for $(\tilde{\mathbf{r}}, \tilde{\mathbf{r}})$ and $\tilde{Z}$ is the typical table for $(\tilde{\mathbf{r}}, \tilde{\mathbf{r}})$. From the construction, it follows that $(\mathbf{r}, \mathbf{c})$ is δ-tame if the symmetric margin $(\tilde{\mathbf{r}}, \tilde{\mathbf{r}})$ is so.

Next, since $\boldsymbol{\alpha}(1) \geq \boldsymbol{\beta}(1)$ and $\boldsymbol{\alpha}(m) = \boldsymbol{\beta}(n)$,

$$\tilde{\mathbf{r}}(1) = \mathbf{r}(1) + k \sum_{1 \leq i \leq m} \psi'(\boldsymbol{\alpha}(1) + \boldsymbol{\alpha}(i)) \geq \mathbf{r}(1) + \sum_{1 \leq i \leq m} \psi'(\boldsymbol{\beta}(1) + \boldsymbol{\alpha}(i)) = \mathbf{r}(1) + k\,\mathbf{c}(1),$$

$$\tilde{\mathbf{r}}(m) = \mathbf{r}(m) + k \sum_{1 \leq i \leq m} \psi'(\boldsymbol{\alpha}(m) + \boldsymbol{\alpha}(i)) = \mathbf{r}(m) + k \sum_{1 \leq i \leq m} \psi'(\boldsymbol{\beta}(n) + \boldsymbol{\alpha}(i)) = \mathbf{r}(m) + k\,\mathbf{c}(n),$$

$$\tilde{\mathbf{r}}(km+1) = k\,\mathbf{c}(1) + \sum_{1 \leq i \leq n} \psi'(\boldsymbol{\beta}(1) + \boldsymbol{\beta}(j)) \leq k\,\mathbf{c}(1) + \mathbf{r}(1),$$

$$\tilde{\mathbf{r}}(km+n) = k\,\mathbf{c}(n) + \sum_{1 \leq j \leq n} \psi'(\boldsymbol{\beta}(n) + \boldsymbol{\beta}(j)) = k\,\mathbf{c}(n) + \sum_{1 \leq j \leq n} \psi'(\boldsymbol{\alpha}(m) + \boldsymbol{\beta}(j)) = k\,\mathbf{c}(n) + \mathbf{r}(m).$$

It follows that the largest row sum of $\tilde{Z}$ is $\tilde{\mathbf{r}}(m) = \tilde{\mathbf{r}}(k_0) \leq k\mathbf{c}(n) + \mathbf{r}(m)$ and its smallest row sum is $\tilde{\mathbf{r}}(km+1) = k\mathbf{c}(1) + M$, where $M := \sum_{1 \leq j \leq n} \psi'(\boldsymbol{\beta}(1) + \boldsymbol{\beta}(j))$ is a constant that does not depend on k. By the hypothesis, $\tilde{\mathbf{r}}(k_0) < ktm + tn \leq tk_0$ and $\tilde{\mathbf{r}}(km+1) \geq ks'm + M$ for some $s' > s$. It follows that for k sufficiently large, $\tilde{\mathbf{r}}(km+1)/k_0 \geq (s+s')/2 > s$, as desired.

For case (2), we switch the roles of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ (e.g., $\tilde{\boldsymbol{\beta}}_k := ((\mathbf{1}_k \otimes \boldsymbol{\beta})^\top, \boldsymbol{\alpha}^\top)^\top$) and apply the same argument for case (1). □

Next, we establish a technical lemma, which essentially reduces the proof of Theorem 1.22 to an extreme Barvinok margin. Recall that we say a function $h(t)$ *log-convex* if $\log h(x)$ is convex.

Lemma 7.2 (Reducing symmetric margin to Barvinok margin)**.** *Suppose ψ'' is log-convex on Θ. Fix a symmetric $n \times n$ margin $(\tilde{\mathbf{r}}, \tilde{\mathbf{r}})$ such that $0 < \tilde{\mathbf{r}}(1) \leq \cdots \leq \tilde{\mathbf{r}}(n)$. Let $(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\alpha}})$ be its symmetric MLE. Then there exists a symmetric margin $(\mathbf{r}^*, \mathbf{r}^*)$ such that*

$$\tilde{\mathbf{r}}(1) \leq \mathbf{r}^*(1) = \cdots = \mathbf{r}^*(n-1) \leq \mathbf{r}^*(n) \leq \tilde{\mathbf{r}}(n) \tag{7.1}$$

and the corresponding symmetric MLE $(\boldsymbol{\alpha}^*, \boldsymbol{\alpha}^*)$ satisfies $\tilde{\boldsymbol{\alpha}}(1) \leq \boldsymbol{\alpha}^*(1) \leq \boldsymbol{\alpha}^*(n) = \tilde{\boldsymbol{\alpha}}(n)$.

Proof. Without loss of generality assume $A \leq 0 < B$ and $n \geq 3$. The overall idea of the construction is to evolve the MLEs smoothly so that the first $n-1$ row sums merge together while the last coordinate of the MLE stays put.

1. Computing the Jacobian.

Next, for each $\boldsymbol{\alpha} \in [\tilde{\boldsymbol{\alpha}}(1), \tilde{\boldsymbol{\alpha}}(n)]^n$, define the corresponding row margin $\mathbf{r}(\boldsymbol{\alpha}) \in \mathbb{R}^n$ by

$$\mathbf{r}(\boldsymbol{\alpha})(i) = \sum_{j=1}^{n} \psi'(\boldsymbol{\alpha}(i) + \boldsymbol{\alpha}(j)) \qquad \text{for } 1 \leq i \leq n.$$

That is, $(\mathbf{r}(\boldsymbol{\alpha}), \mathbf{r}(\boldsymbol{\alpha}))$ is the symmetric margin corresponding to the symmetric MLE $(\boldsymbol{\alpha}, \boldsymbol{\alpha})$. Let $E(\boldsymbol{\alpha}) = (E_{ij})_{i,j}$ denote the $n \times n$ positive symmetric matrix with $E_{ij} = \psi''(\boldsymbol{\alpha}(i) + \boldsymbol{\alpha}(j))$. Let $E_{i\bullet}$ and $E_{\bullet j}$ denotes the ith row sum and the jth column sum of E, respectively. Since E is symmetric, $E_{\bullet i} = E_{i\bullet}$. Then the Jacobian of the margin $\mathbf{r}$ w.r.t. the MLE $\boldsymbol{\alpha}$ is given by

$$J_{\mathbf{r}(\boldsymbol{\alpha});\boldsymbol{\alpha}} = \begin{bmatrix} E_{11} + E_{\bullet 1} & E_{12} & E_{13} & \cdots & E_{1n} \\ E_{21} & E_{22} + E_{\bullet 2} & E_{23} & \cdots & E_{2n} \\ \vdots & & & & \vdots \\ E_{n1} & E_{n2} & \cdots & E_{n,n-1} & E_{nn} + E_{\bullet n} \end{bmatrix}.$$

2. Evolution of the MLE to synchronize the first $n-1$ coordinates.

Now we define a piecewise smooth evolution of the MLE $\boldsymbol{\alpha}_t$, $t \geq 0$ with $\boldsymbol{\alpha}_0 = \tilde{\boldsymbol{\alpha}}$. Our construction is so that $\boldsymbol{\alpha}_t$ converges to $\boldsymbol{\alpha}^*$ where

$$\tilde{\boldsymbol{\alpha}}(1) \leq \boldsymbol{\alpha}^*(1) = \cdots = \boldsymbol{\alpha}^*(n-1) \leq \boldsymbol{\alpha}^*(n) = \tilde{\boldsymbol{\alpha}}(n). \tag{7.2}$$

Furthermore, the corresponding row sum vector, say $\mathbf{r}^*$, should satisfy (7.1). The evolution $\boldsymbol{\alpha}_t$ will be defined inductively. Let $i_1 \in \{1, \ldots, n-2\}$ be such that the gap $\boldsymbol{\alpha}_0(i_1+1) - \boldsymbol{\alpha}_0(i_1)$ is maximized.

Then we evolve $\boldsymbol{\alpha}_t$ so that its i_1th coordinate increases and its i_1+1st coordinate decreases until some time t_1 that the gap becomes zero. During this time interval $[0,t_1]$, we wish the minimum and the maximum row sums move toward each other so that (7.1) holds at intermediate times. The time derivative $\dot{\boldsymbol{\alpha}}_t$ is defined as

$$\dot{\boldsymbol{\alpha}}_t = \mathbf{e}_{i_1} - a(t)\mathbf{e}_{i_1+1} \quad \text{with} \quad a(t) = \frac{E_{n,i_1}(\boldsymbol{\alpha}_t)}{E_{n,i_1+1}(\boldsymbol{\alpha}_t)} \quad \text{as long as } \boldsymbol{\alpha}_t(i_1) < \boldsymbol{\alpha}_t(i_1+1), \tag{7.3}$$

where $\mathbf{e}_i$ denotes the ith standard basis vector in $\mathbb{R}^n$ From now we will supress the dependence on $\boldsymbol{\alpha}_t$ in $E_{ij}(\boldsymbol{\alpha}_t)$. Let $t_1 = \inf\{t \geq 0 : \boldsymbol{\alpha}_t(i_1) = \boldsymbol{\alpha}_t(i_1+1)\}$. Note that $t_1 \leq \boldsymbol{\alpha}_0(i_1+1) - \boldsymbol{\alpha}_0(i_1)$.

Starting from time t_1, let $i_2 \in \{1,\dots,n-2\}$ such that the gap $\boldsymbol{\alpha}_{t_1}(j_2+1) - \boldsymbol{\alpha}_{t_1}(i_2)$ is maximized and we define the dynamics similary using the index i_2 until some finite time $t_2 \geq t_1$ such that $\boldsymbol{\alpha}_{t_2}(i_2) = \boldsymbol{\alpha}_{t_2}(i_2+1)$, and so on. During this process, the last coordinate $\boldsymbol{\alpha}_0(n)$ remains the same.

Next, we will observe that the $\boldsymbol{\alpha}_t$ converges to some $\boldsymbol{\alpha}^*$ satisfying (7.2) as $t \to \infty$. To this effect, we claim

$$\boldsymbol{\alpha}_{t_k}(n-1) - \boldsymbol{\alpha}_{t_k}(i_k) \leq \left(1 - \frac{1}{2n}\right)(\boldsymbol{\alpha}_{t_{k-1}}(n-1) - \boldsymbol{\alpha}_{t_{k-1}}(i_k)). \tag{7.4}$$

To show the claim, write $\mathscr{A} = \boldsymbol{\alpha}_{t_{k-1}}(n-1) - \boldsymbol{\alpha}_{t_{k-1}}(i_k)$ and $\mathscr{A}' = \boldsymbol{\alpha}_{t_k}(n-1) - \boldsymbol{\alpha}_{t_k}(i_k)$. Denote $\Delta_j := \boldsymbol{\alpha}_{t_{k-1}}(j+1) - \boldsymbol{\alpha}_{t_{k-1}}(j) \geq 0$ so that we can write $\mathscr{A} = \sum_{i_k \leq j < n-1} \Delta_j$. Since $a(t) \in [0,1]$ for all $t \geq 0$, for each $k \geq 1$, the two coordinates $\boldsymbol{\alpha}_{t_{k-1}}(i_k) < \boldsymbol{\alpha}_{t_{k-1}}(i_k+1)$ (setting $t_0 = 0$) are replaced by some value between their mean and the larger one $\boldsymbol{\alpha}_{t_{k-1}}(i_k+1)$ at time t_k. In particular,

$$\boldsymbol{\alpha}_{t_k-1}(i_k+1) - \boldsymbol{\alpha}_{t_k}(i_k+1) \leq \Delta_{i_k}/2.$$

Writing $\mathscr{A}' = \Delta_{n-2} + \cdots + \Delta_{i_k+1} + \boldsymbol{\alpha}_{t_{k-1}}(i_k+1) - \boldsymbol{\alpha}_{t_k}(i_k+1)$ (using $\boldsymbol{\alpha}_{t_k}(i_k) = \boldsymbol{\alpha}_{t_k}(i_k+1)$),

$$\mathscr{A}' \leq \Delta_{n-2} + \cdots + \Delta_{i_k+1} + \frac{\Delta_{i_k}}{2} = \mathscr{A} - \frac{\Delta_{i_k}}{2}.$$

Since Δ_{i_k} is the largest gap among the $n-1$ ones at time t_{k-1},

$$\mathscr{A} - \mathscr{A}' \geq \frac{\Delta_{i_k}}{2} \geq \frac{1}{2n} \sum_{i_k \leq j < n-1} \Delta_j = \frac{1}{2n}\mathscr{A}.$$

Simplifying the above shows the claim (7.4).

Now let $\liminf_{k\to\infty} i_k = i^* \geq 1$. Then $i_k \geq i^*$ for all $k \geq k_0$ for some $k_0 \geq 1$. In turn, $\boldsymbol{\alpha}_{t_{k-1}}(n-1) - \boldsymbol{\alpha}_{t_{k-1}}(i^*)$ is a decreasing function in t for $t \geq t_{k_0}$ and it contracts by at least a constant factor due to (7.4) whenever $i_k = i^*$ for $k \geq k_0$. Thus $\boldsymbol{\alpha}_{t_{k-1}}(n-1) - \boldsymbol{\alpha}_{t_{k-1}}(i^*) \to 0$ as $k \to \infty$. In particular, $\boldsymbol{\alpha}_{t_{k-1}}(i_k+1) - \boldsymbol{\alpha}_{t_{k-1}}(i_k) \to 0$ as $k \to \infty$. But since this is the maximum gap at time t_{k-1}, we deduce $\boldsymbol{\alpha}_{t_{k-1}}(n-1) - \boldsymbol{\alpha}_{t_{k-1}}(1) \to 0$ as $k \to \infty$. Also note that $\boldsymbol{\alpha}_{t_{k-1}}(i^*)$ is increasing and $\boldsymbol{\alpha}_{t_{k-1}}(n-1)$ is decreasing for $k \geq k_0$. This shows that $\boldsymbol{\alpha}_{t_{k-1}} \to \boldsymbol{\alpha}^*$ as $t \to \infty$ for some $\boldsymbol{\alpha}^*$ satisfying (7.2).

3. Contraction of the range of the row sums.

Lastly, denoting $\mathbf{r}_t := \mathbf{r}(\boldsymbol{\alpha}_t)$, we will show that $\dot{\mathbf{r}}_t(1) \geq 0$ and $\dot{\mathbf{r}}(n) \leq 0$ during (t_{k-1}, t_k) for each $k \geq 1$. This is enough to conclude (7.1). Fix $k \geq 1$. We wish to show

$$\dot{\mathbf{r}}_t(1) = E_{1,i_k} + E_{1\bullet}\mathbf{1}(i_k = 1) - E_{1,i_k+1}a(t) \geq 0,$$
$$\dot{\mathbf{r}}_t(n) = E_{n,i_k} - E_{n,i_k+1}a(t) \leq 0,$$

The expressions for the time derivates of the extreme row sums are clear by chain rule and (7.3). The choice $a(t) = E_{n,i_k}/E_{n,i_k+1}$ in (7.3) is valid if $\frac{E_{1,i_k}}{E_{1,i_k+1}} \geq \frac{E_{n,i_k}}{E_{n,i_k+1}}$, equivalently,

$$\frac{\psi''(\boldsymbol{\alpha}_t(1) + \boldsymbol{\alpha}_t(i_k))}{\psi''(\boldsymbol{\alpha}_t(1) + \boldsymbol{\alpha}_t(i_k+1))} \geq \frac{\psi''(\boldsymbol{\alpha}_t(n) + \boldsymbol{\alpha}_t(i_k))}{\psi''(\boldsymbol{\alpha}_t(n) + \boldsymbol{\alpha}_t(i_k+1))}. \tag{7.5}$$

Recall that if $h(\cdot)$ is positive and log-convex, then for $x \le y$ and $z \le w$, we have

$$\frac{h(x+z)}{h(x+w)} \ge \frac{h(y+z)}{h(y+w)}.$$

Since the dynamics respects the ordering $\boldsymbol{\alpha}_t(1) \le \cdots \le \boldsymbol{\alpha}_t(n)$, (7.5) follows since ψ'' is positive and log-convexity. This finishes the proof. □

We are now ready to establish the first part of Theorem 1.22.

Proof of the first part of Theorem 1.22. Suppose (1.23) holds. By Lemma 7.1, without loss of generality, we assume $\mathbf{r} = \mathbf{c}$ and $m = n \ge N_0$ for some large constant N_0 to be determined. Let $(\boldsymbol{\alpha}, \boldsymbol{\alpha})$ be the unique symmetric MLE for $(\mathbf{r}, \mathbf{r})$. Without loss of generality, assume $n \ge 3$ and $\mathbf{r}(1) \le \cdots \le \mathbf{r}(n)$. Then $\boldsymbol{\alpha}(1) \le \cdots \le \boldsymbol{\alpha}(n)$. Let $Z = (z_{ij})$ denote the typical table for $(\mathbf{r}, \mathbf{r})$. By the hypothesis, we may choose $\delta > 0$ small enough so that

$$\phi(A) < \phi(A_\delta) < 3\phi(s) - 2\phi(t) \quad \text{and} \quad 2\phi(t) - \phi(s) < \phi(B_\delta) < \phi(B).$$

We will first show $z_{n,n} \le B_\delta$. Let $\boldsymbol{\alpha}^*$ and $\mathbf{r}^*$ be as in Lemma 7.2. Since $\boldsymbol{\alpha}(n) = \boldsymbol{\alpha}^*(n)$, it suffices to upper bound $\boldsymbol{\alpha}^*(n)$. By the margin condition, Lemma 7.2, and Jensen's inequality (ψ' is convex since ψ'' is increasing by the hypothesis),

$$\begin{aligned} s \le n^{-1}\mathbf{r}(1) \le n^{-1}\mathbf{r}^*(1) &= (1-n^{-1})\psi'(2\boldsymbol{\alpha}^*(1)) + n^{-1}\psi'(\boldsymbol{\alpha}^*(1) + \boldsymbol{\alpha}^*(n)), \\ t \ge n^{-1}\mathbf{r}(n) \ge n^{-1}\mathbf{r}^*(n) &= (1-n^{-1})\psi'(\boldsymbol{\alpha}^*(1) + \boldsymbol{\alpha}^*(n)) + n^{-1}\psi'(2\boldsymbol{\alpha}^*(n)) \\ &\ge \psi'\big((1+n^{-1})\boldsymbol{\alpha}(n) + (1-n^{-1})\boldsymbol{\alpha}^*(1)\big) \ge \psi'\big(\boldsymbol{\alpha}^*(n) + \boldsymbol{\alpha}^*(1)\big). \end{aligned} \tag{7.6}$$

Combining the two inequalities, we get $s \le (1-n^{-1})\psi'(2\boldsymbol{\alpha}^*(1)) + n^{-1}t$, which yields

$$\phi\big((1-n^{-1})^{-1}s - n^{-1}t\big) \le 2\boldsymbol{\alpha}^*(1).$$

Then combining the above with $t \ge \psi'(\boldsymbol{\alpha}^*(1) + \boldsymbol{\alpha}^*(n))$ from (7.6),

$$2\boldsymbol{\alpha}^*(n) \le 2\phi(t) - 2\boldsymbol{\alpha}^*(1) \le 2\phi(t) - \phi\big((1-n^{-1})^{-1}s - n^{-1}t\big). \tag{7.7}$$

By continuity of ϕ, the right-hand side above converges to $2\phi(t) - \phi(s)$ as $n \to \infty$. Hence exists a constant N_0 depending only on μ, s, t such that the right-hand side above is well-defined and it is at most $\phi(B_\delta)$ for all $n \ge N_0$. Recalling $\boldsymbol{\alpha}(n) = \boldsymbol{\alpha}^*(n)$, this shows

$$z_{n,n} = \psi'(2\boldsymbol{\alpha}(n)) \le \psi'(\phi(B_\delta)) = B_\delta \qquad \text{for all } n \ge N_0.$$

Next, it now remains to show $z_{11} \ge A_\delta$. Note that

$$s \le n^{-1}\mathbf{r}(1) = n^{-1}\sum_{j=1}^{n} \psi'(\boldsymbol{\alpha}(1) + \boldsymbol{\alpha}(j)) \le \psi'(\boldsymbol{\alpha}(1) + \boldsymbol{\alpha}(n)).$$

Using the upper bound on $\boldsymbol{\alpha}(n)$ in (7.7), this gives

$$2\boldsymbol{\alpha}(1) \ge 2\phi(s) - 2\boldsymbol{\alpha}(n) \ge 2\phi(s) - 2\phi(t) + \phi\big((1-n^{-1})^{-1}s - n^{-1}t\big).$$

The last expression above converges to $3\phi(s) - 2\phi(t)$. By enlarging N_0 if necessary, it follows that the last expression above is at least $\phi(A_\delta)$ for all $n \ge N_0$. This yields $z_{11} = \psi'(2\boldsymbol{\alpha}(1)) \ge A_\delta$, as desired. □

Instead of proving the second part of Theorem 1.22 directly, we will establish a stronger result in Proposition 7.3 below on the sharp phase transition of typical tables for Barvinok margins. Essentially for Barvinok margins, the sufficient condition for tameness in Theorem 1.22 is also necessary. A special case of this result, specifically for contingency tables, was previously established

by Dittmer, Lyu, and Pak [DLP20, Lem. 5.1]. The argument presented here for the general case is significantly simpler. The following result also directly implies Corollary 1.21.

Proposition 7.3 (Sharp phase transition in typical tables for Barvinok margins)**.** *Suppose μ is a mearsure on $\mathbb{R}_{\geq 0}$ such that $A = 0$, $B = \infty$, and $\phi(B) < \infty$. Fix constants $s, t \in (A, B)$ with $s \leq t$ and a $\rho \in [0, 1)$. Fix two converging sequences $s_n \to s$ and $t_n \to t$ in (A, B). Consider the following symmetric margin $(\mathbf{r}, \mathbf{r})$ with (1.22). Then the following hold:*

(i) *If $2\phi(t) - \phi(s) < \phi(B)$, then*

$$z_{11} = s_n + O(n^{\rho-1}), \quad z_{1,n} = t_n + O(n^{\rho-1}),$$
$$z_{nn} = \psi'\left(2\phi(t_n - O(n^{\rho-1})) + O(n^{\rho-1}) - \phi(s_n - O(n^{\rho-1}))\right) = O(1).$$

In particular, for some fixed $\delta > 0$, $(\mathbf{r}, \mathbf{c})$ is δ-tame for all $n \geq 1$.

(ii) *If $2\phi(t) - \phi(s) > \phi(B)$, then*

$$z_{11} = s_n + O(n^{\rho-1}), \quad z_{1,n} = \psi'\left(\frac{\phi(B) + \phi(s_n + \varepsilon)}{2} + o(1)\right),$$
$$\lfloor n^{\rho-1} \rfloor \psi'(z_{nn}) = t_n - \psi'\left(\frac{\phi(B) + \phi(s_n + O(n^{\rho-1}))}{2} + o(1)\right) = \Omega(1).$$

In particular, $(\mathbf{r}, \mathbf{c})$ is not δ-tame for all $n \geq 1$ for any fixed $\delta > 0$.

Proof. Note that for each n, $\mathcal{T}(\mathbf{r}, \mathbf{c})$ contains a positive real-valued matrix since it contains the Fisher-Yates table $(\mathbf{r}(i)\mathbf{c}(j)/N)_{i,j}$ where N denotes the total sum. Hence by Lemma 4.2, there exists a unique typical table $Z = Z^{\mathbf{r},\mathbf{c}}$. By Theorem 1.7, there exists an MLE $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ for margin $(\mathbf{r}, \mathbf{c})$. By shifting the MLE, we may assume that $\boldsymbol{\alpha}(1) = \boldsymbol{\beta}(1)$. By symmetry, Z is symmetric, so this implies $\boldsymbol{\alpha} = \boldsymbol{\beta}$. Denoting $\alpha_1 := \boldsymbol{\alpha}(1)$ and $\alpha_2 := \boldsymbol{\alpha}(n)$. Then $\alpha_1 \leq \alpha_2$, $z_{ij} = \psi'(2\boldsymbol{\alpha}_1)$ for $1 \leq i, j \leq n - \lfloor n^\rho \rfloor$, $z_{ij} = \psi'(\alpha_1 + \alpha_2)$ for $1 \leq i \leq n - \lfloor n^\rho \rfloor$ and $j > n - \lfloor n^\rho \rfloor$, and $z_{ij} = \psi'(2\alpha_2)$ for $i, j > n - \lfloor n^\rho \rfloor$. Hence the margin condition for Z reduces to the following system of equations:

$$\begin{aligned} s_n n &= (n - \lfloor n^\rho \rfloor)\psi'(2\alpha_1) + \lfloor n^\rho \rfloor \psi'(\alpha_1 + \alpha_2) \\ t_n n &= (n - \lfloor n^\rho \rfloor)\psi'(\alpha_1 + \alpha_2) + \lfloor n^\rho \rfloor \psi'(2\alpha_2). \end{aligned} \tag{7.8}$$

Since $\psi' \geq 0$, from the second equation, we deduce $\psi'(\alpha_1 + \alpha_2) \leq t_n = O(1)$. From the first equation,

$$2\alpha_1 = \phi(s_n - O(n^{\rho-1})). \tag{7.9}$$

It follows that, dropping the term $\lfloor n^\rho \rfloor \psi'(2\alpha_2) \geq 0$ from the second equation in (7.8),

$$\alpha_2 \leq \phi(t_n) - \phi(s_n - O(n^{\rho-1}))/2. \tag{7.10}$$

Now suppose $2\phi(t) - \phi(s) < \phi(B)$. Since ϕ is continuous, there exists $\varepsilon > 0$ such that

$$2\phi(t + \varepsilon) + \varepsilon - \phi(s - \varepsilon) < \phi(B_\varepsilon).$$

From (7.10), it follows that for all sufficiently large $n \geq 1$, $2\alpha_2 \leq \phi(B_\varepsilon)$. Hence $\psi'(2\alpha_2) = O(1)$, so from the second equation in (7.8), $\alpha_1 + \alpha_2 = \phi(t_n - O(n^{\rho-1})) + O(n^{\rho-1})$, so using (7.9),

$$\alpha_2 = \phi(t_n - O(n^{\rho-1})) + O(n^{\rho-1}) - \phi(s_n - O(n^{\rho-1}))/2 < \phi(B_\varepsilon)/2$$

for $n \geq 1$ large. Thus in this case, $(\mathbf{r}, \mathbf{r})$ is δ-tame for a fixed δ for all $n \geq 1$ sufficiently large. The asymptotic expressions for z_{ij} follow from $z_{11} = \psi'(2\alpha_1)$, $z_{1,n} = \psi'(\alpha_1 + \alpha_2)$, and $z_{nn} = \psi'(2\alpha_2)$.

Next, suppose $2\phi(t) - \phi(s) > \phi(B)$. Then there exists $\varepsilon > 0$ such that for all n large enough,

$$2\phi(t_n - \varepsilon) - \phi(s_n) > \phi(B). \tag{7.11}$$

Using $2\alpha_2 < \phi(B)$ with (7.9) and (7.11), for all sufficiently large $n \geq 1$,

$$\alpha_1 + \alpha_2 < \alpha_1 + \frac{\phi(B)}{2} = \frac{\phi(s_n - O(n^{\rho-1})) + \phi(B)}{2} \leq \phi(t_n - \varepsilon).$$

Thus, in the second equation in (7.8), $(n - \lfloor n^\rho \rfloor)\psi'(\alpha_1 + \alpha_2)$ can be at most $n(t_n - \varepsilon)$ and the other term $\lfloor n^\rho \rfloor \psi'(2\alpha_2)$ must make up for the linearly large remainder term $\geq \varepsilon n$. Namely,

$$\lfloor n^{\rho-1} \rfloor \psi'(2\alpha_2) = t_n - (1 - \lfloor n^\rho \rfloor / n)\psi'(\alpha_1 + \alpha_2) \geq \varepsilon \tag{7.12}$$

for all n large enough, so we have $2\alpha_2 \geq \phi(\varepsilon n^{1-\rho})$. Hence $z_{nn} \geq \varepsilon n^{1-\rho} \to \infty$ as $n \to \infty$ and $2\alpha_2 = \phi(B) - o(1)$. Thus $(\mathbf{r}, \mathbf{r})$ is not δ-tame for all $n \geq 1$ sufficiently large for any fixed $\delta > 0$. We can further obtain an asymptotic expression for z_{nn}. Using (7.9) and $2\alpha_2 = \phi(B) - o(1)$, $2\alpha_1 + 2\alpha_2 = \phi(B) + o(1) + \phi(s_n - O(n^{\rho-1}))$. Plugging this in to the equality in (7.12),

$$\lfloor n^{\rho-1} \rfloor \psi'(2\alpha_2) = t_n - \psi'\left(\frac{\phi(B) + \phi(s_n - O(n^{\rho-1}))}{2} + o(1) \right).$$

□

Remark 7.4. Recently in [BR24], Barvinok and Rudelson observed that the Barvinok margin $\mathbf{r} = \mathbf{c} = (n, \dots, n, \lambda n)$ is δ-tame if $\lambda < 2$ and it is not if $\lambda > 2$ when μ is the Lebesgue measure on $\mathbb{R}_{\geq 0}$. In particular, they noted that $z_{nn} = \Omega(n)$. Recall that for this base measure, we have $\psi'(\theta) = \phi(\theta) = -1/\theta$ (see Ex. 3.6). Hence our Proposition 7.3 implies more precise asymptotics of the typical table in the supercritical case $\lambda > 2$. For instance, $z_{nn} = \frac{\lambda}{2-\lambda} + O(n^{-1})$ for $\lambda < 2$ and $z_{nn} = (\lambda - 2 + o(1))n + O(n^{-1})$ for $\lambda > 2$.

We can now quickly deduce Corollary 1.23 from the results we established above.

Proof of Corollary 1.23. First, we will show that $t/s < \lambda_c$ implies tameness. For all base measures in the statement, ψ'' is increasing and log-convex (see Sec. 3) with $\phi(A) = \infty$ and $\phi(B) = 0$. Hence by Theorem 1.22, $(\mathbf{r}, \mathbf{c})$ is δ-tame for some $\delta = \delta(\mu, s, t) > 0$ if $2\phi(t) - \phi(s) < 0$. A simple computation using explicit forms of ϕ shows that this condition is simply $t/s < \lambda_c$, where the critical ratio λ_c is given in (1.24). For the other direction, note that by Proposition 7.3, the (n, n) entry of the typical table corresponding to the Barvinok margin (1.22) with $t_n = t + o(1)$, $s_n = s + o(1)$ blows up as $n \to \infty$ if $2\phi(t) - \phi(s) > 0$, which occurs if and only if $t/s > \lambda_c$. This finishes the proof. □

Lastly in this section, we prove the sharp condition for tame margins for $B < \infty$ stated in Theorem 1.20. Our proof is a combination of the symmetrization technique (Lemma 7.1) and a minor modification of the proof of [BH13, Lem. 12.3].

Proof of Theorem 1.20. By shifting, we may assume $A = 0 \leq B < \infty$. Then the inequality (1.21) reduces to $(s+t)^2 < 4Bs$. We first argue that for each $(s, t) \in (0, B)^2$ with $s \leq t$, $(s, t) \notin \Omega(\mu)$ if $(s+t)^2 > 4Bs$. Fix such s and t. Fix $x, y \in (0, 1)$ and $c \in \mathbb{R}$. Consider the following n-dimensional dual variables $\boldsymbol{\alpha} = (-n, \dots, -n, 0, \cdots, 0)$ and $\boldsymbol{\beta} = (n + c, \cdots, n + c, c, \cdots, c)$, where $\boldsymbol{\alpha}$ repeats $-n$ $\lfloor xn \rfloor$ times and $\boldsymbol{\beta}$ repeats $n + c$ $\lfloor yn \rfloor$ times. Let $Z := \psi'(\boldsymbol{\alpha} \oplus \boldsymbol{\beta})$ and let $(\mathbf{r}, \mathbf{c})$ denote the margin of Z. By Theorem 1.7, $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is an MLE for $(\mathbf{r}, \mathbf{c})$ and Z is the typical table for $(\mathbf{r}, \mathbf{c})$. By construction, the margin $(\mathbf{r}, \mathbf{c})$ is not δ-tame for any constant $\delta > 0$ independent of n since the MLEs diverge as $n \to \infty$.

Since $\psi'(-n) \to 0$ and $\psi'(n + c) \to B$ as $n \to \infty$, it is easy to see that the row and column sums rescaled by n^{-1} as $n \to \infty$ have only the following four values:

$$y\psi'(c), \quad yB + (1 - y)\psi'(c), \quad x\psi'(c) + (1 - x)B, \quad (1 - x)\psi'(c).$$

If $y \leq 1 - x$, then the minimum and the maximum among the above values are $y\psi'(c)$ and $x\psi'(c) + (1 - x)B$, respectively. Thus the proof is finished once we choose $x, y \in (0, 1)$ and c so that $y \leq 1 - x$,

$s = y\psi'(c)$, and $t = x\psi'(c) + (1-x)B$. We choose c so that $0 < s \le \psi'(c) \le t < B$ and then choose $x = x(c)$ and $y = y(c)$ to satisfy the required identities. We can then make $\psi'(c)$ is sufficiently close to s to make sure $y \le 1 - x$. This shows $(s,t) \notin \Omega(\mu)$ if $(s+t)^2 > 4Bs$.

Next, we show that $(s+t)^2 < 4Bs$ implies $(s,t) \in \Omega(\mu)$. By Lemma 7.1, without loss of generality, we assume $m = n$ and $\mathbf{r} = \mathbf{c}$. Let $(\boldsymbol{\alpha}, \boldsymbol{\alpha})$ be the unique symmetric MLE for $(\mathbf{r}, \mathbf{r})$. Without loss of generality, assume $n \ge 3$ and $\mathbf{r}(1) \le \cdots \le \mathbf{r}(n)$. Then $\boldsymbol{\alpha}(1) \le \cdots \le \boldsymbol{\alpha}(n)$. We will denote $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$. Since ψ' is increasing and $\alpha_1 \le \cdots \le \alpha_n$ for all i, j, it follows that

$$z_{i,n} = \psi'(\alpha_i + \alpha_n) \ge z_{1,n} \ge \psi'(\alpha_1 + \alpha_j) = z_{1,j},$$
$$z_{n,j} = \psi'(\alpha_n + \alpha_j) \ge z_{n,1} \ge \psi'(\alpha_i + \alpha_1) = z_{i,1}.$$

Fix $\varepsilon > 0$ sufficiently small. Note that $\phi(x) \to -\infty$ as $x \searrow 0$ and $\phi(B-x) \to \infty$ as $x \searrow 0$. Hence there are constants c_1, c_2 (depending on ε) such that

$$\phi(c_1) + \phi(B - \varepsilon) \le \phi(s - \varepsilon), \qquad \phi(c_1) + \phi(B - c_2) \ge \phi(t + \varepsilon).$$

We claim that

$$\phi(c_1) \overset{(a)}{\le} \alpha_1 \le \alpha_n \overset{(b)}{\le} \phi(B - c_2). \tag{7.13}$$

Since $\phi(z_{ij}) = \alpha_i + \alpha_j$ and ϕ is increasing, the assertion follows immediately from this claim. Furthermore, (b) follows from (a). Indeed, suppose (b) is not true while (a) holds. Then

$$t \ge n^{-1}\mathbf{r}(n) = n^{-1}\sum_i z_{i,n} = n^{-1}\sum_i \psi'(\alpha_i + \alpha_n) \ge \psi'(\phi(c_1) + \phi(B - c_2)) \ge (t + \varepsilon),$$

which is a contradiction. Thus it is enough to show (7.13) (a).

Suppose for contradiction (a) does not hold, i.e., $\alpha_1 < \phi(c_1)$. Then necessarily $\alpha_n > \phi(B - \varepsilon)$, since otherwise for all $1 \le j \le n$,

$$z_{1,j} = \psi'(\alpha_1 + \alpha_j) \le \psi'(\alpha_1 + \alpha_n) \le \psi'\big(\phi(c_1) + \phi(B - \varepsilon)\big) \le \psi'(\phi(s - \varepsilon)) = s - \varepsilon.$$

So this implies $s \le \mathbf{r}(1)/n \le s - \varepsilon$, a contradiction.

Now since $\alpha_1 < \phi(c_1)$ and $\alpha_n > \phi(B - \varepsilon)$,

$$z_{i,n} = \psi'(\alpha_i + \alpha_n) \ge \psi'(\alpha_n) \ge \psi'(\phi(B - \varepsilon)) = B - \varepsilon \qquad \text{if } \alpha_i \ge 0\,,$$
$$z_{1,j} = \psi'(\alpha_1 + \alpha_j) \le \psi'(\alpha_1) \le \psi'(\phi(c_1)) = c_1 \qquad \text{if } \alpha_j \le 0.$$

Let $\rho := n^{-1}\{1 \le i \le n : \alpha_i \ge 0\}$. Then we have

$$tn \ge c_n = \sum_i z_{i,n} = \sum_{i;\alpha_i \ge 0} z_{i,n} + \sum_{i;\alpha_i < 0} z_{i,n} \ge n\rho(B - \varepsilon) + (1-\rho)nz_{1,n},$$
$$sn \le r_1 = \sum_j z_{1,j} = \sum_{j;\alpha_j < 0} z_{1,j} + \sum_{j;\alpha_j \ge 0} z_{1,j} \ge (1-\rho)c_1 + \rho n z_{1,n}.$$

It follows that

$$t \ge \rho(B - \varepsilon) + (1-\rho)z_{1,n}, \qquad s \le (1-\rho)c_1 + \rho z_{1,n}. \tag{7.14}$$

Denote $\tau := z_{1,n}$. Since $\rho \in [0,1]$,

$$t + \varepsilon \ge \rho B + (1-\rho)\tau, \qquad s - c_1 \le \rho\tau.$$

This yields

$$t + \varepsilon \ge \rho B + (1-\rho)\tau \ge 2\sqrt{B\rho\tau} - \rho\tau \ge 2\sqrt{B(s - c_1)} - (s - c_1),$$

where second inequality above uses $\frac{\rho B+\tau}{2} \geq \sqrt{\rho B \tau}$ and last inequaliy uses the fact that the function $x \mapsto 2\sqrt{Bx} - x$ is increasing for $x \in [0, B]$ together with the second inequality in (7.14). Thus if

$$t + \varepsilon < 2\sqrt{B(s - c_1)} - (s - c_1), \tag{7.15}$$

then this leads to a contradiction. Note that (7.15) holds for ε, c_1 sufficiently small if $t < 2\sqrt{Bs} - s$, or $(s+t)^2 < 4Bs$. Thus we conclude that (7.13) (a) hold, as desired. □

Lastly in this section, we prove Theorem 1.24.

Proof of Theorem 1.24. We first show the "only if" part. Fix $I \subseteq [n]$. δ-tameness implies that there is a symmetric MLE $(\boldsymbol{\alpha}, \boldsymbol{\alpha})$ for margin $(\mathbf{r}, \mathbf{r})$ such that $c_1 := \delta \leq \psi'(\boldsymbol{\alpha} \oplus \boldsymbol{\alpha}) \leq B - \delta =: c_2$. The typical table $\psi'(\boldsymbol{\alpha} \oplus \boldsymbol{\alpha}) = (z_{ij})$ satisfies the margin $(\mathbf{r}, \mathbf{r})$, so $c_1 \leq \mathbf{r}(i)/n \leq c_2$ for all i, j. Now writing $\mathbf{r}(i) = \sum_{j \in I} z_{ij} + \sum_{j \notin I} z_{ij}$ and using $z_{ij} \leq B - \delta$,

$$B|I|^2 - \sum_{i \in I} \mathbf{r}(i) \geq \delta |I|^2 - \sum_{i \in I} \sum_{j \notin I} z_{ij}.$$

Also, since $B|I| \geq \sum_{i \in I} z_{ij}$ due to $z_{ij} \in [0, B]$,

$$\sum_{j \notin I} B|I| \wedge \mathbf{r}(j) \geq \sum_{j \notin I} \sum_{i \in I} z_{ij}.$$

Combining the above, we deduce

$$B|I|^2 + \sum_{j \notin I} B|I| \wedge \mathbf{r}(j) - \sum_{i \in I} \mathbf{r}(i) \geq \delta |I|^2.$$

Hence (1.25) holds with $c_3 = \delta$.

Next, we show the "if" part, which is the more substantial implication. For the case of $\mu = \mathrm{Uniform}(\{0, 1\})$, Chatterjee, Diaconis, and Sly showed this implication in [CDS11, Lem. 4.1]. The implication for the general case follows from a minor modification of their argument. Below we provide a detailed argument for completeness.

Suppose there exist constants $c_1, c_2, c_3 > 0$ independent of $\mathbf{r}$ such that $c_1 \leq \mathbf{r}(i)/n \leq c_2$ for all i and (1.25) hold. Assume that $\mathbf{r}(1) \leq \cdots \leq \mathbf{r}(n)$ and $\boldsymbol{\alpha}(1) \leq \cdots \leq \boldsymbol{\alpha}(m)$. We will show that $\boldsymbol{\alpha}(n) \leq C$ for some constant $C = C(\mu, c_1, c_2, c_3) > 0$ independent of $\mathbf{r}$. Given this, we also have a lower bound $\boldsymbol{\alpha}(1) \geq \phi(c_1) - C$ since

$$c_1 n \leq \mathbf{r}(1) \leq \sum_{j=1}^{n} \psi'(\boldsymbol{\alpha}(1) + \boldsymbol{\alpha}(j)) \leq n\psi'(\boldsymbol{\alpha}(1) + \boldsymbol{\alpha}(n)) \leq n\psi'(\boldsymbol{\alpha}(1) + C).$$

Then $\psi'(2\phi(c_1) - C) \leq \psi'(\boldsymbol{\alpha} \oplus \boldsymbol{\beta}) \leq \phi'(2C)$. Hence $(\mathbf{r}, \mathbf{r})$ is δ-tame for $\delta > 0$ small enough so that $\delta < \psi'(2\phi(c_1) - C)$ and $\phi'(2C) \leq B - \delta$.

Without loss of generality, we can assume $\boldsymbol{\alpha}(n) \geq C_0$ for any constant $C_0 = C_0(\mu, c_1, c_2, c_3) > 0$. We first claim that a constant fraction of coordinates of $\boldsymbol{\alpha}$ is large:

$$|\{i : \boldsymbol{\alpha}(i) \geq \boldsymbol{\alpha}(n)/4\}| \geq c_1^2 n. \tag{7.16}$$

To see this, let $M := |\{i : \boldsymbol{\alpha}(i) > -\boldsymbol{\alpha}(n)/2\}|$. Note that $M \geq 1$ since $\boldsymbol{\alpha}(n) \geq C_0 > 0$. Since ψ' takes values in $(0, B)$, it follows that

$$c_2 n \geq \mathbf{r}(n) = \sum_{j=1}^{n} \psi'(\boldsymbol{\alpha}(n) + \boldsymbol{\alpha}(j)) \geq M\psi'(\boldsymbol{\alpha}(n)/2).$$

Assuming $C_0 > 2\phi(c_2)$, we can impose $\boldsymbol{\alpha}(n) > 2\phi(c_2)$. Then $\psi'(\boldsymbol{\alpha}(n)/2) > c_2$, so the above yields $M < n$. It follows that there exists an index j such that $\boldsymbol{\alpha}(j) \leq -\boldsymbol{\alpha}(n)/2$. In particular, $\boldsymbol{\alpha}(1) < 0$. Let

$$M_1 := \{i : i \neq j, \boldsymbol{\alpha}(i) < -\boldsymbol{\alpha}(1)/2\}.$$

Then using $\boldsymbol{\alpha}(1) \le -\boldsymbol{\alpha}(n)/2$, note that

$$c_1 n \le \mathbf{c}(1) \le M_1 \psi'(\boldsymbol{\alpha}(1)/2) + B(n - M_1) \le M_1 \psi'(-\boldsymbol{\alpha}(n)/4) + B(n - M_1).$$

Assuming $C_0 > -4\phi(\frac{Bc_1}{1+c_1})$, we can further impose $\boldsymbol{\alpha}(n) \ge -4\phi(\frac{Bc_1}{1+c_1})$ so that $\psi'(-\boldsymbol{\alpha}(n)/4) \le B\delta/(1+c_1)$. Hence

$$M_1 \le \frac{(B-\delta)n}{B - \psi'(-\boldsymbol{\alpha}(n)/4)} \le (1-c_1^2)n.$$

Hence $n - M_1 \ge c_1^2 n$, so there are at least $c_1^2 n$ indices i such that $\boldsymbol{\alpha}(i) \ge -\boldsymbol{\alpha}(1)/2 \ge \boldsymbol{\alpha}(n)/4$. This shows the claim (7.16).

Next, assume $C_0 > 17^2$ so that $h := \sqrt{\boldsymbol{\alpha}(n)} \ge 17$. For each integer k between 0 and $\frac{h}{16} - 1$, define

$$D_k := \left\{ i : -\frac{\boldsymbol{\alpha}(n)}{8} + kh \le \boldsymbol{\alpha}(i) < -\frac{\boldsymbol{\alpha}(n)}{8} + (k+1)h \right\}.$$

Since $D_0, D_1, \dots$ are disjoint, there exists an integer $1 \le k \le \frac{h}{16} - 1$ such that $|D_k| \le \frac{n}{(h/16)-1}$. Fix such an integer k and let

$$I := \left\{ i : \boldsymbol{\alpha}(i) \ge \frac{\boldsymbol{\alpha}(n)}{8} - \left(k + \frac{1}{2}\right) h \right\}.$$

Since $\boldsymbol{\alpha}(n) > 0$, we have $\frac{\boldsymbol{\alpha}(n)}{4} \ge \frac{\boldsymbol{\alpha}(n)}{8} - \left(k + \frac{1}{2}\right) h$, so it follows $|I| \ge c_1^2 n$ by the claim (7.16).

Note that since $k + 1 \le h/16$, we have $\boldsymbol{\alpha}(i) \ge \boldsymbol{\alpha}(n)/16$ for each $i \in I$. Hence get

$$B|I|^2 - \sum_{i,j \in I} z_{ij} \le (B - \psi'(\boldsymbol{\alpha}(n)/8))|I|^2. \tag{7.17}$$

Next, fix $j \notin I$. We consider three cases. First, suppose $\boldsymbol{\alpha}(j) \ge -\frac{\boldsymbol{\alpha}(n)}{8} + (k+1)h$. Then for each $i \in I$, we have $\boldsymbol{\alpha}(i) + \boldsymbol{\alpha}(j) \ge h/2$, so

$$\mathbf{r}(j) \wedge B|I| - \sum_{i \in I} z_{ij} \le B|I| - \sum_{i \in I} z_{ij} \le |I|(B - \psi'(h/2)).$$

Second, suppose $\boldsymbol{\alpha}(j) \ge -\frac{\boldsymbol{\alpha}(n)}{8} + kh$. Then for each $i \notin I$, we have $\boldsymbol{\alpha}(i) + \boldsymbol{\alpha}(j) \le -h/2$, so

$$\mathbf{r}(j) \wedge B|I| - \sum_{i \in I} z_{ij} \le \mathbf{r}(j) - \sum_{i \in I} z_{ij} = \sum_{i \notin I} z_{ij} \le n\psi'(-h/2).$$

Lastly, the third case is the one in which $j \notin I$ does not belong to either of the previous two cases. The set of all such j's is contained in the set D_k and for such j's, we bound $\mathbf{r}(j) \wedge B|I| - \sum_{i \in I} z_{ij} \le B|I| \le Bn$. Hence combining the three cases above with (7.17),

$$\sum_{j \notin I} \left(\mathbf{r}(j) \wedge B|I| \right) + B|I|^2 - \sum_{i \in I} \mathbf{r}(i) \le \left[(B - \psi'(h/2)) + \psi'(-h/2) + \frac{B}{(h/16)-1} + (B - \psi'(h^2/8)) \right] n^2.$$

According to (1.25), the left-hand side of the above inequality is bounded below by $c_3|I|^2 \ge c_3 c_1^2 n^2$. But the coefficient of n^2 on the right-hand side tends to zero as $\boldsymbol{\alpha}(n) = h^2 \to \infty$, which is a constracition. So there exists a constant $C = C(\mu, c_1, c_2, c_3) \ge C_0 > 0$ such that $h > C$ implies that the right-hand side above is at most $c\delta^2 n^2$. Then $\boldsymbol{\alpha}(n) \le C$, as desired. □

8. Proof of convergence of generalized Sinkhorn algorithm

In this section, we establish Theorem 1.25 on the linear convergence of the generalized Sinkhorn algorithm (1.26). The key difficulty is showing that the sequence of dual variables along the trajectory of the Sinkhorn algorithm stays bounded. In the special case of entropic optimal transport, a uniform bound on the norm of the dual variables (e.g., [MG20, Lem. 2.3]) is established relying heavily on the closed form of Sinkhorn iterates (2.6), which is enjoyed only for the special case of the Poisson base measure in our setting. In the lemma below, we establish that the generalized Sinkhorn update weakly contracts in the L^∞-norm.

Lemma 8.1 (L^∞-monotonicity of the Sinkhorn iterates)**.** *Suppose $\mathcal{T}(\mathbf{r},\mathbf{c})\cap(A,B)^{m\times n}$ is non-empty. Let $(\boldsymbol{\alpha}_k,\boldsymbol{\beta}_k)$, $k\ge 0$ denote the iterates produced by the Sinkhorn algorithm (1.26). Let $(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})$ be an arbitrary MLE for the margin $(\mathbf{r},\mathbf{c})$.*

(i) *For each $k\ge 0$, $\|(\boldsymbol{\alpha}_{k+1},\boldsymbol{\beta}_{k+1})-(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})\|_\infty \le \|(\boldsymbol{\alpha}_k,\boldsymbol{\beta}_k)-(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})\|_\infty \le \|\boldsymbol{\alpha}_0-\hat{\boldsymbol{\alpha}}\|_\infty$.*

(ii) *Suppose $(\mathbf{r},\mathbf{c})$ is δ-tame, ψ'' is increasing, and $\boldsymbol{\alpha}_0=\mathbf{0}$. Then $(\boldsymbol{\alpha}_k,\boldsymbol{\beta}_k)$ is δ-tame for all $k\ge 1$.*

Proof. By permuting the rows and columns if necessary, we may assume that $\mathbf{r}(1)\le\cdots\le\mathbf{r}(m)$ and $\mathbf{c}(1)\le\cdots\le\mathbf{c}(n)$. Since ψ' is increasing, it follows that $\hat{\boldsymbol{\alpha}}(1)\le\cdots\le\hat{\boldsymbol{\alpha}}(m)$ and $\hat{\boldsymbol{\beta}}(1)\le\cdots\le\hat{\boldsymbol{\beta}}(n)$. Fix $(\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathbb{R}^m\times\mathbb{R}^n$. Let $\boldsymbol{\beta}\mapsto\xi(\boldsymbol{\beta})=:\boldsymbol{\alpha}'$ denote the Sinkhorn update for the first dual variable given the second one $\boldsymbol{\beta}$. Since ψ' is increasing and since $\mathbf{r}(i)=\sum_j\psi'(\boldsymbol{\alpha}'(i)+\boldsymbol{\beta}(j))$ for all i, it follows that $\boldsymbol{\alpha}'$ has increasing coordinates. We would like to compute the Jacobian of this map. To do so, define the function $F:\mathbb{R}^m\times\mathbb{R}^n\to\mathbb{R}^m$ as

$$F(\boldsymbol{\alpha},\boldsymbol{\beta}):=\left(\mathbf{r}(1)-\sum_j\psi'(\boldsymbol{\alpha}(1)+\boldsymbol{\beta}(j)),\dots,\mathbf{r}(m)-\sum_j\psi'(\boldsymbol{\alpha}(m)+\boldsymbol{\beta}(j))\right).$$

Then $\boldsymbol{\alpha}'=\xi(\boldsymbol{\beta})$ is the unique zero of the equation $F(\cdot,\boldsymbol{\beta})=\mathbf{0}$. Let $E=E(\boldsymbol{\alpha}',\boldsymbol{\beta})$ be the $m\times n$ matrix whose (i,j) entry is $-\psi''(\boldsymbol{\alpha}'(i)+\boldsymbol{\beta}(j))$ and let $E_{i\bullet}$ denote the ith row sum of E for $i=1,\dots,m$. Then the Jacobian of F with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively, are given by

$$[J_{F;\boldsymbol{\alpha}}(\boldsymbol{\alpha}',\boldsymbol{\beta})]_{m\times m}=\operatorname{diag}(E_{1\bullet},\dots,E_{m\bullet}),\qquad [J_{F;\boldsymbol{\beta}}(\boldsymbol{\alpha}',\boldsymbol{\beta})]_{m\times n}=E.$$

The first Jacobian matrix above is always invertible since $\psi''>0$ on the domain. Hence by the implicit function theorem,

$$[J_{\boldsymbol{\alpha}';\boldsymbol{\beta}}]_{m\times n}=-[J_{F;\boldsymbol{\alpha}}(\boldsymbol{\alpha}',\boldsymbol{\beta})]^{-1}_{m\times m}[J_{F;\boldsymbol{\beta}}(\boldsymbol{\alpha}',\boldsymbol{\beta})]_{m\times n}=-\left[E(\boldsymbol{\alpha}',\boldsymbol{\beta})_{ij}/E(\boldsymbol{\alpha}',\boldsymbol{\beta})_{i\bullet}\right]_{m\times n}.$$

Importantly, we observe that $-J_{\boldsymbol{\alpha}';\boldsymbol{\beta}}$ is a row-stochastic matrix.

Now note that $\xi'(\hat{\boldsymbol{\beta}})=\hat{\boldsymbol{\alpha}}$. Let $\gamma(s)=(1-s)\boldsymbol{\beta}+s\hat{\boldsymbol{\beta}}$ denote the linear interpolation between $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$. Then denoting $P_s:=-J_{\xi(\gamma(s));\gamma(s)}$, we have

$$\boldsymbol{\alpha}'-\hat{\boldsymbol{\alpha}}=\xi(\boldsymbol{\beta})-\xi(\hat{\boldsymbol{\beta}})=\underbrace{\left[\int_0^1 P_s\,ds\right]}_{=:P}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}).\tag{8.1}$$

The matrix P defined above is row-stochastic since every intermediate negative Jacobian matrix is row-stochastic by the earlier observation. In particular, since $\|P\|_\infty=1$, this yields

$$\|\boldsymbol{\alpha}'-\hat{\boldsymbol{\alpha}}\|_\infty\le\|P\|_\infty\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_\infty=\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_\infty.$$

By a symmetric argument, it also holds $\|\boldsymbol{\beta}'-\hat{\boldsymbol{\beta}}\|_\infty\le\|\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}\|_\infty$, where $\boldsymbol{\beta}'$ denotes the output of the Sinkhorn update for the second dual variable given the first dual variable $\boldsymbol{\alpha}$. It then follows that, for all $k\ge 0$,

$$\|\boldsymbol{\alpha}_{k+1}-\hat{\boldsymbol{\alpha}}\|_\infty\le\|\boldsymbol{\beta}_{k+1}-\hat{\boldsymbol{\beta}}\|_\infty\le\|\boldsymbol{\alpha}_k-\hat{\boldsymbol{\alpha}}\|_\infty\le\|\boldsymbol{\beta}_k-\hat{\boldsymbol{\beta}}\|_\infty.$$

By induction, from the above, we can deduce **(i)**.

Next, we show **(ii)**. Assume ψ'' is increasing and let $\boldsymbol{\alpha}_0 = \mathbf{0}$. We may shift the MLE (which we will still denote as $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$) if necessary so that $\hat{\boldsymbol{\alpha}} \le \boldsymbol{\alpha}_0$ entrywise. We claim the following:

(a) $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_k$ has increasing nonnegative coordinates for $k \ge 1$.

(b) $\boldsymbol{\alpha}_k - \hat{\boldsymbol{\alpha}}$ has decreasing nonneative coordinates for $k \ge 0$.

Indeed, note that (b) holds for $k = 0$ by the hypothesis. We will next show that (b) holds for $k \ge 1$ if (a) holds for the same k. By (8.1) with $\boldsymbol{\beta} = \boldsymbol{\beta}_k$ and since $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_k$ has nonnegative coordinates,

$$\boldsymbol{\alpha}_k(i) - \hat{\boldsymbol{\alpha}}(i) = \int_0^1 P_s[i,:](\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_k)\, ds \ge 0 \qquad \text{for all } i. \tag{8.2}$$

Next, we argue that $\boldsymbol{\alpha}_k - \hat{\boldsymbol{\alpha}}$ has decreasing coordintes. Denoting $\boldsymbol{\beta}^s := (1-s)\boldsymbol{\beta}_k + s\hat{\boldsymbol{\beta}}$, the corresponding dual variable $\boldsymbol{\alpha}^s := \xi(\boldsymbol{\beta}^s)$ has increasing coordinates by the observation in the first paragraph. Since ψ'' is increasing, it follows that P_s is the row-normalization of an $m \times n$ matrix $E(\boldsymbol{\alpha}^s, \boldsymbol{\beta}^s)$ with coordinates increasing along both row and column indices. So the rows $P_s[i,:]$ and $P_s[i',:]$ for $i < i'$ are probability distributions over $[n]$ where the former assigns smaller weights on lesser indices than the latter. More precisely, there exists an index j^* such that $P_s[i,j] \le P_s[i',j]$ for $j \le j^*$ and the inequality reverses for $j > j^*$. By the induction hypothesis $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_k$ has increasing coordinates, one can easily check that $P_s[i,:](\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_k) \ge P_s[i',:](\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_k)$. Integrating in s, (8.2) implies that $\boldsymbol{\alpha}_k - \hat{\boldsymbol{\alpha}}$ has decreasing coordinates.

By a symmetric argument, one can show that (a) for $k+1$ holds if (b) for k holds. Thus by induction, this shows (a) and (b) above.

Now we are ready to finish the proof of **(ii)**. Observe that

$$\begin{aligned}\boldsymbol{\alpha}_k(m) + \boldsymbol{\beta}_k(n) &= \hat{\boldsymbol{\alpha}}(m) + [\boldsymbol{\alpha}_k(m) - \hat{\boldsymbol{\alpha}}(m)] + \hat{\boldsymbol{\beta}}(n) + \left[\boldsymbol{\beta}_k(n) - \hat{\boldsymbol{\beta}}(n)\right] \\ &= \left(\hat{\boldsymbol{\alpha}}(m) + \hat{\boldsymbol{\beta}}(n)\right) + \underbrace{P[m,:](\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_k) - (\hat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}_k(n))}_{\le 0} \le \hat{\boldsymbol{\alpha}}(m) + \hat{\boldsymbol{\beta}}(n) \le \phi(B_\delta),\end{aligned}$$

where the first inequality follows since $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_k$ has increasing (claim (a) above) coordinates and the second inequality follows since $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ is δ-tame. Similarly,

$$\begin{aligned}\boldsymbol{\alpha}_k(1) + \boldsymbol{\beta}_k(1) &= \hat{\boldsymbol{\alpha}}(1) + [\boldsymbol{\alpha}_k(1) - \hat{\boldsymbol{\alpha}}(1)] + \hat{\boldsymbol{\beta}}(1) + \left[\boldsymbol{\beta}_k(1) - \hat{\boldsymbol{\beta}}(1)\right] \\ &= \left(\hat{\boldsymbol{\alpha}}(1) + \hat{\boldsymbol{\beta}}(1)\right) + \underbrace{P[:,1](\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_k) - (\hat{\boldsymbol{\beta}}(1) - \boldsymbol{\beta}_k(1))}_{\ge 0} \ge \hat{\boldsymbol{\alpha}}(1) + \hat{\boldsymbol{\beta}}(1) \ge \phi(A_\delta).\end{aligned}$$

Since both $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$ have increasing coordinates, this is enough to conclude **(ii)**. □

Proof of Theorem 1.25. We first claim the following: (1.27) holds if all Sinkhorn iterates $(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k)$ for $k \ge k_0$ as well as the MLE $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ are ε-tame. Before proving this claim, we will first deduce parts **(i)**-**(iii)** from this claim. First, we remark that there are well-known results from the optimization literature that are directly applicable to the generalized Sinkhorn algorithm (1.26). Since each sub-problem in (1.26) has a unique solution due to strict concavity of the block-restricted dual objective, asymptotic convergence to the critical point of (1.26) follows from a general result for alternating maximization (e.g., [Ber97, Prop. 2.7.1]). Every critical point of the dual objective is an MLE, which is a global optimum by Theorem 1.7. Thus it follows that $\boldsymbol{\alpha}_k \oplus \boldsymbol{\beta}_k \to \boldsymbol{\alpha}^* \oplus \boldsymbol{\beta}^*$ as $k \to \infty$ entrywise. In particular, if we choose $\varepsilon > 0$ small enough so that $(\mathbf{r}, \mathbf{c})$ is ε-tame (i.e., $\phi(A_\varepsilon) \le \boldsymbol{\alpha}^* \oplus \boldsymbol{\beta}^* \le \phi(B_\varepsilon)$), then there exists $k_0 \ge 1$ such that $\phi(A_{\varepsilon/2}) \le \boldsymbol{\alpha}_k \oplus \boldsymbol{\beta}_k \le \phi(B_{\varepsilon/2})$ for all $k \ge k_0$. Then **(i)** follows from the claim. For **(ii)**, the hypothesis of the claim is directly justified by Lemma 8.1 **(ii)**. For **(iii)**, the hypothesis and Lemma 8.1 **(i)** impliy that entrywise,

$$\boldsymbol{\alpha}_k \oplus \boldsymbol{\beta}_k \le \phi(B_\varepsilon) - 2\|\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}^*\|_\infty + \|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}^*\|_\infty + \|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*\|_\infty \le \phi(B_\varepsilon).$$

The lower bound follows similarly. Thus $(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k)$ is ε-tame for all $k \geq 0$.

It now suffices to show the claim. Our analysis for this is inspired by the proof of linear convergence of the Sinkhorn algorithm for entropic optimal transport due to Carlier [Car22]. For simplicity denote $F := -g^{\mathbf{r},\mathbf{c}}$ (see (1.6)). Consider the following centered Sinkhorn iterates $(\tilde{\boldsymbol{\alpha}}_k, \tilde{\boldsymbol{\beta}}_k)$ for $k \geq 1$ where $(\tilde{\boldsymbol{\alpha}}_0, \tilde{\boldsymbol{\beta}}_0) = (\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ and for $k \geq 1$, $(\tilde{\boldsymbol{\alpha}}_k, \tilde{\boldsymbol{\beta}}_k)$ is obtained from $\tilde{\boldsymbol{\beta}}_{k-1}$ by the same Sinkhorn update in (1.26) but followed by centering (adding and subtracting the same constants to $\tilde{\boldsymbol{\alpha}}_k$ and $\tilde{\boldsymbol{\beta}}_k$, respectively) so that $\sum_i \tilde{\boldsymbol{\alpha}}_k(i) = 0$. By an induction, it is easy to verify

$$\boldsymbol{\alpha}_k \oplus \boldsymbol{\beta}_k = \tilde{\boldsymbol{\alpha}}_k \oplus \tilde{\boldsymbol{\beta}}_k \qquad \text{for all } k \geq 0.$$

We also let $(\tilde{\boldsymbol{\alpha}}^*, \tilde{\boldsymbol{\beta}}^*)$ denote the standard MLE for $(\mathbf{r}, \mathbf{c})$. Note that $\sum_i \tilde{\boldsymbol{\alpha}}^*(i) = 0$ and $\boldsymbol{\alpha}^* \oplus \boldsymbol{\beta}^* = \tilde{\boldsymbol{\alpha}}^* \oplus \tilde{\boldsymbol{\beta}}^*$. In particular, $F(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k) = F(\tilde{\boldsymbol{\alpha}}_k, \tilde{\boldsymbol{\beta}}_k)$ for all $k \geq 0$ and $F(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = F(\tilde{\boldsymbol{\alpha}}^*, \tilde{\boldsymbol{\beta}}^*)$. Thus it is enough to show the claim for the centered iterates $(\tilde{\boldsymbol{\alpha}}_k, \tilde{\boldsymbol{\beta}}_k)$ and standard MLE $(\tilde{\boldsymbol{\alpha}}^*, \tilde{\boldsymbol{\beta}}^*)$. We will omit the tilde notation in the rest of the proof.

Denote $\sigma_\ell^2 = \sigma_\ell^2(\varepsilon)$ for $\ell = 1, 2$, which are defined in the statement. Note that

$$\nabla^2_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \operatorname{diag}\left(\sum_j \psi''(\boldsymbol{\alpha}(i) + \boldsymbol{\beta}(j); i)\right) \quad \text{and} \quad \nabla^2_{\boldsymbol{\beta}} F(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \operatorname{diag}\left(\sum_i \psi''(\boldsymbol{\alpha}(i) + \boldsymbol{\beta}(j); j)\right).$$

If $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $(\boldsymbol{\alpha}', \boldsymbol{\beta}')$ are both ε-tame, then so is their convex combination. Hence $F(\boldsymbol{\alpha}, \cdot)$ is $m\sigma_1^2$-strongly convex and $m\sigma_2^2$-smooth (i.e., $\nabla_{\boldsymbol{\beta}} F(\boldsymbol{\alpha}, \cdot)$ is $m\sigma_2^2$-Lipschitz continuous) on the secant line between $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $(\boldsymbol{\alpha}', \boldsymbol{\beta})$. By the first-order optimality of $\boldsymbol{\beta}_{k+1}$, $\nabla_{\boldsymbol{\beta}} F(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_{k+1}) = \mathbf{0}$. Hence we deduce the second-order growth property

$$F(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k) - F(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_{k+1}) \geq \frac{n\sigma_1^2}{2} \|\boldsymbol{\alpha}_{k+1} - \boldsymbol{\alpha}_k\|^2.$$

Similarly, using $\nabla_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha}_{k+1}, \boldsymbol{\beta}_{k+1}) = \mathbf{0}$, we get

$$F(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_{k+1}) - F(\boldsymbol{\alpha}_{k+1}, \boldsymbol{\beta}_{k+1}) \geq \frac{m\sigma_1^2}{2} \|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k\|^2.$$

Combining the two inequalities above and recalling $\Delta_k = F(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k) - F(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$, we obtain

$$\Delta_k - \Delta_{k+1} \geq \frac{n\sigma_1^2}{2} \|\boldsymbol{\alpha}_{k+1} - \boldsymbol{\alpha}_k\|^2 + \frac{m\sigma_1^2}{2} \|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k\|^2. \tag{8.3}$$

Next, note that ψ is σ_1^2-strongly convex and σ_2^2-smooth on $[\phi(A_\varepsilon), \phi(B_\varepsilon)]$. In particular, for each x, y in that interval,

$$\psi(x) - \psi(y) \geq \psi'(y)(x - y) + \frac{\sigma_1^2}{2}(x - y)^2. \tag{8.4}$$

Then ε-tameness, (8.4), and $\sum_i \boldsymbol{\alpha}_k = 0 = \sum_i \boldsymbol{\alpha}^*(i)$ give

$$\begin{aligned} &\sum_{i,j} \left[\psi(\boldsymbol{\alpha}^*(i) + \boldsymbol{\beta}^*(j)) - \psi(\boldsymbol{\alpha}_k(i) + \boldsymbol{\beta}_k(j))\right] \\ &\geq \sum_{i,j} \left[\psi'(\boldsymbol{\alpha}_k(i) + \boldsymbol{\beta}_k(j))(\boldsymbol{\alpha}^*(i) + \boldsymbol{\beta}^*(j) - \boldsymbol{\alpha}_k(i) - \boldsymbol{\beta}_k(j))\right] + \frac{\sigma_1^2}{2} \underbrace{\|(\boldsymbol{\alpha}^* \oplus \boldsymbol{\beta}^*) - (\boldsymbol{\alpha}_k \oplus \boldsymbol{\beta}_k)\|_F^2}_{= n\|\boldsymbol{\alpha}^* - \boldsymbol{\alpha}\|^2 + m\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|^2} \end{aligned} \tag{8.5}$$

Also, recall that

$$\nabla_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \left(\mathbf{r}(i) - \sum_j \psi'(\boldsymbol{\alpha}(i) + \boldsymbol{\beta}(j)); i\right) \quad \text{and} \quad \nabla_{\boldsymbol{\beta}} F(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \left(\mathbf{c}(j) - \sum_i \psi'(\boldsymbol{\alpha}(i) + \boldsymbol{\beta}(j)); j\right).$$

Hence we have

$$\langle \nabla_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k), \boldsymbol{\alpha}^* - \boldsymbol{\alpha}_k \rangle + \langle \nabla_{\boldsymbol{\beta}} F(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k), \boldsymbol{\beta}^* - \boldsymbol{\beta}_k \rangle = \langle (\mathbf{r}, \mathbf{c}), (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) - (\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k) \rangle - I_1, \tag{8.6}$$

where I_1 denotes the first term in (8.5).

Then we can deduce the following strong-convexity-type inequality

$$\begin{aligned}
-\Delta_k &\overset{(a)}{=} \langle (\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k) - (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*), (\mathbf{r}, \mathbf{c}) \rangle + \sum_{i,j} \left[\psi(\boldsymbol{\alpha}^*(i) + \boldsymbol{\beta}^*(j)) - \psi(\boldsymbol{\alpha}_k(i) + \boldsymbol{\beta}_k(j)) \right] \\
&\overset{(b)}{\geq} \langle \underbrace{\nabla_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k)}_{=\mathbf{0}}, \boldsymbol{\alpha}^* - \boldsymbol{\alpha}_k \rangle + \langle \nabla_{\boldsymbol{\beta}} F(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k), \boldsymbol{\beta}^* - \boldsymbol{\beta}_k \rangle + \frac{\sigma_1^2}{2} \left[n \|\boldsymbol{\alpha}^* - \boldsymbol{\alpha}_k\|^2 + m \|\boldsymbol{\beta}^* - \boldsymbol{\beta}_k\|^2 \right] \\
&\overset{(c)}{\geq} -\frac{1}{2m\sigma_1^2} \|\nabla_{\boldsymbol{\beta}} F(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k) - \nabla_{\boldsymbol{\beta}} F(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_{k+1})\|^2 \\
&\overset{(d)}{\geq} -\frac{\sigma_2^4}{2\sigma_1^2} \left(n \|\boldsymbol{\alpha}_{k+1} - \boldsymbol{\alpha}_k\|^2 + m \|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k\|^2 \right),
\end{aligned} \tag{8.7}$$

where (a) follows from the definition of $g^{\mathbf{r},\mathbf{c}}$ in (1.6), (b) follows from combining (8.5) and (8.6). For (c), we used $\nabla_{\boldsymbol{\beta}} F(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_{k+1}) = \mathbf{0}$ and Young's inequality $ab \leq \frac{a^2}{2\lambda} + \frac{\lambda b^2}{2}$ with $a = \|\nabla_{\boldsymbol{\beta}} F(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k)\|$, $b = \|\boldsymbol{\beta}^* - \boldsymbol{\beta}_k\|$, and $\lambda = m\sigma_1^2$, which give

$$\begin{aligned}
\langle \nabla_{\boldsymbol{\beta}} F(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k), \boldsymbol{\beta}^* - \boldsymbol{\beta}_k \rangle &\geq -\|\nabla_{\boldsymbol{\beta}} F(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k)\| \cdot \|\boldsymbol{\beta}^* - \boldsymbol{\beta}_k\| \\
&\geq -\frac{1}{2m\sigma_1^2} \|\nabla_{\boldsymbol{\beta}} F(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k) - \nabla_{\boldsymbol{\beta}} F(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_{k+1})\|^2 - \frac{m\sigma_1^2}{2} \|\boldsymbol{\beta}^* - \boldsymbol{\beta}_k\|^2.
\end{aligned}$$

Lastly, (d) follows from the Lipschitz continuity of $\nabla_{\boldsymbol{\beta}} F$ and including an additional nonpositive term for the lower bound.

Combining (8.7) with (8.3), we get

$$\Delta_k \leq (\sigma_1/\sigma_2)^4 (\Delta_k - \Delta_{k+1}).$$

Rearranging, this yields $\Delta_{k+1} \leq \left(1 - (\sigma_1/\sigma_2)^4\right) \Delta_k$. This shows the second inequality in (1.27).

To show the first inequality in (1.27), we switch $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ and $(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k)$ in (8.5) to get

$$\begin{aligned}
&\sum_{i,j} \left[\psi(\boldsymbol{\alpha}_k(i) + \boldsymbol{\beta}_k(j)) - \psi(\boldsymbol{\alpha}^*(i) + \boldsymbol{\beta}^*(j)) \right] \\
&\geq \underbrace{\sum_{i,j} \left[\psi'(\boldsymbol{\alpha}^*(i) + \boldsymbol{\beta}^*(j)) (\boldsymbol{\alpha}_k(i) + \boldsymbol{\beta}_k(j) - \boldsymbol{\alpha}^*(i) - \boldsymbol{\beta}^*(j)) \right]}_{= \langle (\mathbf{r}, \mathbf{c}), (\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k) - (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \rangle} + \frac{\sigma_1^2}{2} \|(\boldsymbol{\alpha}^* \oplus \boldsymbol{\beta}^*) - (\boldsymbol{\alpha}_k \oplus \boldsymbol{\beta}_k)\|_F^2,
\end{aligned}$$

where for the summation we used $\nabla F(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = \mathbf{0}$ and (8.6) with $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ and $(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k)$ switched. Rearranging terms then gives

$$\Delta_k \geq \frac{\sigma_1^2}{2} \|(\boldsymbol{\alpha}^* \oplus \boldsymbol{\beta}^*) - (\boldsymbol{\alpha}_k \oplus \boldsymbol{\beta}_k)\|_F^2.$$

Combining the above we obtain (1.27) as claimed. □

9. Proof of applications

In this section, we prove all results stated in Section 1.7.

Proof of Theorem 1.27. We first show that there exists a constant $C = C(\mu,\delta) > 0$ such that for ν-almost all margins $(\mathbf{r},\mathbf{c})$ with an MLE $(\boldsymbol{\alpha},\boldsymbol{\beta})$, the following holds for each $t \geq 0$:

$$d_{TV}\left(\widetilde{\xi}_{I,J}, \widetilde{\mu}_{\boldsymbol{\alpha}(I)\oplus\boldsymbol{\beta}(J)}\right) \leq t + \left[\exp(C\rho)\,\mathbb{P}\left(Y \in \mathscr{T}_\rho(\mathbf{r},\mathbf{c})\right)^{-1} \wedge \sup_{\bar{\mathbf{y}}} p_{\bar{\mathbf{y}}}(\mathbf{r},\mathbf{c})\right] \exp\left(-2t^2|I\times J|\right), \tag{9.1}$$

where $p_{\bar{\mathbf{y}}}(\mathbf{r},\mathbf{c})$ is defined in Theorem 1.11. To see this, fix a measurable set $D \subseteq \mathbb{R}$ and for each matrix $W \in \mathbb{R}^{m\times n}$, denote

$$S_D(W) := \frac{1}{|I\times J|}\sum_{(i,j)\in I\times J} \mathbf{1}(W_{ij} \in D).$$

Then $\widetilde{\mu}_{I,J}(D) = \mathbb{E}[S_D(X)]$ and $\widetilde{\xi}_{\boldsymbol{\alpha}(I)\oplus\boldsymbol{\beta}(J)}(D) = \mathbb{E}[S_D(Y)]$. Denoting the expression in the bracket in (9.1) by K and using Theorems 1.10 and 1.11,

$$\begin{aligned}
|\widetilde{\xi}_{I,J}(D) - \widetilde{\mu}_{\boldsymbol{\alpha}(I)\oplus\boldsymbol{\beta}(J)}(D)| &= \mathbb{E}\left[|S_D(X) - \mathbb{E}[S_D(Y)]|\right] \\
&\leq t\,\mathbb{P}\left(|S_D(X) - \mathbb{E}[S_D(Y)]| \leq t\right) + \mathbb{P}\left(|S_D(X) - \mathbb{E}[S_D(Y)]| > t\right) \\
&\leq t + K\,\mathbb{P}\left(|S_D(Y) - \mathbb{E}[S_D(Y)]| > t\right).
\end{aligned}$$

By Hoeffding's inequality,

$$\mathbb{P}\left(|S_D(Y) - \mathbb{E}[S_D(Y)]| > t\right) \leq 2\exp\left(-2t^2|I|\cdot|J|\right).$$

This shows (9.1).

Note that K is at most $\exp(F_\mu(\mathbf{r},\mathbf{c}))$ due to Theorems 1.10, 1.14, and 1.13. Then the bound (1.29) follows from (9.1) by choosing $t = O(\sqrt{F_\mu(\mathbf{r},\mathbf{c})/|I\times J|})$. □

Proof of Theorem 1.29. By Theorem 1.27, triangle inequality, and the convexity of the TV distance, it suffices to show that

$$\mathbb{E}\left[d_{TV}\left(\mu_{\overline{\boldsymbol{\alpha}}_m(U)+\overline{\boldsymbol{\beta}}_n(V)}, \mu_{\boldsymbol{\alpha}(U)+\boldsymbol{\beta}(V)}\right)^4\right] = O\left(\|(\mathbf{r},\mathbf{c}) - (\bar{\mathbf{r}}_m,\bar{\mathbf{c}}_n)\|_1\right).$$

By Pinsker's inequality, for each $\theta,\theta' \in [\phi(A_\delta),\phi(B_\delta)]$, for $C_\delta := \sup_{\phi(A_\delta)\leq w\leq\phi(B_\delta))}|\psi'(w)|$,

$$2d_{TV}(\mu_\theta,\mu_{\theta'})^2 \leq D(\mu_\theta\,\|\,\mu_{\theta'}) = (\theta-\theta')\psi'(\theta) - \psi(\theta) + \psi(\theta') \leq C_\delta|\theta-\theta'|.$$

It follows that

$$4d_{TV}\left(\mu_{\bar{\boldsymbol{\alpha}}_m(U)+\bar{\boldsymbol{\beta}}_n(V)}, \mu_{\boldsymbol{\alpha}(U)+\boldsymbol{\beta}(V)}\right)^4 \leq C_\delta^2|\bar{\boldsymbol{\alpha}}_m(U) + \bar{\boldsymbol{\beta}}_n(V) - \boldsymbol{\alpha}(U) + \boldsymbol{\beta}(V)|^2.$$

Using the fact that $\mathbb{E}[\bar{\boldsymbol{\alpha}}_m] = \mathbb{E}[\boldsymbol{\alpha}] = 0$ and Theorem 1.19, the expectation of the right-hand side is at most

$$C_\delta^2\left(\|\bar{\boldsymbol{\alpha}}_m - \boldsymbol{\alpha}\|_2^2 + \|\bar{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\|_2^2\right) = O(\|(\bar{\mathbf{r}}_m,\bar{\mathbf{c}}_n) - (\mathbf{r},\mathbf{c})\|_1).$$

□

Proof of Corollary 1.30. Recall from Ex. 1.9 that the sequence of k-cloning of $(\mathbf{r}_0,\mathbf{c}_0)$ is uniformly δ-tame for some δ depending only on $(\mathbf{r}_0,\mathbf{c}_0)$. Specifically, if $(\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)$ is an MLE for $(\mathbf{r}_0,\mathbf{c}_0)$, then its concatenation $(\boldsymbol{\alpha}_0\otimes\mathbf{1}_k, \boldsymbol{\beta}_0\otimes\mathbf{1}_k)$ is an MLE for the k-cloning of $(\mathbf{r}_0,\mathbf{c}_0)$. By symmetry, the entries in the first $k\times k$ submatrix of X have the same distribution. The same holds for $Y \sim \mu_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}$ since the first $k\times k$ submatrix of $\boldsymbol{\alpha}\oplus\boldsymbol{\beta}$ has constant entries, namely, $\boldsymbol{\alpha}_0(1) + \boldsymbol{\beta}_0(1)$. Then the assertion follows from Theorem 1.27. □

Next, we establish the scaling limit of X in the cut norm stated in Theorem 1.19. We begin with a simple observation of the cut norm.

Proposition 9.1. *Let $A \in \mathbb{R}^{m\times n}$ and let W_A denote the corresponding kernel. Then*

$$mn\|W_A\|_\square = \sup_{\mathbf{x}\in\{0,1\}^m, \mathbf{y}\in\{0,1\}^n} |\mathbf{x}^\top A\mathbf{y}|.$$

Proof. Write $I_i = ((i-1)/m, i/m]$ for $i = 1,\dots,m$ and $J_j = ((j-1)/n, j/n]$ for $j = 1,\dots,n$. We claim that the supremum in $\|W_A\|_\square$ in (1.30) is attained over a measurable rectangle $S\times T \subseteq [0,1]^2$, where S is the disjoint union of some of the intervals I_i's and T is the disjoint union of some of the intervals J_j's. This will be enough to conclude.

Now we show the claim. Since W_A takes the constant value A_{ij} over the rectangle $I_i \times J_j$,

$$\int_{S\times T} W_A\,dx\,dy = \sum_i |S\cap I_i|\left(\sum_j |T\cap J_j| A_{ij}\right).$$

If the quantity in the parenthesis on the right-hand side is nonnegative, then replacing $S\cap I_i$ with I_i can only increase the total value; otherwise, replace $S\cap I_i$ with $\emptyset$. In this way, we find a new set $\bar{S}$ that is the disjoint union of some intervals I_i's such that

$$\int_{S\times T} W_A\,dx\,dy \le \int_{\bar{S}\times T} W_A\,dx\,dy.$$

Similarly, we can find a disjoint union $\bar{T}$ of some of J_j's such that

$$\int_{S\times T} W_A\,dx\,dy \le \int_{\bar{S}\times T} W_A\,dx\,dy \le \int_{\bar{S}\times\bar{T}} W_A\,dx\,dy.$$

Repeating the same argument with $-W_A$, we can deduce the claim. □

We also record a simple consequence of Lemma 5.1 that the $(\boldsymbol{\alpha},\boldsymbol{\beta})$-model concentrates around its expectation in cut norm.

Lemma 9.2 (Concentration of the $(\boldsymbol{\alpha},\boldsymbol{\beta})$-model in cut norm)**.** *Let $(\mathbf{r},\mathbf{c})$ be an $m\times n$ δ-tame margin and $Y \sim \mu_{\boldsymbol{\alpha}\oplus\boldsymbol{\beta}}$, where $(\boldsymbol{\alpha},\boldsymbol{\beta})$ is an MLE for the margin $(\mathbf{r},\mathbf{c})$. Define positive constants L_1 and L_2 as in Lemma 5.1. Denote $\tilde{Y} = Y - \mathbb{E}[Y]$. For each $s \in [0, L_1L_2]$,*

$$\mathbb{P}\left(\|W_{\tilde{Y}}\|_\square \ge s\right) \le 2^{m+n+1}\exp\left(-\frac{s^2mn}{2L_2}\right),$$

$$\mathbb{P}\left(\|Y - \Gamma_{\mathbf{r},\mathbf{c}}(\overline{Y})\|_1 \ge smn\right) \le 3^{m+n}\exp\left(-\frac{s^2mn}{18L_2}\right). \tag{9.2}$$

Proof. Using Proposition 9.1, union bound, and Lemma 5.1 with $t = s\frac{mn}{\|\mathbf{x}\|^2\|\mathbf{y}\|^2}$ (see (5.5)),

$$\mathbb{P}\left(\|W_{\tilde{Y}}\|_\square \ge s\right) \le \sum_{\mathbf{x}\in\{0,1\}^m\,\mathbf{y}\in\{0,1\}^n} \mathbb{P}\left(|\mathbf{x}^\top\tilde{Y}\mathbf{y}| \ge smn\right) \le 2^{m+n+1}\exp\left(-\frac{s^2mn}{2L_2}\right).$$

To show (9.2), note that

$$\begin{aligned}\|Y - \Gamma_{\mathbf{r},\mathbf{c}}(\overline{Y})\|_1 &= \|r(Y)-\mathbf{r}\|_1 + \|c(Y)-\mathbf{c}\|_1 + |Y_{mn} - \overline{Y}_{\bullet\bullet} + \sum_{i=1}^{m-1}\mathbf{r}(i) - \mathbf{c}(n)| \\ &\le \|r(Y)-\mathbf{r}\|_1 + \|c(Y)-\mathbf{c}\|_1 + |Z_{\bullet\bullet} - Y_{\bullet\bullet}|\end{aligned}$$

with $Z = \mathbb{E}[Y]$. Hence by using the observation (5.6) and Lemma 5.1,

$$\mathbb{P}\left(\|Y - \Gamma_{\mathbf{r},\mathbf{c}}(\overline{Y})\|_1 \ge smn\right) \le 3\mathbb{P}\left(\max_{\mathbf{x}\in\{-1,0,1\}^m,\mathbf{y}\in\{-1,0,1\}^n} \mathbf{x}^\top\tilde{Y}\mathbf{y} \ge \frac{smn}{3}\right) \le 3^{m+n}\exp\left(-\frac{s^2mn}{18L_2}\right).$$

□

Proof of Theorem 1.19. Let $(\boldsymbol{\alpha}_m, \boldsymbol{\beta}_n)$ be the standard MLE for $(\mathbf{r}_m, \mathbf{c}_n)$ and let $Y \sim \mu_{\boldsymbol{\alpha}_m \oplus \boldsymbol{\beta}_n}$. By Lemma 9.2, there exists a constant $C = C(\mu, \delta) > 0$ such that

$$\mathbb{P}(\|W_Y - W_{\mathbb{E}[Y]}\|_\square \geq t) \leq \exp\left((m+n+1)\log 2 - \frac{t^2 mn}{C}\right).$$

By Proposition 9.1 and Lemma 9.2, whenever $t \in [0, L_1 L_2]$,

$$\mathbb{P}\left(\|W_Y - W_{\Gamma_{\mathbf{r}_m,\mathbf{c}_n}(\overline{Y})}\|_1 \geq t\right) \leq 3^{m+n} \exp\left(-\frac{t^2 mn}{18 L_2}\right).$$

Combining the above by a union bound,

$$\mathbb{P}\left(\|W_{\Gamma_{\mathbf{r}_m,\mathbf{c}_n}(\overline{Y})} - W_{\mathbb{E}[Y]}\|_\square \geq t\right) \leq \exp\left((m+n+1)\log 3 - \frac{t^2 mn}{C'}\right)$$

for some constant $C' = C'(\mu, \delta) > 0$.

Then by the transference principles (Theorems 1.10 and 1.11), for ν-almost surely for each $t \geq 0$,

$$\mathbb{P}_X\left(\|W_X - W_{Z^{\mathbf{r}_m,\mathbf{c}_n}}\|_\square \geq t\right) \leq 3e\left[\exp(C_1\rho)\,\mathbb{P}\left(Y \in \mathscr{T}_\rho(\mathbf{r},\mathbf{c})\right)^{-1} \wedge \sup p_{\mathbf{r},\mathbf{c}}(\cdot)\right] \exp\left(m+n-\frac{t^2 mn}{C'}\right).$$

Hence by using a similar argument as in the proof of the second part of Theorem 1.27,

$$\mathbb{P}\left(\|W_X - W_{Z^{\mathbf{r}_m,\mathbf{c}_n}}\|_\square \geq \sqrt{\frac{F_\mu(\mathbf{r},\mathbf{c})}{mn}}\right) \leq 2\exp\left(-F_\mu(\mathbf{r},\mathbf{c})\right).$$

By triangle inequality and Theorem 1.19,

$$\|W_X - W^{\mathbf{r},\mathbf{c}}\|_\square \leq \|W_X - W_{Z^{\mathbf{r}_m,\mathbf{c}_n}}\|_\square + C''\|(\bar{\mathbf{r}}_m, \bar{\mathbf{c}}_n) - (\mathbf{r},\mathbf{c})\|_1$$

for some constant $C''' = C''(\mu, \delta)$. This is enough to conclude (1.31). By Borel-Cantelli lemma, this yields that $W_X \to W^{\mathbf{r},\mathbf{c}}$ in cut norm almost surely as $m, n \to \infty$. □

Next, we establish the limit of the ESD of X.

Proof of Theorem 1.33. In this proof, we will use 'ESD' to refer to the empirical eigenvalue (spectral) distribution of a matrix. Without loss of generality, we assume the sequence of δ-tame margins $(\mathbf{r}_m, \mathbf{c}_n)$ is indexed by k so that $m = m(k)$ and $n = n(k)$ and simply denote $(\mathbf{r}_m, \mathbf{c}_n) = (\mathbf{r}_k, \mathbf{c}_k)$. Let $(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k)$ denote the standard MLE for margin $(\mathbf{r}_k, \mathbf{c}_k)$ and let $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ denote the limiting MLE from Theorem 1.19. Let $\hat{\xi}_k$ denote the ESD of the gram matrix $\hat{\Xi}_k := \tilde{Y}_k \tilde{Y}_k^*$. Similarly, let ξ_k denote the ESD of $\hat{\tilde{X}}_k \tilde{X}_k^*$. Let $R_k(z) := (\hat{\Xi}_k - zI)^{-1}$, $z \in \mathbb{C} \setminus \mathbb{R}$ denote the resolvent of $\hat{\Xi}_k$.

We will first argue for the unique existence and boundedness of the solution τ for the Dyson equation (1.32). By Proposition 6.8 the limiting continuum margin $(\mathbf{r}, \mathbf{c})$ is δ-tame so $\phi(A_\delta) \leq \boldsymbol{\alpha}(x) + \boldsymbol{\beta}(y) \leq \phi(B_\delta)$ for all $(x, y) \in [0,1]^2$. It follows that the kernel $\psi''(\boldsymbol{\alpha} \oplus \boldsymbol{\beta})$ for the integral operator S is uniformly bounded away from zero and from ∞. According to the discussion in [AEK17, Sec. 3.1] (with a slight modification for the operator setting), the problem reduces to showing the same statement about the quadratic vector equation (QVE) $-\frac{1}{\sigma_\cdot(z)} = z + \mathscr{S}\sigma_\cdot(z)$ with $\mathscr{S}$ the symmetrization of S. The hypothesis of [AEK19, Thm. 2.1] is easily verified, so the solution σ is unique and uniformly bounded for every $z \in \mathbb{H}$.

Let S_k denote the integral operator with step-kernel $\psi''(\bar{\boldsymbol{\alpha}}_k \oplus \bar{\boldsymbol{\beta}}_k) : (0,1]^2 \to \mathbb{R}$, where $(\bar{\boldsymbol{\alpha}}_k, \bar{\boldsymbol{\beta}}_k)$ is as in Theorem 1.19. By δ-tameness, this kernel is uniformly bounded away from zero and from ∞ with the bounds depending only on δ and μ. By the same argument, the solution σ^k to the QVE

$-\frac{1}{\sigma^k_\cdot} = \zeta + \mathscr{S}_k \sigma^k_\cdot$ is unique and uniformly bounded. Since ψ'' is Lipschitz continuous on bounded domain, by Theorem 1.19

$$\|\psi''(\boldsymbol{\alpha} \oplus \boldsymbol{\beta}) - \psi''(\bar{\boldsymbol{\alpha}}_k \oplus \bar{\boldsymbol{\beta}}_k)\|_2^2 \le C\|(\mathbf{r}, \mathbf{c}) - (\bar{\mathbf{r}}_k, \bar{\mathbf{c}}_k)\|_1 = o(1)$$

for some constant $C = C(\mu, \delta) > 0$. Thus the variance kernels $\psi''(\bar{\boldsymbol{\alpha}}_k \oplus \bar{\boldsymbol{\beta}}_k)$ converge to $\psi''(\boldsymbol{\alpha} \oplus \boldsymbol{\beta})$ in L^2. This and the stability theorem [AEK19, Thm. 2.13] imply that σ^k converges to σ. Following the discussion in [AEK17, Thm. 3.1], we can transform σ^ks to solutions of the Dyson equation (1.32) with S_k in place of its symmetrization $\mathscr{S}_k$, which we denote as τ^ks. This shows τ^k converges to τ pointwise.

Now we deduce the convergence of the expected singular value distribution of $\widetilde{Y}_k$. The entries of Y_k have uniform subexponential norms due to the δ-tameness. This and the uniform boundedness of the variance kernel verifies the hypothesis in [AEK17]. So by [AEK17, Thm. 2.2], the Stieltjes transform of $\hat{\xi}_k$, $m^{-1}\mathrm{tr}R_k(z)$, is very close to $\langle \tau^k(z) \rangle$ in the half-plane $\mathrm{Im}(z) \ge \varepsilon^*$ for any fixed $\varepsilon^* > 0$. (see the reference for a precise statement). Since we know $\tau^k \to \tau$ pointwise, it follows that the Stieltjes transforms of $\hat{\xi}_k$ converge pointwise to τ. Hence $\hat{\xi}_k$ converges weakly to the probability measure ξ given by the inverse Stieltjes transform of τ. In fact, by the convergence of local laws in [AEK17, Thm.2.7, 2.9], it also follows that $\mathbb{E}[\hat{\xi}_k] \to \xi$ weakly. At this point, we have shown **(i)** for the tilted models $\tilde{Y}_k$.

To show **(ii)**, we wish to show that ξ_k converges to ξ in probability as $k \to \infty$. Let $d_{\mathscr{W}_1}$ denote the Wasserstein-1 distance between probability measures. Fix $\varepsilon > 0$. Then by the transference principles that hold under (A1) (see Theorem 1.10 and Theorem 1.14),

$$\mathbb{P}\left(d_{\mathscr{W}_1}(\xi_k, \mathbb{E}[\hat{\xi}_k]) \ge \varepsilon\right) \le \exp(Cn^{3/2}(\log n)^2)\, \mathbb{P}\left(d_{\mathscr{W}_1}(\hat{\xi}_k, \mathbb{E}[\hat{\xi}_k]) \ge \varepsilon\right)$$

for some constant $C = C(\mu, \delta, \kappa) > 0$. We will first consider the special case when μ has compact support. According to [GZ00, Cor. 1.8] of Guionnet and Zeitouni (also see the remark following the statement), we have the following sub-Gaussian concentration of ESD of the Wishart matrix $\hat{\Xi}_k$:

$$\mathbb{P}\left(d_{\mathscr{W}_1}(\hat{\xi}_k, \mathbb{E}[\hat{\xi}_k]) \ge \varepsilon\right) \le \exp(-c(\varepsilon)(m+n)^2). \tag{9.3}$$

Combining the above, for some constant $D = D(\mu, \delta, \kappa, \varepsilon) > 0$,

$$\mathbb{P}\left(d_{\mathscr{W}_1}(\xi_k, \mathbb{E}[\hat{\xi}_k]) \ge \varepsilon\right) \le \exp(-D(m+n)^2).$$

Since $\mathbb{E}[\hat{\xi}_k] \to \xi$ weakly, it follows that $\xi_k \to \xi$ weakly in probability, as desired.

It remains to justify the sub-Gaussian ESD concentration for $\hat{\Xi}_k$ (9.3) for the more general case when μ is sub-Gaussian. We remark that the sub-Gaussian ESD concentration for Wishart matrices in [GZ00, Cor. 1.8] is a direct consequence of the similar result for the Wigner matrices stated in [GZ00, Thm. 1.1] using Girko's symmetrization trick (see the discussion above [GZ00, Cor. 1.8]). This result for the Wigner matrices holds when the laws of the entries have common compact support or satisfy the log-Sobolev inequality with a uniform constant. Klochkov and Zhivotovskiy [KZ20, Lem. 1.4] showed that such a result holds when the laws of the entries are uniformly sub-Gaussian. It is an elementary fact that the bounded exponential tilt of a sub-Gaussian distribution is sub-Gaussian with a uniform sub-Gaussian norm. It follows that the sub-Gaussian norms of the entries of Y are uniformly bounded by a constant depending only on μ and δ. This is enough to conclude. □

Proof of Corollary 1.34. This follows immediately from Theorem 1.33 by the argument sketched above the statement of Corollary 1.34. □

Proof of Corollary 1.35. When μ is standard Gaussian, then according to the computations in Ex. 3.1, the variance matrix $\psi''(\boldsymbol{\alpha} \oplus \boldsymbol{\beta})$ is always the all-ones matrix $\mathbf{1}_m \mathbf{1}_n^\top$. Hence, the Dyson equation

(1.32) is the same as the one that characterizes the limiting ESD of a $m \times n$ matrix with independent and unit variance entries. The limiting ESD for such matrices is known to be the Marchenko-Pastur distribution given in the statement (see, e.g., [AEK17]). □

10. Concluding remarks

Finally, we provide some concluding remarks and discuss several intriguing open problems.

Random measure perspective. Our main result stated in Theorem 1.19 assumes that the sequence of margins $(\mathbf{r}_m, \mathbf{c}_n)$ converges (after rescaling) in L^1. This assumption is natural from the perspective of viewing random matrices as random functions. A natural question that we have not investigated in this work is to view the random matrices as random measures on $[0,1]^2$ (as done in the permutation limit theory [HKM+13, BDMW24]) and ask if the marginal measures converge weakly, then the joint random measure should also converge to a limiting measure, possibly a deterministic one. We conjecture that if the marginal measures converge to limiting ones with density with respect to the Lebesgue measure, then the joint random measure in the limit should be given by the deterministic measure whose density is the limiting typical kernel as in Theorem 1.19.

Conditioning on other constraints. In this work, we only considered conditioning a large random matrix on dense margins, in which the row and column sums are proportional to the number of columns and rows, respectively. Can one develop an analogous concentration and limit theory of random matrices conditioned on sparse margins? Also, note that row and column margin of an $m \times n$ matrix $\mathbf{X}$ are particular linear constraints of the form $\mathbf{X}\mathbf{1}_m = \mathbf{r}$ and $\mathbf{1}_n\mathbf{X} = \mathbf{c}^\top$. Can one develop an analogous concentration and limit theory of random matrices conditioned on general linear constraints of the form $\mathbf{X}\mathbf{u} = \mathbf{b}$ and $\mathbf{u}'\mathbf{X} = \mathbf{b}'$ for vectors $\mathbf{u}, \mathbf{u}', \mathbf{b}, \mathbf{b}'$ of appropriate dimensions? Moreover, can one establish similar results for symmetric random matrices under constraints? For instance, this will give a concentration and limit theory for Wigner matrices with a given row sum, which aligns well with the limit theory of random graphs with given degree sequences [CDS11].

Large deviations principle. Dhara and Sen [DS22] established a large deviations principle for random graphs with prescribed degree sequences, building upon foundational work by Chatterjee, Diaconis, and Sly [CDS11]. We anticipate a similar large deviations principle for random matrices conditioned on given margins, with the information projection (2.7) likely serving as the rate function. This is currently an ongoing work by the authors.

Schrödinger bridge and optimal transport. Drawing an analogy from the connection between the static Schrödinger bridges and the entropic optimal transport (see Sec. 2.1), we can also consider an 'entropic optimal transport' version of our conditioned random matrix problem. Namely, let $\gamma : \mathbb{R}^{m\times n} \to \mathbb{R}_{\geq 0}$ be a cost function on $m \times n$ real matrices. Then replacing the base model $\mathscr{R}$ with a probability measure proportional to $e^{-\gamma/\varepsilon}\mathscr{R}$, (2.7) becomes

$$\min_{\mathscr{H}\in\mathscr{P}^{m\times n}} \int_{\mathbb{R}^{m\times n}} \gamma(\mathbf{x})\mathscr{H}(d\mathbf{x}) + \varepsilon\, D_{KL}(\mathscr{H}\,\|\,\mathscr{R}) \quad \text{subject to} \quad \mathbb{E}_{X\sim\mathscr{H}}[(r(X), c(X))] = (\mathbf{r}, \mathbf{c}).$$

It remains to be seen whether the 'typical tables' in this generalized setting exhibit interesting structures.

Acknowledgements

HL is partially supported by NSF DMS-2206296 and DMS-2232241. SM is partially supported by NSF-2113414. The authors thank Alexander Barvinok, Peter Winkler, Igor Pak, Hongchang Ji, and Rami Tabri for helpful discussions.

REFERENCES

[AEK17] Johannes Alt, László Erdős, and Torben Krüger, *Local law for random gram matrices*, Electronic Journal of Probability **22** (2017), no. 25, 1–41.

[AEK19] Oskari Ajanki, László Erdős, and Torben Krüger, *Quadratic vector equations on complex upper half-plane*, vol. 261, American Mathematical Society, 2019.

[AGZ10] Greg W Anderson, Alice Guionnet, and Ofer Zeitouni, *An introduction to random matrices*, no. 118, Cambridge university press, 2010.

[Bar09] Alexander Barvinok, *Asymptotic estimates for the number of contingency tables, integer flows, and volumes of transportation polytopes*, International Mathematics Research Notices **2009** (2009), no. 2, 348–385.

[Bar10a] ———, *On the number of matrices and a random matrix with prescribed row and column sums and 0–1 entries*, Advances in Mathematics **224** (2010), no. 1, 316–339.

[Bar10b] ———, *What does a random contingency table look like?*, Combinatorics, Probability and Computing **19** (2010), no. 4, 517–539.

[BC89] Julian Besag and Peter Clifford, *Generalized monte carlo significance tests*, Biometrika **76** (1989), no. 4, 633–642.

[BCL+08] Christian Borgs, Jennifer T Chayes, László Lovász, Vera T Sós, and Katalin Vesztergombi, *Convergent sequences of dense graphs i: Subgraph frequencies, metric properties and testing*, Advances in Mathematics **219** (2008), no. 6, 1801–1851.

[BCL+12] ———, *Convergent sequences of dense graphs ii. multiway cuts and statistical physics*, Annals of Mathematics (2012), 151–219.

[BD95] Bhaskar Bhattacharya and Richard Dykstra, *A general duality approach to i-projections*, Journal of statistical planning and inference **47** (1995), no. 3, 203–216.

[BDMW24] Jacopo Borga, Sayan Das, Sumit Mukherjee, and Peter Winkler, *Large deviation principle for random permutations*, International Mathematics Research Notices **2024** (2024), no. 3, 2138–2191.

[Ber97] Dimitri P Bertsekas, *Nonlinear programming*, Journal of the Operational Research Society **48** (1997), no. 3, 334–334.

[BH10] Alexander Barvinok and JA Hartigan, *Maximum entropy gaussian approximations for the number of integer points and volumes of polytopes*, Advances in Applied Mathematics **45** (2010), no. 2, 252–289.

[BH12] Alexander Barvinok and J Hartigan, *An asymptotic formula for the number of non-negative integer matrices with prescribed row and column sums*, Transactions of the American Mathematical Society **364** (2012), no. 8, 4323–4368.

[BH13] Alexander Barvinok and John A Hartigan, *The number of graphs and a random graph with a given degree sequence*, Random Structures & Algorithms **42** (2013), no. 3, 301–348.

[BH20] Petter Brändén and June Huh, *Lorentzian polynomials*, Annals of Mathematics **192** (2020), no. 3, 821–891.

[BLP23] Petter Brändén, Jonathan Leake, and Igor Pak, *Lower bounds for contingency tables via lorentzian polynomials*, Israel Journal of Mathematics **253** (2023), no. 1, 43–90.

[BLSY10] Alexander Barvinok, Zur Luria, Alex Samorodnitsky, and Alexander Yong, *An approximation algorithm for counting contingency tables*, Random Structures & Algorithms **37** (2010), no. 1, 25–66.

[BR24] Alexander Barvinok and Mark Rudelson, *A quick estimate for the volume of a polyhedron*, Israel Journal of Mathematics (2024), 1–25.

[BT13] Amir Beck and Luba Tetruashvili, *On the convergence of block coordinate descent type methods*, SIAM journal on Optimization **23** (2013), no. 4, 2037–2060.

[Car22] Guillaume Carlier, *On the linear convergence of the multimarginal sinkhorn algorithm*, SIAM Journal on Optimization **32** (2022), no. 2, 786–794.

[CDG+06] Mary Cryan, Martin Dyer, Leslie Ann Goldberg, Mark Jerrum, and Russell Martin, *Rapidly mixing markov chains for sampling contingency tables with a constant number of rows*, SIAM Journal on Computing **36** (2006), no. 1, 247–278.

[CDHL05] Yuguo Chen, Persi Diaconis, Susan P Holmes, and Jun S Liu, *Sequential monte carlo methods for statistical analysis of tables*, Journal of the American Statistical Association **100** (2005), no. 469, 109–120.

[CDPS17] Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer, *Convergence of entropic schemes for optimal transport and gradient flows*, SIAM Journal on Mathematical Analysis **49** (2017), no. 2, 1385–1418.

[CDS11] Sourav Chatterjee, Persi Diaconis, and Allan Sly, *Random graphs with a given degree sequence*, The Annals of Applied Probability **21** (2011), no. 4, 1400–1435.

[CDS14] ———, *Properties of random doubly stochastic matrices*, Ann. de lInst. Henri Poincaré (2014).

[CEW22] Mark Colarusso, William Erickson, and Jeb Willenbring, *Contingency tables and the generalized littlewood–richardson coefficients*, Proceedings of the American Mathematical Society **150** (2022), no. 1, 79–94.

[CG08] Fan Chung and Ron Graham, *Quasi-random graphs with given degree sequences*, Random Structures & Algorithms **32** (2008), no. 1, 1–19.

[CGP16] Yongxin Chen, Tryphon Georgiou, and Michele Pavon, *Entropic and displacement interpolation: a computational approach using the hilbert metric*, SIAM Journal on Applied Mathematics **76** (2016), no. 6, 2375–2396.

[CL02] Fan Chung and Linyuan Lu, *Connected components in random graphs with given expected degree sequences*, Annals of combinatorics **6** (2002), no. 2, 125–145.

[CM07] E Rodney Canfield and Brendan D McKay, *The asymptotic volume of the birkhoff polytope*, arXiv preprint arXiv:0705.2422 (2007).

[CM10] ———, *Asymptotic enumeration of integer matrices with large equal row and column sums*, Combinatorica **30** (2010), no. 6, 655.

[CP97] Joseph T Chang and David Pollard, *Conditioning as disintegration*, Statistica Neerlandica **51** (1997), no. 3, 287–317.

[CS79] Edward F Connor and Daniel Simberloff, *The assembly of species communities: chance or competition?*, Ecology **60** (1979), no. 6, 1132–1140.

[Csi75] Imre Csiszár, *I-divergence geometry of probability distributions and minimization problems*, The annals of probability (1975), 146–158.

[Csi84] ———, *Sanov property, generalized i-projection and a conditional limit theorem*, The Annals of Probability (1984), 768–793.

[Cut13] Marco Cuturi, *Sinkhorn distances: Lightspeed computation of optimal transport*, Advances in neural information processing systems **26** (2013).

[DE85] Persi Diaconis and Bradley Efron, *Testing for independence in a two-way table: new interpretations of the chi-square statistic*, The Annals of Statistics (1985), 845–874.

[DG95] Persi Diaconis and Anil Gangolli, *Rectangular arrays with fixed margins*, Discrete probability and algorithms, Springer, 1995, pp. 15–41.

[DG00] Martin Dyer and Catherine Greenhill, *Polynomial-time counting and sampling of two-rowed contingency tables*, Theoretical Computer Science **246** (2000), no. 1-2, 265–278.

[DGKTB10] Charo I Del Genio, Hyunju Kim, Zoltán Toroczkai, and Kevin E Bassler, *Efficient and exact sampling of simple graphs with given arbitrary degree sequence*, PloS one **5** (2010), no. 4, e10012.

[DKM97] Martin Dyer, Ravi Kannan, and John Mount, *Sampling contingency tables*, Random Structures & Algorithms **10** (1997), no. 4, 487–506.

[DLP20] Samuel Dittmer, Hanbaek Lyu, and Igor Pak, *Phase transition in random contingency tables with non-uniform margins*, Transactions of the American Mathematical Society **373** (2020), no. 12, 8313–8338.

[DP18] Sam Dittmer and Igor Pak, *Contingency tables have empirical bias*, In preparation (2018).

[DS98] Persi Diaconis and Bernd Sturmfels, *Algebraic algorithms for sampling from conditional distributions*, The Annals of statistics **26** (1998), no. 1, 363–397.

[DS22] Souvik Dhara and Subhabrata Sen, *Large deviation for uniform graphs with given degrees*, Ann. Appl. Probab. **32** (2022), no. 3, 2327–53.

[DSC95] Persi Diaconis and Laurent Saloff-Coste, *Random walk on contingency tables with fixed row and column sums*, Tech. report, Technical Report, Department of Mathematics, Harvard University, 1995.

[Dye03] Martin Dyer, *Approximate counting by dynamic programming*, Proceedings of the thirty-fifth annual ACM symposium on Theory of computing, 2003, pp. 693–699.

[EG60] P Erdos and Tibor Gallai, *Graphen mit punkten vorgeschriebenen grades*, Mat. Lapok **11** (1960), 264–274.

[FL89] Joel Franklin and Jens Lorenz, *On the scaling of multidimensional matrices*, Linear Algebra and its applications **114** (1989), 717–735.

[For40] Robert Fortet, *Résolution d'un système d'équations de m. Schrödinger*, Journal de Mathématiques Pures et Appliquées **19** (1940), no. 1-4, 83–105.

[GIM21] Catherine Greenhill, Mikhail Isaev, and Brendan D McKay, *Subgraph counts for dense random graphs with specified degrees*, Combinatorics, Probability and Computing **30** (2021), no. 3, 460–497.

[GM09] Catherine Greenhill and Brendan D McKay, *Random dense bipartite graphs and directed graphs with specified degrees*, Random Structures & Algorithms **35** (2009), no. 2, 222–249.

[Goo50] Isidore Jacob Good, *Probability and the weighing of evidence*, C. Griffin London, 1950.

[Goo63] Irving J Good, *Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables*, The Annals of Mathematical Statistics **34** (1963), no. 3, 911–934.
[GZ00] Alice Guionnet and Ofer Zeitouni, *Concentration of the spectral measure for large matrices*.
[HKM+13] Carlos Hoppen, Yoshiharu Kohayakawa, Carlos Gustavo Moreira, Balázs Ráth, and Rudini Menezes Sampaio, *Limits of permutation sequences*, Journal of Combinatorial Theory, Series B **103** (2013), no. 1, 93–113.
[Ide16] Martin Idel, *A review of matrix scaling and sinkhorn's normal form for matrices and positive maps*, arXiv preprint arXiv:1609.06349 (2016).
[KTV99] Ravi Kannan, Prasad Tetali, and Santosh Vempala, *Simple markov-chain algorithms for generating bipartite graphs and tournaments*, Random Structures & Algorithms **14** (1999), no. 4, 293–308.
[KZ20] Yegor Klochkov and Nikita Zhivotovskiy, *Uniform hanson-wright type concentration inequalities for unbounded entries via the entropy method*, Electron. J. Probab **25** (2020), no. 22, 1–30.
[Léo13] Christian Léonard, *A survey of the Schrödinger problem and some of its connections with optimal transport*, arXiv preprint arXiv:1308.0215 (2013).
[LM22] Dirk Lorenz and Hinrich Mahler, *Orlicz space regularization of continuous optimal transport problems*, Applied Mathematics & Optimization **85** (2022), no. 2, 14.
[Lov12] László Lovász, *Large networks and graph limits*, vol. 60, American Mathematical Soc., 2012.
[LP22] Hanbaek Lyu and Igor Pak, *On the number of contingency tables and the independence heuristic*, Bulletin of the London Mathematical Society **54** (2022), no. 1, 242–255.
[Lur08] Zur Luria, *Counting contingency tables with balanced margins*, manuscript (2008).
[MG20] Simone Di Marino and Augusto Gerolin, *An optimal transport approach for the Schrödinger bridge problem and convergence of sinkhorn algorithm*, Journal of Scientific Computing **85** (2020), no. 2, 27.
[Mor02] Ben J Morris, *Improved bounds for sampling contingency tables*, Random Structures & Algorithms **21** (2002), no. 2, 135–146.
[MP67] VA Marchenko and Leonid A Pastur, *Distribution of eigenvalues for some sets of random matrices*, Mat. Sb.(NS) **72** (1967), no. 114, 4.
[MR95] Michael Molloy and Bruce Reed, *A critical point for random graphs with a given degree sequence*, Random structures & algorithms **6** (1995), no. 2-3, 161–180.
[MR98] ———, *The size of the giant component of a random graph with a given degree sequence*, Combinatorics, probability and computing **7** (1998), no. 3, 295–305.
[MW90a] Brendan D McKay and Nicholas C Wormald, *Asymptotic enumeration by degree sequence of graphs of high degree*, European Journal of Combinatorics **11** (1990), no. 6, 565–580.
[MW90b] ———, *Uniform generation of random regular graphs of moderate degree*, Journal of Algorithms **11** (1990), no. 1, 52–67.
[Ngu14] Hoi H Nguyen, *Random doubly stochastic matrices: the circular law*.
[NN20] Evita Nestoridi and Oanh Nguyen, *On the mixing time of the diaconis–gangolli random walk on contingency tables over z/qz*.
[Nut21] Marcel Nutz, *Introduction to entropic optimal transport*, Lecture notes, Columbia University (2021).
[NW22] Marcel Nutz and Johannes Wiesel, *Entropic optimal transport: Convergence of potentials*, Probability Theory and Related Fields **184** (2022), no. 1, 401–424.
[ODCdS+15] Chiara Orsini, Marija M Dankulov, Pol Colomer-de Simón, Almerima Jamakovic, Priya Mahadevan, Amin Vahdat, Kevin E Bassler, Zoltán Toroczkai, Marián Boguná, Guido Caldarelli, et al., *Quantifying randomness in real networks*, Nature communications **6** (2015), no. 1, 8627.
[PTT21] Michele Pavon, Giulio Trigila, and Esteban G Tabak, *The data-driven Schrödinger bridge*, Communications on Pure and Applied Mathematics **74** (2021), no. 7, 1545–1573.
[Ren88] James Renegar, *A polynomial-time algorithm, based on newton's method, for linear programming*, Mathematical programming **40** (1988), no. 1, 59–93.
[Rus95] Ludger Ruschendorf, *Convergence of the iterative proportional fitting procedure*, The Annals of Statistics (1995), 1160–1174.
[San61] Ivan Nicolaevich Sanov, *On the probability of large deviations of random variables*, Selected Translations in Mathematical Statistics and Probability **1** (1961), 213–244.
[Sin64] Richard Sinkhorn, *A relationship between arbitrary positive matrices and doubly stochastic matrices*, The annals of mathematical statistics **35** (1964), no. 2, 876–879.
[Sni91] Tom AB Snijders, *Enumeration and simulation methods for 0–1 matrices with given marginals*, Psychometrika **56** (1991), 397–417.
[Tal96] Michel Talagrand, *A new look at independence*, The Annals of probability (1996), 1–34.

[TGS22] Dávid Terjék and Diego González-Sánchez, *Optimal transport with f-divergence regularization and generalized sinkhorn algorithm*, International Conference on Artificial Intelligence and Statistics, PMLR, 2022, pp. 5135–5165.

[Ver08] Norman D Verhelst, *An efficient mcmc algorithm to sample binary matrices with fixed marginals*, Psychometrika **73** (2008), no. 4, 705–728.

[Ver18] Roman Vershynin, *High-dimensional probability: An introduction with applications in data science*, vol. 47, Cambridge university press, 2018.

[Vil21] Cédric Villani, *Topics in optimal transportation*, vol. 58, American Mathematical Soc., 2021.

[Wan20] Guanyang Wang, *A fast mcmc algorithm for the uniform sampling of binary matrices with fixed margins*.

[Wu20] Da Wu, *On properties of random binary contingency tables with non-uniform margin*, arXiv preprint arXiv:2002.12559 (2020).

[Wu23] ———, *Asymptotic properties of random contingency tables with uniform margin*, Journal of Theoretical Probability **36** (2023), no. 4, 2066–2092.

[YW09] Xiaowei Ying and Xintao Wu, *Graph generation with prescribed feature constraints*, Proceedings of the 2009 SIAM International Conference on Data Mining, SIAM, 2009, pp. 966–977.

HANBAEK LYU, DEPARTMENT OF MATHEMATICS, UNIVERSITY OF WISCONSIN - MADISON, WI, 53717, USA
Email address: hlyu@math.wisc.edu

SUMIT MUKHERJEE, DEPARTMENT OF STATISTICS, COLUMBIA UNIVERSITY, NEW YORK, NY 10027, USA
Email address: sm3949@columbia.edu