ALEA, Lat. Am. J. Probab. Math. Stat. 19, 1045–1070 (2022)
DOI: 10.30757/ALEA.v19-43

Maximum likelihood estimation for discrete exponential families and random graphs

Krzysztof Bogdan, Michał Bosy and Tomasz Skalski

Wrocław University of Science and Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
E-mail address: krzysztof.bogdan@pwr.edu.pl
URL: http://prac.im.pwr.edu.pl/∼bogdan/

School of Computer Science and Mathematics, Kingston University London, Penrhyn Road, KT1 2EE Kingston upon Thames, UK
E-mail address: m.bosy@kingston.ac.uk
URL: https://michalbosy.wixsite.com/mathematician

Wrocław University of Science and Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
& LAREMA – UMR CNRS 6093, Université d'Angers, 2 bd Lavoisier, 49045 Angers Cedex 01, France
E-mail address: tomasz.skalski@pwr.edu.pl
URL: http://prac.im.pwr.edu.pl/∼skalski/

Abstract. We characterize the existence of the maximum likelihood estimator for discrete exponential family on finite set. Our criterion is simple to apply as we show in various settings, most notably for exponential models of random graphs. As an application, we point out the size of independent identically distributed samples for which the maximum likelihood estimator exists with high probability.

1. Introduction and preliminaries

Exponential families are of paramount importance in probability and statistics. They were introduced by Fisher, Pitman, Darmois and Koopman in 1934-36 and have many properties that make them indispensable in theory and applications, see Lehmann and Casella (1998, Section 2.7), Barndorff-Nielsen (1978, Chapter 9), Andersen (1970), Diaconis (1988, Chapter 9.E), Diaconis and Freedman (1984), and Lauritzen (1984). In this paper we study *discrete* exponential families, more specifically, exponential families on *finite* sets, and give a new

Received by the editors December 29th, 2021; accepted April 13th, 2022.
2010 *Mathematics Subject Classification.* 05C80, 62H12.
Key words and phrases. Maximum likelihood, discrete exponential family, random graph.
The first author was supported in part by NCN (National Science Center, Poland) grant 2018/31/G/ST1/02252 and grant 049U/0052/19 from WUST. The third author was supported in part by grant 049M/0010/19 from WUST and a French Government Scholarship.

characterization of the existence of the maximum likelihood estimator (MLE) for exponential family and the data at hand. We also present applications, in particular for specific exponential families we give a threshold of the sample size sufficient for the existence of MLE with high probability for *i.i.d.* samples.

The computation of MLE is in general difficult with the number of variables increasing. On the other hand, for given data and an exponential family, MLE may fail to exist. In particular, Crain (1974, 1976) pointed out to problems with the maximum likelihood estimation when the number of parameters is too large for the sample size. He also gave a sufficient condition for MLE to exist almost surely – the Haar condition.

A complete characterization of the existence of MLE for rather general exponential families was given by Barndorff-Nielsen. Namely, by Barndorff-Nielsen (1978, Theorem 9.13), MLE for a sample and an exponential family exists if and only if the vector of the sample means calculated for a basis of the linear space of exponents belongs to the interior of the convex hull of the pointwise range of the basis.

This beautiful criterion is alas cumbersome to apply. Therefore, Jacobsen (1989) gives an alternative condition for discrete exponential families, together with applications to Cox regression, logistic regression and multiplicative Poisson models. Similar condition is presented by Albert and Anderson (1984) for log-linear model. Haberman (1974) gives a characterization of the existence of MLE for hierarchical log-linear models. His conditions can be interpreted in terms of polytope geometry, see also Eriksson et al. (2006), and Fienberg and Rinaldo (2012). Brown (1986) characterizes the existence of MLE when the log-partition function is steep and regularly convex, and interprets the problem of finding MLE as the optimization of the Kullback-Leibler divergence. Darroch et al. (1980) connect the properties of MLE in decomposable models with graph-theoretical notions, thus starting the theory of graphical models in statistics. Sufficient conditions for the existence of MLE in specific exponential families are also given by Stone (1990) and Bogdan and Ledwina (1996). Geyer (1990) looks for MLE in the closure of convex exponential families, relates the existence of MLE with the linear programming feasibility problem, and in the case of nonexistent MLE, reduces the considered exponential family until MLE exists for the family. He also applies MCMC algorithms to calculate MLE. A comparison between the conditions of Barndorff-Nielsen and Jacobsen is discussed in Konis (2007). In addition, Konis presents an implementation of Jacobsen's test using linear programming. A broad survey of the history of log-linear models and further motivation for the study of the existence of MLE can be found in Fienberg and Rinaldo (2007, 2012).

The main inspiration for our work is Bogdan and Bogdan (2000, Theorem 2.3) on the existence of MLE for exponential families of continuous functions on finite interval. In Theorem 2.2 below we propose a similar characterization, which is new in the setting of discrete exponential families. We obtain the result by a straightforward, self-contained approach, which does not depend on the delicate convex analysis of Barndorff-Nielsen (1978).

The paper is composed as follows. In Section 2 we state and prove our criterion, using the notion of *set of uniqueness*. The criterion is restated in Section 2.2 as a linear programming problem. In Section 3 we give applications to exponential families spanned by Rademacher and Walsh functions, and to exponential families of random graphs. In particular we give sharp or plain thresholds for the sample size to secure the existence of MLE with high probability. In Appendix A we give auxiliary results and reformulations of our criterion and pin down its connections with the criterion of Barndorff-Nielsen.

Acknowledgments: We are grateful to Małgorzata Bogdan, Piotr Ciołek, Persi Diaconis, Hélène Massam, Sumit Mukherjee, Krzysztof Oleszkiewicz, Krzysztof Samotij, Maciej Wilczyński and anonymous referees for comments, corrections, references and discussion. The third author was partially supported by the BGF Cotutelle scholarship and would like to thank Campus France for financial support.

1.1. *Discrete exponential family.* Consider a finite set $\mathcal{X} \neq \emptyset$ and *weight* function $\mu : \mathcal{X} \to (0, \infty)$. As usual, $\mathbf{R}^{\mathcal{X}}$ is the family of all the real-valued functions on $\mathcal{X}$. For $\phi \in \mathbf{R}^{\mathcal{X}}$ we define the *partition* and the *log-partition* functions,

$$Z(\phi) = \sum_{x \in \mathcal{X}} e^{\phi(x)} \mu(x), \quad \psi(\phi) = \log Z(\phi), \tag{1.1}$$

respectively, and the *exponential density*

$$p = e(\phi) = e^{\phi - \psi(\phi)} = e^{\phi}/Z(\phi). \tag{1.2}$$

Clearly, $p > 0$ and $\sum_{x \in \mathcal{X}} p(x)\mu(x) = 1$. For arbitrary real number c we have $\psi(\phi+c) = \psi(\phi)+c$, hence

$$e(\phi + c) = e(\phi). \tag{1.3}$$

Moreover, for $\phi_1, \phi_2 \in \mathbf{R}^{\mathcal{X}}$ we have $e(\phi_1) = e(\phi_2)$ if and only if $\phi_1 - \phi_2$ is constant. Consider $x_1, \ldots, x_n \in \mathcal{X}$, a *sample*. For $\phi \in \mathbf{R}^{\mathcal{X}}$ we denote, as usual,

$$\bar{\phi} = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i).$$

The *likelihood function* of $p = e(\phi)$ is defined as

$$L_{e(\phi)}(x_1, \ldots, x_n) = L_p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i),$$

and the *log-likelihood function* is

$$l_{e(\phi)}(x_1, \ldots, x_n) := \log L_{e(\phi)}(x_1, \ldots, x_n) = n\left(\bar{\phi} - \psi(\phi)\right). \tag{1.4}$$

Of course, for every $c \in \mathbf{R}$ we have

$$l_{e(\phi+c)}(x_1, \ldots, x_n) = l_{e(\phi)}(x_1, \ldots, x_n). \tag{1.5}$$

We note that the likelihood functions are uniformly bounded. Indeed, for every $\phi \in \mathbf{R}^{\mathcal{X}}$,

$$\psi(\phi) = \log \sum_{x \in \mathcal{X}} e^{\phi(x)} \mu(x) \geq \max_{\mathcal{X}} \phi + \min_{\mathcal{X}} \log \mu, \tag{1.6}$$

and so by (1.4) and (1.6),

$$l_{e(\phi)}(x_1, \ldots, x_n) \leq -n \min_{\mathcal{X}} \log \mu \quad \text{and} \quad L_{e(\phi)}(x_1, \ldots, x_n) \leq (\min_{\mathcal{X}} \mu)^{-n}. \tag{1.7}$$

We fix a linear subspace $\mathcal{B} \subset \mathbf{R}^{\mathcal{X}}$. The *exponential family* spanned by $\mathcal{B}$ is

$$e(\mathcal{B}) := \{p = e(\phi) : \phi \in \mathcal{B}\}. \tag{1.8}$$

Since $\mathcal{X}$ is a finite set, $e(\mathcal{B})$ will be called *discrete exponential family* (we do not consider infinite countable sets, for which see Jacobsen (1989)).

We call $\hat{p} \in e(\mathcal{B})$ an MLE for $x_1, \ldots, x_n$ and $e(\mathcal{B})$ if

$$L_{\hat{p}}(x_1, \ldots, x_n) = \sup_{p \in e(\mathcal{B})} L_p(x_1, \ldots, x_n),$$

or, equivalently,

$$l_{\hat{p}}(x_1, \ldots, x_n) = \sup_{p \in e(\mathcal{B})} l_p(x_1, \ldots, x_n).$$

The following result is well known (see, e.g., Johansen (1979, Theorem 2.1) or Diaconis (1988, p. 177)), but for the reader's convenience we give a proof in Appendix A.1.

Lemma 1.1. *If MLE exists, then it is unique.*

Despite the boundedness (1.7), MLE may fail to exist, as shown by the following example.

Example 1.2. Let $\mathcal{X} = \{0, 1\}$, $\mu \equiv 1$, $\mathcal{B} = \mathbf{R}^{\mathcal{X}}$, $n = 1$ and $x_1 = 1$. Let $a, b \in \mathbf{R}$ and $\phi = a + b\mathbb{1}_{\{1\}}$. Then $Z(\phi) = e^a(1 + e^b)$, $e(\phi) = e^{b\mathbb{1}_{\{1\}}}/(1 + e^b)$, and $L_{e(\phi)}(x_1) = e(\phi)(1) = e^b/(1 + e^b)$. Thus, $\sup L_{e(\phi)}(x_1) = 1$, but the supremum is not attained for any $a, b \in \mathbf{R}$, so MLE does not exist in this case. On the other hand, if $n = 3$, $x_1 = x_2 = 0$, and $x_3 = 1$, then $L_{e(\phi)}(x_1, x_2, x_3) = e^b/(1 + e^b)^3$. By calculus, the maximum is attained when $e^b = 1/2$, therefore $\hat{p} = (2 - \mathbb{1}_{\{1\}})/3$ is the MLE in this case.

We note that the first supremum in Example 1.2 is approached when $b \to \infty$, or for the density $p = \mathbb{1}_{\{1\}}$, which, however, is not in $e(\mathbf{R}^{\mathcal{X}})$ but rather in $e(\mathbf{R}^{\{1\}})$. Below in Theorem 2.2 we characterize the situation when the genuine MLE exists, and in Theorem 2.6 we treat, by a suitable reduction of $\mathcal{X}$, the case when the supremum of the likelihood function is "attained at infinity". Before we proceed, we owe the reader some comments on the notation used in this paper and in the literature.

1.2. *Alternative setting.* Let d be a natural number. Consider a nonempty finite set $S \subset \mathbf{R}^d$, weight m on S and the linear space spanned by the coordinate functions on $\mathbf{R}^d$. The corresponding exponential densities have the form

$$\pi_\theta(y) = e^{\theta \cdot y}/\zeta(\theta), \quad y \in S, \tag{1.9}$$

where $\theta \in \mathbf{R}^d$, $\cdot$ is the scalar product in $\mathbf{R}^d$ and $\zeta(\theta) = \sum_{y \in \mathcal{S}} e^{\theta \cdot y} m(y)$. Thus, (1.9) is a *natural*, or *standard*, exponential family, see Letac (1992) or Brown (1986). Since the range of the vector of parameters θ is the whole of $\mathbf{R}^d$, which is open, the exponential family (1.9) is *regular*, see Lauritzen (1996, Appendix D.1). The setting is actually generic, as we explain momentarily. If functions $\phi_1, \ldots, \phi_d$ span the linear space $\mathcal{B}$ in the general discussion above and we let $T(x) = (\phi_1(x), \ldots, \phi_d(x))$ for $x \in \mathcal{X}$, then for every $\phi \in \mathcal{B}$ there is $\theta \in \mathbf{R}^d$ such that $\phi(x) = \theta \cdot T(x)$ for $x \in \mathcal{X}$, and

$$e(\phi) = e^{\theta \cdot T}/Z(\theta \cdot T). \tag{1.10}$$

This is the form used by most authors, see Lauritzen (1996) or Johansen (1979), and T is called the *canonical* statistics. Furthermore, we let $S = T(\mathcal{X}) \subset \mathbf{R}^d$ and $m(y) = \sum_{x: T(x) = y} \mu(x)$ for $y \in S$. With the notation of (1.9) and (1.10) we have

$$\pi_\theta(y) = e(\phi)(x) \quad \text{if} \quad T(x) = y. \tag{1.11}$$

If $x_1, \ldots, x_n \in \mathcal{X}$ is the sample and we denote $y_1 = T(x_1), \ldots, y_n = T(x_n)$, then the corresponding likelihoods are equal, too. Therefore $\pi_{\hat{\theta}}$ is the maximum likelihood estimator for

$y_1, \ldots, y_n$ and $\{\pi_\theta,\ \theta \in \mathbf{R}^d\}$ if and only if $e(\hat{\theta} \cdot T)$ is the maximum likelihood estimator for $x_1, \ldots, x_n$ and $\{e(\phi),\ \phi \in \mathcal{B}\}$. This makes a complete connection between our setting and the setting of natural exponential families with finite support S. The same setting of discrete exponential families on finite set is described, using slightly different language, in Sullivant (2018, §6.2). We also recall that if $\phi_1, \ldots, \phi_d$ are affinely independent, then the representation (1.10) is *minimal*, see Johansen (1979, Chapter 1) or Lauritzen (1996), where the affine independence means that $\theta \cdot T = const$ implies $\theta = 0$. In general, one allows the representation to be nonminimal because over-parametrization is often natural in applications. We shall return to this discussion again in Section A.6, but for now we get back to the setting of $\mathcal{B}$ and (1.8). The latter allows to work without coordinates and benefit from properties of specific linear spaces $\mathcal{B}$, which could otherwise be obscured by an arbitrary choice of T and S.

2. Main results

Let $\mathbb{1}$ denote the function on $\mathcal{X}$ identically equal to 1. Assume that $\mathbb{1} \in \mathcal{B}$. This entails no restriction on the considered exponential families $e(\mathcal{B})$, but allows an elegant formulation of the criterion of existence of MLE in terms of $\mathcal{B}$, in fact in terms of the cone of nonnegative functions in $\mathcal{B}$:

$$\mathcal{B}_+ := \{\phi \in \mathcal{B} : \phi \geq 0\}.$$

We note in passing that Appendix A.6 gives a reformulation of our criterion for the existence of MLE without requiring that $\mathbb{1} \in \mathcal{B}$.

Let $U \subset \mathcal{X}$. We say that U is a *set of uniqueness* for $\mathcal{B}$ if $\phi = 0$ is the only function in $\mathcal{B}$ such that $\phi = 0$ on U. Similarly, we say that U is a *set of uniqueness* for $\mathcal{B}_+$ if $\phi = 0$ is the only function in $\mathcal{B}_+$ such that $\phi = 0$ on U. Put differently, U is of uniqueness for $\mathcal{B}_+$ if the conditions $\phi \in \mathcal{B}_+$ and $\phi = 0$ on U imply that $\phi = 0$ on $\mathcal{X}$. Of course, if U is a set of uniqueness for $\mathcal{B}$, then U is a set of uniqueness for $\mathcal{B}_+$.

Example 2.1. Let $\mathcal{X} = \{-2, -1, 0, 1, 2\} \subset \mathbf{R}$. Let $\mathcal{B}$ denote the class of all real functions on $\mathcal{X}$ that are of the form $a + bx$ on $\{-2, -1, 0\}$ and $a + cx$ on $\{0, 1, 2\}$ with some $a, b, c \in \mathbf{R}$. Then $\{-1, 2\}$ is a set of uniqueness for $\mathcal{B}_+$ but $\{-2, 2\}$ is not. We also observe that $\{-1, 2\}$ is not a set of uniqueness for $\mathcal{B}$, so the nonnegativity of functions in $\mathcal{B}_+$ plays a role here.

Being a set of uniqueness is a monotone property in the sense that every set larger than a set of uniqueness is also of uniqueness. Furthermore, if U is a set of uniqueness for $\mathcal{B}_+$ and $\mathcal{A}$ is a linear subspace of $\mathcal{B}$, then U is of uniqueness for $\mathcal{A}_+$.

The following is a crucial definition: For $U \subset \mathcal{X}$ and $\phi \in \mathcal{B}$ we let

$$\lambda_U(\phi) = \max_{\mathcal{X}} \phi - \min_U \phi.$$

Here is our characterization of the existence of MLE for discrete exponential families.

Theorem 2.2. *MLE for $e(\mathcal{B})$ and $x_1, \ldots, x_n \in \mathcal{X}$ exists if and only if $\{x_1, \ldots, x_n\}$ is of uniqueness for $\mathcal{B}_+$.*

Proof: Let us start with the "only if" part. If $U = \{x_1, \ldots, x_n\}$ is not a set of uniqueness for $\mathcal{B}_+$, then there is a nonzero function $f \in \mathcal{B}_+$ such that $f(x_1) = \ldots = f(x_n) = 0$. Let $\phi \in \mathcal{B}$ be arbitrary. Let $\varphi = \phi - f$. We have $\bar{\varphi} = \bar{\phi}$, but $\psi(\varphi) < \psi(\phi)$, where ψ is defined in (1.1). So, by (1.4), $l_{e(\phi)}(x_1, \ldots, x_n) < l_{e(\varphi)}(x_1, \ldots, x_n)$. Therefore no $\phi \in \mathcal{B}$ is MLE for $x_1, \ldots, x_n$.

To prove the other implication, we let U be a set of uniqueness for $\mathcal{B}_+$. By (1.4) for $\varphi \in \mathcal{B}$,

$$l_{e(\varphi)}(x_1,\dots,x_n) = n\left(\overline{\varphi} - \psi(\varphi)\right) \le n\left(\frac{1}{n}\left(\min_U \varphi + (n-1)\max_{\mathcal{X}} \varphi\right) - \psi(\varphi)\right).$$

Let $C = \min_{x\in\mathcal{X}} \log \mu(x)$. By (1.6), (1.5),

$$\begin{aligned} l_{e(\varphi)}(x_1,\dots,x_n) &\le \min_U \varphi + (n-1)\max_{\mathcal{X}} \varphi - n\max_{\mathcal{X}} \varphi - nC \\ &= -\lambda_U(\varphi) - nC \to -\infty, \end{aligned}$$

as $\lambda_U(\varphi) \to \infty$. By Lemma A.1, $\lambda_U(\varphi) \to \infty$ if $\lambda_{\mathcal{X}}(\varphi) \to \infty$. In particular, there exists $M > 0$ such that if $\lambda_{\mathcal{X}}(\varphi) > M$, then

$$l_{e(\varphi)}(x_1,\dots,x_n) < l_{e(0)}(x_1,\dots,x_n) = -n\log\mu(\mathcal{X}).$$

By (1.5) and continuity, the maximum of $l_{e(\varphi)}(x_1,\dots,x_n)$ is attained on the compact set $\{\varphi \in \mathcal{B} : 0 \le \varphi \le M\}$. □

The above proof is different from that of Bogdan and Bogdan (2000, Theorem 2.3), Barndorff-Nielsen (1978, Theorem 9.13) and Sullivant (2018, Theorem 8.2.1); the use of λ_U makes our arguments more direct.

Remark 2.3. By Theorem 2.2 we see that the existence of MLE depends on the sequence $(x_1,\dots,x_n)$ only through the set $\{x_1,\dots,x_n\}$. Furthermore, the existence of MLE does not depend on μ, i.e., we may take constant μ without loosing generality. Summarizing, the *existence* of MLE depends only on $\mathcal{B}$ and the *set* $\{x_1,\dots,x_n\}$. Of course, the actual MLE, say $\widehat{p}$, does depend on the *sequence* $(x_1,\dots,x_n)$, the weight μ and $\mathcal{B}$.

2.1. *Nonexistence of MLE.* In this section we elaborate on the case of nonexistence of MLE in the spirit of Geyer (1990). To this end we fix $x_1,\dots,x_n \in \mathcal{X}$ and assume that there is a nontrivial $\delta \in \mathcal{B}_+$ such that $\delta(x_1) = \ldots = \delta(x_n) = 0$. By Theorem 2.2, $\sup_{p\in e(\mathcal{B})} l_p(x_1,\dots,x_n)$ is not attained at any $p \in e(\mathcal{B})$. However, the supremum is "attained at infinity", in fact for an exponential density on a proper subset of the state space $\mathcal{X}$. Indeed, fix δ as above. If $\varphi \in \mathcal{B}$ and $k \in (0,\infty)$, then

$$l_{e(\varphi)}(x_1,\dots,x_n) \le l_{e(\varphi-k\delta)}(x_1,\dots,x_n),$$

see the first part of the proof of Theorem 2.2. Furthermore,

$$\psi(\varphi - k\delta) \to \log \sum_{x\in\mathcal{X}:\delta(x)=0} e^{\varphi(x)}\mu(x), \quad \text{as } k \to \infty. \tag{2.1}$$

We let $\widetilde{\mathcal{X}} = \{x \in \mathcal{X} : \delta(x) = 0\}$ and carrying on with the notation for $\widetilde{\mathcal{X}}$ we obtain measure $\widetilde{\mu}$, linear space $\widetilde{\mathcal{B}}$ with cone $\widetilde{\mathcal{B}}_+$, log-partition function $\widetilde{\psi}$, likelihood function $\widetilde{L}$, log-likelihood function $\widetilde{l}$ and exponential family $e(\widetilde{\mathcal{B}})$. Put simpler, we discard $\{x \in \mathcal{X} : \delta(x) > 0\}$ and achieve the following reduction.

Lemma 2.4. $\sup_{\widetilde{p}\in e(\widetilde{\mathcal{B}})} \widetilde{l}_{\widetilde{p}}(x_1,\dots,x_n) = \sup_{p\in e(\mathcal{B})} l_p(x_1,\dots,x_n)$.

Proof: For $\phi \in \mathcal{B}$ we let $\widetilde{\phi} = \phi|_{\widetilde{\mathcal{X}}}$. Since $\{x_1,\dots,x_n\} \subset \widetilde{\mathcal{X}}$,

$$\overline{\widetilde{\phi}} = \frac{1}{n}\sum_{i=1}^n \widetilde{\phi}(x_i) = \frac{1}{n}\sum_{i=1}^n \phi(x_i) = \overline{\phi}. \tag{2.2}$$

Furthermore,

$$\psi(\phi) = \log\left(\sum_{x\in\mathcal{X}} e^{\phi(x)}\mu(x)\right) \geq \log\left(\sum_{x\in\widetilde{\mathcal{X}}} e^{\phi(x)}\mu(x)\right) = \widetilde{\psi}(\widetilde{\phi}).$$

Thus $\overline{\phi} - \psi(\phi) \leq \overline{\widetilde{\phi}} - \widetilde{\psi}(\widetilde{\phi})$, and so

$$\sup_{p\in e(\mathcal{B})} l_p(x_1,\ldots,x_n) \leq \sup_{\widetilde{p}\in e(\widetilde{\mathcal{B}})} \widetilde{l}_{\widetilde{p}}(x_1,\ldots,x_n).$$

Let $\delta \in \mathcal{B}_+$ and k be as in (2.1). Using (2.1) and (2.2),

$$l_{e(\phi-k\delta)}(x_1,\ldots,x_n) \to \widetilde{l}_{e(\widetilde{\phi})}(x_1,\ldots,x_n), \qquad \text{as } k\to\infty.$$

Therefore,

$$\sup_{p\in e(\mathcal{B})} l_p(x_1,\ldots,x_n) \geq \sup_{\widetilde{p}\in e(\widetilde{\mathcal{B}})} \widetilde{l}_{\widetilde{p}}(x_1,\ldots,x_n).$$

□

Motivated by Lemma 2.4, we define

$$\{x_1,\ldots,x_n\}_{\mathcal{B}_+} = \bigcap \phi^{-1}(\{0\}), \tag{2.3}$$

where the intersection is taken over all $\phi \in \mathcal{B}_+$ such that $\phi(x_1) = \ldots = \phi(x_n) = 0$. Thus for all $\phi \in \mathcal{B}_+$, if ϕ vanishes on $\{x_1,\ldots,x_n\}$, then it vanishes on $\{x_1,\ldots,x_n\}_{\mathcal{B}_+}$, and the latter is the largest such set. Put differently, if there is $\delta \in \mathcal{B}_+$ such that $\delta(x_1) = \ldots = \delta(x_n) = 0$ but $\delta(x) > 0$, then $x \notin \{x_1,\ldots,x_n\}_{\mathcal{B}_+}$, and conversely. In particular, $U \subset \mathcal{X}$ is set of uniqueness for $\mathcal{B}_+$ if and only if $U_{\mathcal{B}_+} = \mathcal{X}$.

Example 2.5. In the setting of Example 2.1 we have $\{-2\}_{\mathcal{B}_+} = \{-2\}$ and $\{-1\}_{\mathcal{B}_+} = \{-2,-1,0\}$.

We note that if $x \notin \{x_1,\ldots,x_n\}_{\mathcal{B}_+}$, then there is $\phi \in \mathcal{B}_+$ such that $\phi = 0$ on $\{x_1,\ldots,x_n\}$ but $\phi(x) > 0$. Since $\mathcal{X}$ is finite, by adding such functions we can construct $\delta \in \mathcal{B}_+$ that vanishes precisely on $\{x_1,\ldots,x_n\}_{\mathcal{B}_+}$, i.e., $\delta^{-1}(\{0\}) = \{x_1,\ldots,x_n\}_{\mathcal{B}_+}$. We adopt the setting of Lemma 2.4 with this δ, in particular with $\widetilde{\mathcal{X}} = \{x_1,\ldots,x_n\}_{\mathcal{B}_+}$, and we get the following result.

Theorem 2.6. *There is a unique* $\widetilde{p} \in e(\widetilde{\mathcal{B}})$ *such that* $\widetilde{l}_{\widetilde{p}}(x_1,\ldots,x_n) = \sup_{p\in e(\mathcal{B})} l_p(x_1,\ldots,x_n)$.

Proof: By the definition of $\{x_1,\ldots,x_n\}_{\mathcal{B}_+}$ and by Theorem 2.2, Lemma 1.1 and 2.4, there is a unique $\widetilde{p} \in e(\widetilde{\mathcal{B}})$ such that

$$\widetilde{l}_{\widetilde{p}}(x_1,\ldots,x_n) = \sup_{\hat{p}\in e(\widetilde{\mathcal{B}})} \widetilde{l}_{\hat{p}}(x_1,\ldots,x_n) = \sup_{p\in e(\mathcal{B})} l_p(x_1,\ldots,x_n).$$

□

Example 2.7. For the first sample in Example 1.2 we get $\tilde{\mathcal{X}} = \{x_1\}_{\mathcal{B}_+} = \{1\}$, and $\tilde{p} = 1$ on $\tilde{\mathcal{X}}$.

For more substantial applications of Theorem 2.6 we refer to Example 3.2 and Example 3.9.

2.2. *Linear programming.* Before we address special spaces $\mathcal{B}$, we offer the reader a down-to-earth perspective. To start with, by a comment at the beginning of Section 2, we get the following simple result.

Corollary 2.8. *If $\{x_1, \ldots, x_n\}$ is of uniqueness for $\mathcal{B}$ then MLE exists for $\mathrm{e}(\mathcal{B})$ and $x_1, \ldots, x_n$.*

Notably, the condition in Corollary 2.8 may be verified by solving the following linear problem:

$$\begin{gathered}\phi \in \mathcal{B},\\ \phi(x_1) = \ldots = \phi(x_n) = 0.\end{gathered}$$

Indeed, $\{x_1, \ldots, x_n\}$ is of uniqueness for $\mathcal{B}$ if and only if the homogeneous linear system has only the trivial solution. In contrast, Theorem 2.2 is a linear programming problem. Indeed, $\{x_1, \ldots, x_n\}$ is of uniqueness for $\mathcal{B}_+$ if and only if the supremum of the (objective) function $\sum_{x\in\mathcal{X}} \phi(x)$ is zero for the class of functions satisfying

$$\begin{gathered}\phi \in \mathcal{B},\\ \phi(x_1) = \ldots = \phi(x_n) = 0,\\ \phi \geq 0.\end{gathered}$$

In this vein Rinaldo et al. (2009, Appendix C) observe that the condition of Barndorff-Nielsen is actually a linear programming problem and make connections to the geometry (of the convex hull of the set S in Section 1.2). The linear programming also occurs in the study of the closures of convex exponential families Geyer (1990) or binary logistic regression models Konis (2007). Furthermore, Wang et al. (2019) consider the linear programming in the case when MLE fails to exist. See also Sullivant (2018) for further information on linear programming and cases of nonexistence of MLE for discrete exponential families. Since the linear programming in general runs in polynomial time, see Schrijver (1986), it should be the method of choice when verifying the existence of MLE for discrete exponential families and data at hand. Having said this, for special linear spaces $\mathcal{B}$ one can come across interesting mathematics, as we demonstrate below. We also remark in passing that the linear problem in Corollary 2.8 is the Haar condition of Crain (1976) in our setting. Quite generally, the sufficient Haar condition of Crain for the existence of MLE is in the uniqueness of a linear problem while our necessary and sufficient condition is in the uniqueness of a linear-programming problem. The latter is still computationally manageable but more subtle (and optimal); see also the last sentence in Example 2.1 for a difference between these two conditions in a very simple setting.

3. Applications

Maximization of likelihood is fundamental in estimation, model selection and testing. In many procedures it is important to know if MLE actually exists for given data $x_1, \ldots, x_n$ and the linear space of exponents $\mathcal{B}$; see Fienberg and Rinaldo (2012, Introduction) for a list of such problems. Fienberg and Rinaldo (2012) interpret the existence of MLE by using the geometry of the polyhedral cone spanned by the rows of a specific design matrix. This result is connected with the criterion of Barndorff-Nielsen (1978). They also inquire which parameters are estimable when MLE is missing.

Below we show that the notion of the set of uniqueness is useful in characterizing the existence of MLE in discrete exponential families for specific spaces $\mathcal{B}$. There are two types of results we propose:

(1) conditions for the existence of MLE for a given sample,
(2) probability bounds for the existence of MLE for independent identically distributed samples.

To this end let $\mathcal{X}$ and $\mathcal{B}$ be as in Section 1.1. Let $X_1, X_2, \dots$ be *i.i.d.* random variables with values in $\mathcal{X}$. We define the random (stopping) time

$$\nu_{\text{uniq}} = \inf\{n \geq 1 : \{X_1, \dots, X_n\} \text{ is a set of uniqueness for } \mathcal{B}_+\}.$$

We will estimate tails of the distribution of ν_{uniq} in terms of $\mathcal{X}$, $\mathcal{B}$ and n. Typically we are interested in uniformly distributed X_i's: $\mathbb{P}(X_i = x) = 1/K$, $x \in \mathcal{X}$, $i = 1, 2, \dots$, where $K = |\mathcal{X}|$. In the setting of Theorem 2.2 we consider $\mathcal{B} = \mathbf{R}^{\mathcal{X}}$. We fix arbitrary $\mu > 0$ on $\mathcal{X}$, see Remark 2.3. Here is a trivial observation.

Lemma 3.1. *MLE for $e(\mathbf{R}^{\mathcal{X}})$ and $x_1, \dots, x_n$ exists if and only if $\{x_1, \dots, x_n\} = \mathcal{X}$.*

Proof: By Theorem 2.2 it is enough to verify that $\mathcal{X}$ is the only set of uniqueness for $\mathbf{R}_+^{\mathcal{X}}$. Obviously, $\mathcal{X}$ is a set of uniqueness for $\mathbf{R}_+^{\mathcal{X}}$ (in fact for $\mathbf{R}^{\mathcal{X}}$). On the other hand, if $U \subset \mathcal{X}$ and $x_0 \in \mathcal{X} \setminus U$, then $\mathbb{1}_{x_0}$ vanishes on U but not on $\mathcal{X}$, hence U is not of uniqueness for $\mathbf{R}_+^{\mathcal{X}}$ (neither it is for $\mathbf{R}^{\mathcal{X}}$). □

Example 3.2. Using notation of Section 2.1, we have $U_{\mathcal{B}_+} = U$, for every $U \subset \mathcal{X}$. Clearly, $U \subset U_{\mathcal{B}_+}$. On the other hand, using Equation (2.3), one may observe that for every $x \notin U$ the function $\phi(x) = \mathbb{1}_{\{x\}} \in \mathcal{B}_+$ and $\phi = \{0\}$ on U, but $x \notin \phi^{-1}(\{0\})$, so $U_{\mathcal{B}_+} \subset U$. In particular, $\{x_1, \dots, x_n\}_{\mathcal{B}_+} = \{x_1, \dots, x_n\}$ is the new state space $\widetilde{\mathcal{X}}$.

Later on we give examples which use the full strength of Theorem 2.2 and the nonnegativity of functions in $\mathcal{B}_+$ therein. For now we propose a probabilistic consequence of Lemma 3.1.

Corollary 3.3. *Let $\mathcal{B} = \mathbf{R}^{\mathcal{X}}$ and $K = |\mathcal{X}|$. Let $X_1, X_2, \dots$ be independent random variables, each with uniform distribution on $\mathcal{X}$. Then, for every $c \in \mathbf{R}$,*

$$\lim_{K \to \infty} \mathbb{P}\left(\nu_{uniq} < K \log K + Kc\right) = e^{-e^{-c}}.$$

Proof: Let $\nu_{\mathcal{X}} = \inf\{n \geq 1 : \{X_1, \dots, X_n\} = \mathcal{X}\}$. The random variable $\nu_{\mathcal{X}}$ yields a connection to the classical Coupon Collector Problem, see Erdős and Rényi (1961), and Pósfai (2010). Namely, by Erdős and Rényi (1961),

$$\lim_{K \to \infty} \mathbb{P}\left(\nu_{\mathcal{X}} < K \log K + Kc\right) = e^{-e^{-c}}.$$

By Lemma 3.1, $\nu_{\mathcal{X}} = \nu_{\text{uniq}}$, and the proof is complete. □

We aim to cover with large probability the whole of $\mathcal{X}$ by a sample of suitable size depending on K.

Corollary 3.4. *Let $\varepsilon \in (0, 1)$, $K = |\mathcal{X}|$ and $\mathcal{B} = \mathbf{R}^{\mathcal{X}}$. Let $X_1, X_2, \dots$ be independent random variables, each with uniform distribution on $\mathcal{X}$. If $K \to \infty$, then*

$$\mathbb{P}\left(\nu_{uniq} < (1-\varepsilon) K \log K\right) \to 0 \quad \text{and} \quad \mathbb{P}\left(\nu_{uniq} < (1+\varepsilon) K \log K\right) \to 1. \tag{3.1}$$

Proof: By Lemma 3.1 and Corollary 3.3, for every $c \in \mathbf{R}$ we get

$$\begin{aligned}\limsup_{K \to \infty} \mathbb{P}\left(\nu_{\text{uniq}} < (1-\varepsilon) K \log K\right) &\leq \limsup_{K \to \infty} \mathbb{P}\left(\nu_{\text{uniq}} < K \log K + Kc\right) \\ &= e^{-e^{-c}}.\end{aligned}$$

Thus $\lim_{K\to\infty} \mathbb{P}\left(\nu_{\mathrm{uniq}} < (1-\varepsilon)\, K \log K\right) = 0$. The second part of (3.1) is obtained analogously. $\square$

Remark 3.5. We summarize (3.1) by saying that $K \log K$ is a *sharp threshold* of the sample size for the existence of MLE for $e(\mathbf{R}^{\mathcal{X}})$ and uniform *i.i.d.* samples. Sharp thresholds are widely used in the theory of random graphs, see Erdős and Rényi (1960, Equation 3). It is also convenient to use them here to indicate the minimal size of *i.i.d.* samples that guarantees the existence of MLE with high probability.

3.1. *Rademacher functions.* For $k \in \mathbf{N}$, let us consider $\mathcal{X} = Q_k := \{-1,1\}^k$, the k-dimensional discrete cube with, say, the uniform weight $\mu(\chi) = 2^{-k}$, $\chi \in Q_k$ (but see Remark 2.3). Thus, $K = |\mathcal{X}| = 2^k$. For $j = 1,\ldots,k$ and $\chi = (\chi_1,\ldots,\chi_k) \in Q_k$ we define the Rademacher functions:

$$r_j(\chi) = \chi_j,$$

and we denote $r_0(\chi) = 1$. Let

$$\mathcal{B}^k = \mathrm{Lin}\{r_0, r_1, \ldots, r_k\}.$$

We define, as usual, the exponential family

$$e(\mathcal{B}^k) = \{e(r) : r \in \mathcal{B}^k\}.$$

Theorem 3.6. *MLE for $e(\mathcal{B}^k)$ and $x_1,\ldots,x_n \in Q_k$ exists if and only if for all $j = 1,\ldots,k$ we have $\{r_j(x_1),\ldots,r_j(x_n)\} = \{-1,1\}$.*

Proof: By Theorem 2.2 we only need to prove that the above condition characterizes the sets of uniqueness for $\mathcal{B}^k_+$. If $j \in \{1,\ldots,k\}$ is such that $r_j(x_1) = \ldots = r_j(x_n) = 1$, then we let $r = r_0 - r_j$. Clearly, $r \in \mathcal{B}^k_+$ and r is not identically zero, but $r(x_i) = 0$ for all $i = 1,\ldots,n$. Thus, $\{x_1,\ldots,x_n\}$ is not a set of uniqueness for $\mathcal{B}^k_+$. Similarly, if $r_j(x_1) = \ldots = r_j(x_n) = -1$, then we consider the function $r = r_0 + r_j \in \mathcal{B}^k_+$. For the converse implication we consider arbitrary

$$r = \sum_{j=0}^{k} a_j r_j \in \mathcal{B}^k_+.$$

Let $\chi = -(\mathrm{sign}(a_1),\ldots,\mathrm{sign}(a_k))$, where, say, $\mathrm{sign}(0) = 1$. Obviously, $\chi \in Q_k$, and since $r(\chi) \geq 0$, we get

$$a_0 \geq \sum_{j=1}^{k} |a_j|. \tag{3.2}$$

Assume that $r = 0$ on $\{x_1,\ldots,x_n\}$. Let $j \in \{1,\ldots,k\}$. There are $x, x' \in \{x_1,\ldots,x_n\}$ such that $r_j(x) = 1$ and $r_j(x') = -1$. We have

$$0 = r(x) + r(x') = 2a_0 + \sum_{i\neq j} a_i[r_i(x) + r_i(x')].$$

It follows that

$$a_0 \leq \sum_{i\neq j} |a_i|.$$

By (3.2), $a_j = 0$, for every $j \geq 1$. Thereby $a_0 = 0$ and $r \equiv 0$. We see that $\{x_1,\ldots,x_n\}$ is a set of uniqueness for $\mathcal{B}^k_+$. $\square$

Example 3.7. Let $x \in Q_k$ be arbitrary. By Theorem 3.6, MLE for $e\left(\mathcal{B}^k\right)$ and $\{x, -x\}$ exists.

We define the *positive* and *negative half-cubes*, respectively:

$$H_j^+ = \{\chi \in Q_k : r_j(\chi) = 1\}, \quad H_j^- = \{\chi \in Q_k : r_j(\chi) = -1\}, \quad j = 1, \dots, k. \tag{3.3}$$

We note that $\mathcal{B}^k$ is also spanned by the indicator functions of half-cubes, namely $\mathbb{1}_j^+ = (r_0 + r_j)/2$ and $\mathbb{1}_j^- = (r_0 - r_j)/2$, $j = 1, \dots, k$.

Corollary 3.8. *MLE for $e(\mathcal{B}^k)$ and $x_1, \dots, x_n \in Q_k$ exists if and only if $\{x_1, \dots, x_n\}$ has a nonempty intersection with each half-cube.*

The proof of Corollary 3.8 is immediate from Theorem 3.6 and the discussion above.

Example 3.9. If MLE fails to exist for $e(\mathcal{B}^k)$ and $x_1, \dots, x_n \in Q_k$, then the following analysis may shed some light on Theorem 2.6. Let

$$J = \{j \in \{1, \dots, k\} : \{r_j(x_1), \dots, r_j(x_n)\} = \{-1, 1\}\}, \quad J' = \{1, \dots, k\} \setminus J.$$

Since we consider the case when MLE does not exist, by Theorem 3.6, $J' \neq \emptyset$. For $j \in J'$ we let

$$H_j = \{\chi \in Q_k : r_j(\chi) = r_j(x_1) = \ldots = r_j(x_n)\}.$$

Clearly, this is a half-cube, see (3.3). We will show that

$$\{x_1, \dots, x_n\}_{\mathcal{B}_+^k} = \bigcap_{j \in J'} H_j. \tag{3.4}$$

We note that for $j \in J'$, r_j is constant on the right-hand side of (3.4). Accordingly, the right-hand side of (3.4) is isomorphic to $\{-1, 1\}^{|J|}$ or to $Q_{|J|}$.

Now if $r = \sum_{j=0}^k a_j r_j \in \mathcal{B}_+^k$ and $r(x_1) = \ldots = r(x_n) = 0$, then $r = \sum_{j \in J} a_j r_j + c \geq 0$ on $\{-1, 1\}^{|J|}$, where $c = a_0 + \sum_{j \in J'} a_j r_j(x_1)$ is the sum of terms which are constant on $\bigcap_{j \in J'} H_j$. In the case when $J = \emptyset$, it is obvious that $\{x_1, \dots, x_n\}_{\mathcal{B}_+^k} = \bigcap_{j \in J'} H_j = \{x_1\}$, since $x_1 = \ldots = x_n$. However, if $J \neq \emptyset$, then by definition of J and Theorem 3.6 with $k = |J|$, $r = 0$ on $\bigcap_{j \in J'} H_j$. Thus $\bigcap_{j \in J'} H_j \subset \{x_1, \dots, x_n\}_{\mathcal{B}_+^k}$. On the other hand, we observe that for each $j \in J'$, $\mathbb{1}_{H_j^c} = 0$ on the sample and $\mathbb{1}_{H_j^c} > 0$ on H_j^c, hence $H_j^c \cap \{x_1, \dots, x_n\}_{\mathcal{B}_+^k} = \emptyset$ and $\{x_1, \dots, x_n\}_{\mathcal{B}_+^k} \subset \bigcap_{j \in J'} H_j$.

By Theorem 2.6, MLE exists for $e(\widetilde{\mathcal{B}}^k)$ and $x_1, \dots, x_n$ with the measure $\widetilde{\mu} := \mu|_{\widetilde{\mathcal{X}}}$ on $\widetilde{\mathcal{X}} := \bigcap_{j \in J'} H_j$. Of course, $\widetilde{X}$ is isomorphic with $Q_{|J|}$, if we ignore the J' coordinates of the points in $\widetilde{X}$. In this way we may also think that $\widetilde{\mu}$ and $x_1, \dots, x_n$ are on $Q_{|J|}$. Thus, one may calculate the supremum of the log-likelihood function for $e(\mathcal{B}^k)$, $x_1, \dots, x_n$ and μ as the maximum of a log-likelihood function on $Q_{|J|}$. Of course, the total mass of $\widetilde{\mu}$ is a fraction of that of μ. For instance, if μ is the uniform probability weight on Q_k then $\widetilde{\mu}$ is uniform with the total mass $2^{-|J'|}$, which adds $n|J'| \log 2$ to the log-likelihood that would be obtained for $Q_{|J|}$ with the uniform probability weight, see, e.g., (1.2).

Here is a probabilistic application of Theorem 3.6.

Corollary 3.10. *Let $k \in \mathbf{N}$ and $X_1, X_2, \ldots, X_n$ be independent random variables, each with uniform distribution on Q_k. Then,*

$$\mathbb{P}\left(\textit{MLE exists for } e(\mathcal{B}^k) \textit{ and } X_1, \ldots, X_n\right) = \left(1 - \frac{1}{2^{n-1}}\right)^k \geq 1 - \frac{k}{2^{n-1}} \to 1, \ \textit{as } n \to \infty.$$

Proof: We have $\mathbb{P}(X_i = x) = 2^{-k}$ for all $x \in Q_k$ and $i = 1, \ldots, n$. We let $R_{ij} = r_j(X_i)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, k$. Thus, $\mathbb{P}(R_{ij} = 1) = \mathbb{P}(R_{ij} = -1) = \frac{1}{2}$ and $\{R_{ij}\}_{i,j}$ are independent. By Theorem 3.6,

$$\begin{aligned}&\mathbb{P}\left(\text{MLE exists for } e(\mathcal{B}^k) \text{ and } X_1, \ldots, X_n\right)\\ =&\mathbb{P}\big(\{R_{ij} : i = 1, \ldots, n\} = \{-1, 1\} \text{ for } j = 1, \ldots, k\big) = \left(1 - \frac{2}{2^n}\right)^k.\end{aligned}$$

Applying the Bernoulli inequality finishes the proof. □

Corollary 3.11. *For $k \in \mathbf{N}$ let $X_1, \ldots, X_{n(k)}$ be independent random variables, each with uniform distribution on Q_k. If $n(k) = \log_2 k + b + o(1)$ for some $b \in \mathbf{R}$ as $k \to \infty$, then*

$$\lim_{k\to\infty} \mathbb{P}\left(\textit{MLE exists for } e(\mathcal{B}^k) \textit{ and } X_1, \ldots, X_{n(k)}\right) = e^{-2^{1-b}}.$$

Proof: By Corollary 3.10,

$$\mathbb{P}\left(\text{MLE exists for } e(\mathcal{B}^k) \text{ and } X_1, \ldots, X_{n(k)}\right) = \left(1 - \frac{1}{k\ 2^{b-1+o(1)}}\right)^k \to e^{-2^{1-b}}, \ \text{as } k \to \infty. \tag{3.5}$$

□

Corollary 3.12. $\log_2 k$ *is a sharp threshold of the sample size for the existence of MLE for $e(\mathcal{B}^k)$ and i.i.d. uniform samples on Q_k.*

Proof: Let $\varepsilon \in (0, 1)$ and (the sample size) $n = n(k) < (1 - \varepsilon) \log_2 k$. Then,

$$\mathbb{P}(\nu_{\text{uniq}} < n) \leq \mathbb{P}(\nu_{\text{uniq}} < (1 - \varepsilon) \log_2 k).$$

For every $b \in \mathbf{R}$ by the equation in (3.5) we have

$$\limsup_{k\to\infty} \mathbb{P}(\nu_{\text{uniq}} < (1 - \varepsilon) \log_2 k) \leq \limsup_{k\to\infty} \mathbb{P}(\nu_{\text{uniq}} < \log_2 k + b) = e^{-2^{1-b}}.$$

Since b is arbitrary, we conclude that $\limsup_{k\to\infty} \mathbb{P}(\nu_{\text{uniq}} < n(k)) = 0$. Analogously, for the sample size $n = n(k) > (1 + \varepsilon) \log_2 k$ we get

$$\liminf_{k\to\infty} \mathbb{P}(\nu_{\text{uniq}} > n(k)) = 1,$$

which ends the proof. □

The above is in stark contrast to Corollary 3.4, as summarized in Remark 3.5. Indeed, in the present setting we have $K = |Q_k| = 2^k$, so the sharp threshold for the sample size needed for the existence of MLE is $\log_2 \log_2 K$. The following result on the expectation of ν_{uniq} agrees well with the sharp threshold.

Lemma 3.13. *Let ν_{uniq} be as in Corollary 3.11. Let $H_k = \sum_{i=1}^{k} \frac{1}{k}$ be the k-th harmonic number. Then,*

$$\frac{H_k}{\log 2} + 1 \leq \mathbb{E}(\nu_{uniq}) < \frac{H_k}{\log 2} + 2, \quad k = 1, 2, \ldots.$$

Proof: Observe that $\nu_{\mathrm{uniq}} = \max\{\tau_1, \ldots, \tau_k\}$, where

$$\tau_j = \min\{n \geq 1 : \{r_j(X_1), \ldots, r_j(X_n)\} = \{-1, 1\}\}, \quad j = 1, \ldots, k.$$

From the fact that $X_1, X_2, \ldots$ are independent and uniformly distributed, we deduce that

$$\mathbb{1}_{r_j(X_i) \neq r_j(X_1)}, \quad i = 2, 3, \ldots, \quad j = 1, 2 \ldots,$$

are independent with symmetric Bernoulli distribution. Then $\tau_1, \ldots, \tau_k$ are independent, and

$$\tau_j + 1 \sim \mathit{Geom}\,(1/2)$$

for $j = 1, \ldots, k$. The result follows from Eisenberg (2008). □

In Section 5 we return to Rademacher functions, but for now we turn to exponential families of random graphs, a major motivation for this work.

4. Random graphs

In this section we focus on random graphs. Their various applications can be found in Rinaldo et al. (2009), Schweinberger et al. (2020) and Mukherjee et al. (2018). What is important for us, many such models are indeed discrete exponential families. As usual, maximum likelihood can be used to select a suitable graph model within the exponential family, see, e.g., Pitman (1979, Chapter 1 and 8) and Bezáková et al. (2006). In this section we characterize the existence of MLE in such context. The theory of random graphs started with probabilistic proofs of the existence or nonexistence of specific graphs by Erdős, see, e.g., Bollobás (1998). Asymptotic properties of random graphs were developed in the seminal papers of Erdős and Rényi (1959, 1960) and Gilbert (1959). Rinaldo et al. (2009) discuss geometric interpretations of the existence of MLE for discrete exponential families with applications to random graphs and social networks. Chatterjee and Diaconis (2013) give normalizing constants that are crucial for the computation of MLE for exponential random graph models. Furthermore, they include examples when MLE fails to exist. The same authors together with Sly discuss in Chatterjee et al. (2011) the asymptotic probability of the existence and uniqueness of MLE for the β-model of graphs. This allows to connect the β-model with a random uniform model of graphs with a given degree sequence, which is then explored using graphons (graph limits, see Lovász and Szegedy (2006)). They also present an algorithm for the computation of MLE in the β-model.

Perry and Wolfe (2012) put nonasymptotic conditions for the existence of MLE in various random graph models parameterized by vertex-specific parameters. Rinaldo et al. (2013) characterize the existence of MLE for β-models. They interpret the Barndorff-Nielsen's criterion using the geometry of multidimensional polytopes of vertex-degree sequences, see also Fienberg and Rinaldo (2012). Wang et al. (2019) transfer the criterion into discrete hierarchical models, using the notion of simplicial complices. These models include, e.g., graphical models and Ising models. Wang, Rauh and Massam also improve the approximation of the set of estimable parameters in the case of the nonexistence of MLE, which is discussed in the setting of marginal polytopes.

Let us start with the notation. Graph is a pair $G = (V, E)$, where $V = \{1, \dots, N\}, N \in \mathbf{N}$, is the set of nodes and E is the set of edges, i.e.,

$$E \subset \binom{V}{2} := \{(r,s) : 1 \le r < s \le N\}.$$

We only consider simple undirected graphs (containing no loops or multiple edges). Let $m = |E|$. If $m = \binom{N}{2}$, then the graph is called complete and is denoted as K_N. On the other hand, the empty graph (with $m = 0$) is denoted as $\overline{K_N}$. For graphs $G = (V, E_1)$ and $H = (V, E_2)$ we let, as usual,

$$G \cup H := (V, E_1 \cup E_2), \qquad G \cap H := (V, E_1 \cap E_2).$$

Furthermore, $G \subset H$ means that $E_1 \subset E_2$. Let $\mathcal{G}_N$ be the family of all the graphs with N nodes, i.e., with $V = \{1, \dots, N\}$. By a *random graph* we understand a random variable $\mathbb{G}$ with values in $\mathcal{G}_N$. The families of distributions of such random variables are called *random graph models*. We focus on the exponential model of random graphs $\mathcal{G}_{N,c}$ defined as follows.

For $1 \le r < s \le N$ and $G \in \mathcal{G}_N$, we let

$$\mathbb{1}_G(r,s) = \begin{cases} 1, & \text{if } (r,s) \in E, \\ 0, & \text{otherwise.} \end{cases}$$

We define $\chi_{r,s} : \mathcal{G}_N \to \{-1, 1\}$ by $\chi_{r,s}(G) = 1 - 2\mathbb{1}_G(r,s)$. We consider the linear space

$$\mathcal{B}^{\mathcal{G}_N} = \mathrm{Lin}\left\{ 1, \chi_{r,s}(G) : 1 \le r < s \le N \right\}.$$

Let $c \in \mathbf{R}^{\binom{V}{2}}$ be a corresponding vector of coefficients. Following the setting of Section 1.1 we let $\mu(G) = 1$ for each $G \in \mathcal{G}_N$ (but see Remark 2.3) and consider the exponential family

$$\mathcal{G}_{N,c} := e(\mathcal{B}^{\mathcal{G}_N}) = \left\{ p_c := e^{\phi_c - \psi(\phi_c)} : c \in \mathbf{R}^{\binom{V}{2}} \right\}, \tag{4.1}$$

where

$$\phi_c(G) = \sum_{(r,s) \in \binom{V}{2}} c_{r,s} \chi_{r,s}(G), \qquad \psi(\phi_c) = \log \sum_{G \in \mathcal{G}_N} e^{\phi_c(G)},$$

for $G \in \mathcal{G}_N$, see also (1.3). As usual, for $p_c \in \mathcal{G}_{N,c}$ we let $L_{p_c}(G_1, \dots, G_n) = \prod_{i=1}^n p_c(G_i)$, etc.

Lemma 4.1. *Let $c \in \mathbf{R}^{\binom{V}{2}}$ and let $\mathbb{G}$ be a random graph with distribution $\mathcal{G}_{N,c}$. Let $1 \le r < s \le N$. Then the probability of the appearance of the edge (r,s) in $\mathbb{G}$ equals*

$$p_{r,s} = \frac{e^{c_{r,s}}}{1 + e^{c_{r,s}}}. \tag{4.2}$$

The result is well known but for convenience a proof is given in Appendix A.3.

Lemma 4.2. *Let $c \in \mathbf{R}^{\binom{V}{2}}$ and let $\mathbb{G}$ be a random graph with distribution $\mathcal{G}_{N,c}$. Let $1 \le r_1, s_1, r_2, s_2 \le N$, $r_1 < s_1, r_2 < s_2$, and $(r_1, s_1) \neq (r_2, s_2)$. Then the appearances of edges (r_1, s_1) and (r_2, s_2) in $\mathbb{G}$ are independent events.*

The proof of the result is similar to that of Lemma 4.1, and can be found in Appendix A.4. For instance, if $p_{r,s} = p \in (0,1)$ for every edge (r,s), then the exponential random graph with distribution $\mathcal{G}_{N,c}$ is the Erdős-Rényi random graph $\mathcal{G}_{N,p}$ in Erdős and Rényi (1959, 1960). The latter means that $\mathbb{P}(e \in E(\mathbb{G})) = p$ for every edge $e \in \binom{V}{2}$, and the events $e \in E(\mathbb{G})$ and $f \in E(\mathbb{G})$ are independent for different edges e, f.

Theorem 4.3. *MLE for $e(\mathcal{B}^{\mathcal{G}_N})$ and $G_1,\dots,G_n \in \mathcal{G}_N$ exists if and only if*

$$\bigcup_{i=1}^{n} G_i = K_N \qquad \text{and} \qquad \bigcap_{i=1}^{n} G_i = \overline{K_N}.$$

Proof: By Theorem **2.2**, MLE exists if and only if $\{G_1,\dots,G_n\}$ is of uniqueness for $\mathcal{B}_+^{\mathcal{G}_N}$.

We first prove the "only if" part of Theorem 4.3. Let us assume that there exists an edge $(r_0,s_0) \notin \bigcup_{i=1}^n G_i$. Then the function $\chi_{r_0,s_0} \in \mathcal{B}_+^{\mathcal{G}_N}$ equals zero on $G_1,\dots,G_n$, but not on the whole $\mathcal{G}_N$. In addition, if there is an edge $(r_0,s_0) \in \bigcap_{i=1}^n G_i$, then the function $(1+\chi_{r_0,s_0}) \in \mathcal{B}_+^{\mathcal{G}_N}$ vanishes for $G_1,\dots,G_n$, but it is not equal to zero, e.g., for the graph $\overline{K_N}$.

We next prove the 'if' part of the theorem. Let $\phi = k_0 + \sum_{r<s} k_{r,s}\chi_{r,s} \in \mathcal{B}_+^{\mathcal{G}_N}$, where $k_0, k_{r,s} \in \mathbf{R}$ for all $1 \le r < s \le N$. Since $\phi(G) \ge 0$ for every $G \in \mathcal{G}_N$,

$$k_0 \ge \sum_{r<s} |k_{r,s}|. \tag{4.3}$$

Let $(r_0,s_0) \in \binom{V}{2}$. Let $\phi(G_1) = \ldots = \phi(G_n) = 0$. Since $\bigcup_{i=1}^n G_i = K_N$ and $\bigcap_{i=1}^n G_i = \overline{K_N}$, there exists a pair of graphs $G', G'' \in \{G_1,\dots,G_n\}$ such that $\chi_{r_0,s_0}(G') = 1$, $\chi_{r_0,s_0}(G'') = -1$. Therefore,

$$\begin{aligned} 0 = \phi(G') + \phi(G'') &= 2k_0 + \sum_{r<s} k_{r,s}\left(\chi_{r,s}(G') + \chi_{r,s}(G'')\right) \\ &= 2k_0 + \sum_{\substack{r<s \\ (r,s)\neq(r_0,s_0)}} k_{r,s}\left(\chi_{r,s}(G') + \chi_{r,s}(G'')\right). \end{aligned}$$

It follows that $k_0 \le \sum_{(r,s)\neq(r_0,s_0)} |k_{r,s}|$ and eventually we get $k_{r_0,s_0} = 0$, thanks to (4.3). Since (r_0,s_0) is arbitrary, $k_{r,s} = 0$ for every $1 \le r < s \le N$. Then also $c_0 = 0$, and thus $\phi \equiv 0$. □

In the above random graph model it is possible to compute explicitly the probability of the existence of MLE for *i.i.d.* samples of graphs in $\mathcal{G}_N$. To this end, for $1 \le r < s \le N$ we fix $c_{r,s} \in \mathbf{R}$. By Lemma 4.1 the probability of the appearance of the edge (r,s) in random graph $\mathbb{G}$ with distribution $\mathcal{G}_{N,c}$ is

$$p_{r,s} = \frac{e^{c_{r,s}}}{1+e^{c_{r,s}}}.$$

Lemma 4.4. *Let $\{\mathbb{G}_1,\dots,\mathbb{G}_n\}$ be i.i.d. with distribution $\mathcal{G}_{N,c}$. Then the probability of the existence of MLE for $e(\mathcal{B}^{\mathcal{G}_N})$ equals*

$$\prod_{1\le r<s\le N} \left(1 - p_{r,s}^n - (1-p_{r,s})^n\right). \tag{4.4}$$

Proof: By Theorem **4.3**, MLE for $e(\mathcal{B}^{\mathcal{G}_N})$ exists if and only if among the random graphs $\mathbb{G}_1,\dots,\mathbb{G}_n$ every edge (r,s), $1 \le r < s \le N$, appears at least once, but not n times. For every edge (r,s) the above condition is satisfied with probability $1-(1-p_{r,s})^n - (p_{r,s})^n$. The independence of the occurrences of different edges in $\mathcal{G}_{N,c}$ yields the product (4.4). □

In particular, if $c = 0$, then the probability of the existence of MLE for $e(\mathcal{B}^{\mathcal{G}_N})$ equals

$$\left(1-2^{1-n}\right)^{\binom{N}{2}},$$

which is an analogue of Corollary 3.11. From the above results we can deduce asymptotic bounds for the *i.i.d.* sample size for which MLE exists with high probability. To this end we recall the classical result on $p = p(N) \in (0,1)$ such that $\mathbb{G}$ from $\mathcal{G}_{N,p}$ has at least one edge with high probability.

Remark 4.5. Frieze and Karoński (2016, Lemma 1.10) Let $\mathbb{G}_{N,p(N)}$ be a random graph with distribution $\mathcal{G}_{N,p(N)}$. Then

$$\lim_{N\to\infty} \mathbb{P}\left(\mathbb{G}_{N,p(N)} \text{ has at least one edge}\right) = \begin{cases} 0 & \text{if} \quad p(N) = o\left(N^{-2}\right), \\ 1 & \text{if} \quad N^{-2} = o\left(p(N)\right). \end{cases}$$

The above may be summarized by saying that N^{-2} is a *threshold* for the probability p such that $\mathbb{G}$ with distribution $\mathcal{G}_{N,p}$ has at least one edge. For more information on threshold functions in the theory of random graphs see, e.g., Frieze and Karoński (2016). In particular, a sharp threshold (mentioned previously) is a threshold but the converse is not true in general.

Lemma 4.6. *Let $\mathbb{G}_1, \ldots, \mathbb{G}_n$ be i.i.d. random variables with distribution $\mathcal{G}_{N,c}$. Then* $\log N$ *is a threshold of the sample size n for the existence of MLE for $e(\mathcal{B}^{\mathcal{G}_N})$.*

Proof: According to Lemma 4.4, the probability of the existence of MLE for $e(\mathcal{B}^{\mathcal{G}_N})$ and $\mathbb{G}_1, \ldots, \mathbb{G}_n$ equals

$$P_{\text{MLE}} = \prod_{1\le r<s\le N} \left(1 - p_{r,s}^n - (1-p_{r,s})^n\right).$$

We define the function

$$f(x) = 1 - x^w - (1-x)^w, \qquad x \in (0,1),\ w \ge 2. \tag{4.5}$$

Clearly, $f(x) = f(1-x)$ and for $w \ge 2$ we have f increasing when $0 < x < \frac{1}{2}$ and decreasing when $\frac{1}{2} < x < 1$. Using (4.5) we can bound P_{MLE} from above by

$$P_{\text{BIG}} := \left(1 - 2^{1-n}\right)^{\binom{N}{2}}.$$

Applying Corollary 3.10 and the equality in (3.5) for $k = \binom{N}{2}$, we observe that for every $b \in \mathbf{R}$ and for $n = n(N) = \log_2 \binom{N}{2} + b + o(1)$ we have $P_{\text{BIG}} \to e^{-2^{1-b}}$, as $N \to \infty$. Therefore, for $n(N) = o(\log N)$ we obtain $P_{\text{MLE}} \le P_{\text{BIG}} \to 0$, as $N \to \infty$.

We consider the sample size $n = n(N)$ (depending on N). We will prove that if $\log N/n \to 0$ as $N \to \infty$, then $P_{\text{MLE}} \to 1$. To this end we bound P_{MLE} from below by

$$P_{\text{SMALL}} := \left(1 - p_{\max}^n - (1-p_{\max})^n\right)^{\binom{N}{2}},$$

where $c_{\max} = \max_{1\le r<s\le N} |c_{r,s}|$ and $p_{\max} = e^{c_{\max}}/(1+e^{c_{\max}})$.

Take n independent Erdős-Rényi random graphs $\mathbb{H}_1, \ldots, \mathbb{H}_n$ with distribution $\mathcal{G}_{N,p_{\max}}$. Then the probability of the existence of MLE for $e(\mathcal{B}^{\mathcal{G}_N})$ and for $\mathbb{H}_1, \ldots, \mathbb{H}_n$ equals exactly P_{SMALL}. Note that intersection and union of the graphs are also Erdős-Rényi random graphs, namely

$$\bigcap_{i=1}^n \mathbb{H}_i \sim \mathcal{G}_{N,p_{\max}^n}, \qquad \bigcup_{i=1}^n \mathbb{H}_i = \overline{\bigcap_{i=1}^n \overline{\mathbb{H}_i}} \sim \mathcal{G}_{N,1-q_{\max}^n},$$

where

$$q_{\max} := 1 - p_{\max} = \frac{e^{-c_{\max}}}{1+e^{-c_{\max}}}.$$

From Remark 4.5, with high probability we have

$$\bigcap_{i=1}^{n} \mathbb{H}_i = \overline{K_N} \qquad \text{and} \qquad \overline{\bigcup_{i=1}^{n} \mathbb{H}_i} = \overline{K_N},$$

provided

$$p_{\max}^n = o(N^{-2}) \qquad \text{and} \qquad q_{\max}^n = o(N^{-2}).$$

By definition, $c_{\max} > 0$, so $p_{\max} > q_{\max}$. In order to get $P_{\text{SMALL}} \to 1$ as $n \to \infty$, it suffices to have $p_{\max}^n = o(N^{-2})$. If $n(N)/\log N \to \infty$ as $N \to \infty$, then the above condition is satisfied. Therefore $\log N$ is a threshold of the sample size for existence of MLE for $e(\mathcal{B}^{\mathcal{G}_N})$ and independent $\mathbb{G}_1, \ldots, \mathbb{G}_n$ from $\mathcal{G}_{N,c}$. $\square$

5. Applications to Walsh functions

We return to Rademacher functions to discuss the spaces spanned by their products. Let $k \in \mathbf{N}$, $1 \le q \le k$, and

$$\mathcal{B}_q^k = \text{Lin}\left\{ w_S : S \subset \{1, \ldots, k\} \text{ and } |S| \le q \right\},$$

where

$$w_S(x) = \prod_{i \in S} r_i(x), \quad x \in Q_k, \quad S \subset \{1, \ldots, k\},$$

are the Walsh functions, see, e.g., Jendrej et al. (2015).

The case $\mathcal{B}_1^k = \mathcal{B}^k$ was discussed in Section 3.1 and the case $q = 2$ is related to the Ising model of ferromagnetism in statistical mechanics, see Wainwright and Jordan (2008, Example 3.1).

Lemma 5.1. *The dimension of the linear space $\mathcal{B}_q^k$ is $\sum_{j=0}^{q} \binom{k}{j}$.*

The proof of Lemma 5.1 is given in Appendix A.5.

Corollary 5.2. *For $q \le \frac{k}{2}$ we have*

$$dim\left(\mathcal{B}_q^k\right) \le 2^{kH_2(\frac{q}{k})} \le \left(\frac{ek}{q}\right)^q,$$

where $H_2(p) = -p\log_2 p - (1-p)\log_2(1-p)$ is the binary entropy function.

The proof follows from Lemma 5.1 and the entropy bound for the sum of binomial coefficients, see, e.g., Galvin (2014, Theorem 3.1).

Characterization of the existence of MLE for $e(\mathcal{B}_q^k)$ and the related sharp thresholds seem to be hard for general q, even for $q = 2$, see Remark 5.4. In the next section we discuss the products of $k - q$ Rademacher functions for fixed $q \in \mathbf{N}$ ($q \le k$). We especially focus on the products of $k - 1$ and k Rademacher functions. Below we characterize the existence of MLE for $e(\mathcal{B}_{k-1}^k)$. As we will see, we get a qualitatively different result than that in Section 3.1. Let $\mathcal{E}$ and $\mathcal{O}$ be the sets of all those points in Q_k that have an even and odd number of positive coordinates, respectively.

Theorem 5.3. *MLE exists for $e(\mathcal{B}_{k-1}^k)$ and $x_1, \ldots, x_n \in Q_k$ if and only if $\mathcal{E}$ or $\mathcal{O} \subset \{x_1, \ldots, x_n\}$.*

Proof: Thanks to Theorem 2.2, we only need to characterize the sets of uniqueness for $\left(\mathcal{B}^k_{k-1}\right)_+$. To this end, we consider the hypercube G_{Q_k}, defined as the graph with vertices in Q_k and edges between all pairs of points which differ by exactly one coordinate. Thus,

$$V(G_{Q_k}) = Q_k \text{ and } E(G_{Q_k}) = \{\{x,y\} \in Q_k \times Q_k : |\{j : r_j(x) \neq r_j(y)\}| = 1\}.$$

Let $U = \{x_1, \ldots, x_n\}$. Assume that U is a set of uniqueness. Let $e \in \mathcal{E}$ and $o \in \mathcal{O}$. The hypercube graph G_{Q_k} is connected, so there exists a path $(e, v_1, v_2, \ldots, v_{2p}, o)$ in G_{Q_k}. Then

$$\left(\mathbb{1}_{\{e,v_1\}} + \mathbb{1}_{\{v_2,v_3\}} + \ldots + \mathbb{1}_{\{v_{2p},o\}}\right) - \left(\mathbb{1}_{\{v_1,v_2\}} + \mathbb{1}_{\{v_3,v_4\}} + \ldots + \mathbb{1}_{\{v_{2p-1},v_{2p}\}}\right) \tag{5.1}$$
$$= \mathbb{1}_{\{e\}} + \mathbb{1}_{\{o\}} \tag{5.2}$$

is a nontrivial nonnegative function on Q_k. Therefore, we must have $\{e, o\} \cap U \neq \emptyset$. Then we easily conclude that $\mathcal{E} \subset U$ or $\mathcal{O} \subset U$.

For the converse implication, we consider $q \in \{0, \ldots, k\}$ and $(k-q)$-subcubes defined as follows,

$$\bigcap_{i=1}^{q} H_{j_i}, \tag{5.3}$$

where $1 \le j_1 < j_2 < \ldots < j_q \le k$ and $H_{j_i} = H^+_{j_i}$ or $H^-_{j_i}$, see (3.3). When $q = k-1$, the intersection, or a 1-cube, is a pair of points in Q_k which differ by exactly one coordinate, so they have a different parity. Moreover, each such pair can be obtained in this way. Using (5.3), as in the proof of Lemma 5.1 we see that $\mathbb{1}_{\{e,o\}} \in \mathcal{B}^k_{k-1}$ for each $e \in \mathcal{E}$ and $o \in \mathcal{O}$. Furthermore, each q-subcube of Q_k with $q \ge 1$ can be covered by disjoint pairs $\{e, o\}$ as above. Therefore, the functions $\mathbb{1}_{\{e,o\}} \in \mathcal{B}^k_{k-1}$ with $e \in \mathcal{E}$ and $o \in \mathcal{O}$ span the linear space $\mathcal{B}^k_{k-1}$.

We next claim that for every $f \in \mathcal{B}^k_{k-1}$,

$$\sum_{x \in \mathcal{O}} f(x) = \sum_{x \in \mathcal{E}} f(x). \tag{5.4}$$

Indeed, if $f = \mathbb{1}_{\{e,o\}}$ with $e \in \mathcal{E}$ and $o \in \mathcal{O}$, then the equality is true because both sides of (5.4) are equal to 1. Since such functions span $\mathcal{B}^k_{k-1}$ it follows that (5.4) is true for every $f \in \mathcal{B}^k_{k-1}$.

Finally, if nonnegative $f \in \mathcal{B}^k_{k-1}$ vanishes on $\mathcal{E}$, then the sum over $\mathcal{O}$ also equals zero, hence $f \equiv 0$, and the same conclusion holds if we assume that $f = 0$ on $\mathcal{O}$. Thus U is the set of uniqueness if $\mathcal{O} \subset U$ or $\mathcal{E} \subset U$. □

Remark 5.4. A naïve extension of Corollary 3.8 fails for $e(\mathcal{B}^k_2)$, if we try to replace the half-cubes with $(k-2)$-subcubes, that is, quarter-cubes. This is seen from Theorem 5.3 for $k = 3$. Indeed, the set

$$\{(1,1,-1),(1,-1,1),(-1,1,1),(1,-1,-1),(-1,1,-1),(-1,-1,1)\}$$

is not of uniqueness for $(\mathcal{B}^3_2)_+$, as follows from (5.1) with $e = (-1,-1,-1)$ and $o = (1,1,1)$, even though the set has nonmpty intersection with each quarter-cube.

We will briefly treat the case of $e(\mathcal{B}^k_k)$, as follows.

Corollary 5.5. *$k2^k \log 2$ is a sharp threshold of the sample size for the existence of MLE for $e(\mathcal{B}^k_k)$ and i.i.d. samples uniform on Q_k.*

Proof: Observe that $e(\mathcal{B}_k^k)$ is isomorphic to $e(\mathbf{R}^{\mathcal{X}})$ for $|\mathcal{X}| = 2^k$. The existence of MLE for $e(\mathcal{B}_k^k)$ is characterized in (more general) Lemma 3.1, and the sharp threshold is given after Corollary 3.4. □

Corollary 5.5 is in stark contrast with the result for the (smaller) space $e(\mathcal{B}_1^k)$ because for $e(\mathcal{B}_1^k)$ the sharp threshold, and so the threshold, equal $\log_2 k$, by Corollary 3.12.

Remark 5.6. Let $1 \le q_1 \le q_2 \le k$. Then every set U of uniqueness for $(\mathcal{B}_{q_2}^k)_+$ is of uniqueness for $(\mathcal{B}_{q_1}^k)_+$, because $(\mathcal{B}_{q_1}^k)_+ \subset (\mathcal{B}_{q_2}^k)_+$.

A characterization of the existence of MLE for $e(\mathcal{B}_q^k)$ for arbitrary q, even for $q = 2$, turned out to be difficult. Accordingly, we do not give a sharp threshold for the size of the uniform *i.i.d.* sample needed for the existence of MLE for $e(\mathcal{B}_q^k)$. However, the case of $e(\mathcal{B}_{k-q}^k)$ seems a little easier in the sense that we are able to give the less precise threshold for the existence of MLE for $e(\mathcal{B}_{k-q}^k)$. Moreover, for each fixed q the threshold for $e(\mathcal{B}_{k-q}^k)$ is the same as for $e(\mathcal{B}_k^k)$, namely $k2^k$ as $k \to \infty$.

Lemma 5.7. *Fix $q \in \mathbf{N}$. Then $k2^k$ is a threshold of the sample size for the existence of MLE for $e(\mathcal{B}_{k-q}^k)$ and i.i.d. sample uniform on Q_k.*

Proof: If $\lim_{k\to\infty} n(k)/(k2^k) = \infty$, then by Remark 5.6 and Corollary 5.5, for $k \to \infty$ we get

$$\begin{aligned}&\mathbb{P}\left(\left\{X_1, \ldots, X_{n(k)}\right\} \text{ is of uniqueness for } \left(\mathcal{B}_{k-q}^k\right)_+\right)\\ &\ge \mathbb{P}\left(\left\{X_1, \ldots, X_{n(k)}\right\} \text{ is of uniqueness for } \mathcal{B}_k^k\right) \to 1,\end{aligned}$$

as needed. On the other hand, every set U of uniqueness for $(\mathcal{B}_{k-q}^k)_+$ must intersect with every subcube defined by fixing last $k-q$ coordinates, because each q-subcube is the support of a function in $(\mathcal{B}_{k-q}^k)_+$, to wit, of its indicator. There are 2^{k-q} such q-subcubes, each of which we can suggestively denote by $(*, \ldots, *, \varepsilon_{q+1}, \ldots, \varepsilon_k)$, where $\varepsilon_{q+1}, \ldots, \varepsilon_k = \pm 1$. Observe that the family of the above subcubes is a partition of Q_k. We consider each q-subcube as a coupon in the Coupon Collector Problem. If a sample point falls into the q-subcube, we consider the coupon as collected. The probability of collecting a given coupon is 2^{q-k}. Therefore, if $n(k) = o\left(2^k k\right)$, hence $n(k) = o\left(2^{k-q}(k-q)\right)$, then

$$\mathbb{P}\left(\left\{X_1, \ldots, X_{n(k)}\right\} \text{ is of uniqueness for } (\mathcal{B}_{k-q}^k)_+\right) \to 0, \quad \text{as } k \to \infty,$$

as needed. □

Appendix A. Appendix

A.1. *Proof of Lemma 1.1.* Let $\hat{p} = e(\phi_0), \widetilde{p} = e(\phi_1) \in e(\mathcal{B})$ and $\widehat{p} \ne \widetilde{p}$, so that $\phi_1 - \phi_0 \ne const$. Let $\phi_t = \phi_0 + t(\phi_1 - \phi_0)$, $p_t = e(\phi_t)$ for $t \in \mathbf{R}$ and $l(t) = l_{p_t}(x_1, \ldots, x_n)$. We claim that l is strictly concave, that is $l'' < 0$. Indeed, since $\overline{\phi_t} = \overline{\phi_0} + t\overline{\phi_1}$ is a linear function, by (1.4) we get

$$l''(t) = -n\frac{d^2}{dt^2}\log Z(\phi_t).$$

Let X be a random variable with values in $\mathcal{X}$ such that $\mathbb{P}(X = x) = p(x)\mu(x)$. As usual, for every $f : \mathcal{X} \to \mathbf{R}$ we have

$$\mathbb{E}f(X) = \sum_{x\in\mathcal{X}} f(x)p(x)\mu(x).$$

Clearly, $(\log Z(\phi_t))' = \frac{Z(\phi_t)'}{Z(\phi_t)}$ and $(\log Z(\phi_t))'' = \frac{Z(\phi_t)''}{Z(\phi_t)} - \left(\frac{Z(\phi_t)'}{Z(\phi_t)}\right)^2$. Hence, thanks to (1.1),

$$Z(\phi_t)' = \sum_{x\in\mathcal{X}} e^{\phi_t(x)}\mu(x)\left(\phi_1(x) - \phi_0(x)\right)$$
$$Z(\phi_t)'' = \sum_{x\in\mathcal{X}} e^{\phi_t(x)}\mu(x)\left(\phi_1(x) - \phi_0(x)\right)^2.$$

Thus,

$$\frac{Z(\phi_t)'}{Z(\phi_t)} = \mathbb{E}[\phi_1(X) - \phi_0(X)] \qquad\qquad \frac{Z(\phi_t)''}{Z(\phi_t)} = \mathbb{E}[\phi_1(X) - \phi_0(X)]^2$$

and so

$$\frac{d^2}{dt^2}\log Z(\phi_t) = \mathbb{E}\left[\phi_1(X) - \phi_0(X) - \mathbb{E}(\phi_1(X) - \phi_0(X))\right]^2 > 0,$$

since $\phi_1 - \phi_0$ is not constant. Hence, l is strictly concave, in particular $l(1/2) > (l(0)+l(1))/2$. If $\sup_{p\in e(\mathcal{B})} L_p(x_1,\ldots,x_n) = L_{\widehat{p}}(x_1,\ldots,x_n) = L_{\widetilde{p}}(x_1,\ldots,x_n)$, then $l(1/2) > \sup_{p\in e(\mathcal{B})} l_p(x_1,\ldots,x_n)$, which is absurd; thus at most one of $\widetilde{p}$ and $\widehat{p}$ can be the MLE.

A.2. *Control by oscillations.* λ_U defined in Section 2 may be thought of as a specific measure of oscillation of ϕ. Of course, $\lambda_U \ge 0$. Furthermore, for every $c \in \mathbf{R}$,

$$\lambda_U(\phi + c) = \lambda_U(\phi), \quad \phi \in \mathcal{B}, \tag{A.1}$$

and for every (positive number) $k > 0$ we have (homogeneity),

$$\lambda_U(k\phi) = k\lambda_U(\phi), \quad \phi \in \mathcal{B}, k \ge 0. \tag{A.2}$$

If $U = \mathcal{X}$, then $\lambda_{\mathcal{X}}(-\phi) = \lambda_{\mathcal{X}}(\phi)$ for $\phi \in \mathcal{B}$, and so $\lambda_{\mathcal{X}}$ is a seminorm. Clearly, $\lambda_U \le \lambda_{\mathcal{X}}$. However, if there is a nontrivial $\phi \in \mathcal{B}_+$ such that $\phi = 0$ on U, then $\lambda_U(\phi) = \sup_{\mathcal{X}} \phi > 0$ but $\lambda_U(-\phi) = 0$. The following result is the engine of Theorem 2.2.

Lemma A.1. *$U \subset \mathcal{X}$ is the set of uniqueness for $\mathcal{B}_+$ if and only if λ_U is comparable with $\lambda_{\mathcal{X}}$ on $\mathcal{B}$, i.e., there exist constants $c_1, c_2 > 0$ such that $c_1\lambda_{\mathcal{X}}(\phi) \le \lambda_U(\phi) \le \lambda_{\mathcal{X}}(\phi)$ for all $\phi \in \mathcal{B}$.*

Proof: We first prove the "if" part. Assume U is not a set of uniqueness for $\mathcal{B}_+$. Then there exists a nonzero function $\phi \in \mathcal{B}_+$ such that $\phi = 0$ on U. We have $\lambda_U(-\phi) = 0$ and $\lambda_{\mathcal{X}}(-\phi) > 0$, hence λ_U and $\lambda_{\mathcal{X}}$ are not comparable on $\mathcal{B}$.

We now prove the "only if" part, which is delicate. For all $\vartheta, \phi \in \mathcal{B}$ we have

$$\begin{aligned}\lambda_U(\vartheta + \phi) &\le \max_{\mathcal{X}} \vartheta + \max_{\mathcal{X}} \phi - \min_U \vartheta - \min_U \phi \\ &= \lambda_U(\vartheta) + \lambda_U(\phi) \le \lambda_U(\vartheta) + \lambda_{\mathcal{X}}(\phi).\end{aligned}$$

It follows that $\lambda_U(\vartheta) \ge \lambda_U(\vartheta - \phi) - \lambda_{\mathcal{X}}(\phi)$, hence

$$\lambda_U(\vartheta + \phi) \ge \lambda_U(\vartheta) - \lambda_{\mathcal{X}}(\phi).$$

Therefore, $|\lambda_U(\vartheta + \phi) - \lambda_U(\vartheta)| \le \lambda_{\mathcal{X}}(\phi)$. As a consequence, λ_U is continuous on $\mathcal{B}$.

We will prove that there is a number $h > 0$ such that $\lambda_U(\phi) \ge h\lambda_{\mathcal{X}}(\phi)$ for every $\phi \in \mathcal{B}$. Let $\mathcal{S} = \{\phi \in \mathcal{B} : \min_{\mathcal{X}} \phi = 0 \text{ and } \max_{\mathcal{X}} \phi = 1\}$. Let $\phi \in \mathcal{S}$. If $\lambda_U(\phi) = 0$, then $\phi = 1$ on U.

Consider $\varphi = 1 - \phi$. Clearly, $\varphi \geq 0$ and $\varphi = 0$ on U. It follows that $\varphi = 0$ on $\mathcal{X}$, because U is of uniqueness. Then $\phi \equiv 1$, which contradicts the assumption $\phi \in \mathcal{S}$. Therefore, $\lambda_U(\phi) > 0$. Since $\mathcal{S}$ is compact and λ_U is continuous, $h := \min_{\mathcal{S}} \lambda_U > 0$. By (A.2) and (A.1) we obtain $\lambda_U(\phi) \geq h\lambda_{\mathcal{X}}(\phi)$ for all $\phi \in \mathcal{B}$. □

A.3. *Proof of Lemma 4.1.* By (4.1), each $G \in \mathcal{G}_N$ appears in $\mathcal{G}_{N,c}$ with probability $p_c(G) = e^{\phi_c(G)-\psi(\phi_c)}$. Then,

$$p_{r,s} = \mathbb{P}\left((r,s) \in E\left(\mathbb{G}\right)\right) = \sum_{\substack{G \in \mathcal{G}_N \\ (r,s) \in E(G)}} \frac{e^{\phi_c(G)}}{\sum_{G \in \mathcal{G}_N} e^{\phi_c(G)}}$$

$$= \frac{\sum_{\substack{G \in \mathcal{G}_N \\ (r,s) \in E(G)}} e^{\phi_c(G)}}{\sum_{\substack{G \in \mathcal{G}_N \\ (r,s) \in E(G)}} e^{\phi_c(G)} + \sum_{\substack{G \in \mathcal{G}_N \\ (r,s) \notin E(G)}} e^{\phi_c(G)}}$$

$$= \frac{\sum_{\substack{G \in \mathcal{G}_N \\ (r,s) \in E(G)}} e^{\sum_{(k,l) \in \binom{V}{2}} c_{k,l}\chi_{k,l}(G)}}{\sum_{\substack{G \in \mathcal{G}_N \\ (r,s) \in E(G)}} e^{\sum_{(k,l) \in \binom{V}{2}} c_{k,l}\chi_{k,l}(G)} + \sum_{\substack{G \in \mathcal{G}_N \\ (r,s) \notin E(G)}} e^{\sum_{(k,l) \in \binom{V}{2}} c_{k,l}\chi_{k,l}(G)}}. \tag{A.3}$$

Note that

$$\sum_{(k,l) \in \binom{V}{2}} c_{k,l}\chi_{k,l}(G) = c_{r,s}\chi_{r,s}(G) + C(G),$$

where

$$C(G) = \sum_{\substack{(k,l) \in \binom{V}{2} \\ (k,l) \neq (r,s)}} c_{k,l}\chi_{k,l}(G).$$

Therefore

$$e^{\sum_{(k,l) \in \binom{V}{2}} c_{k,l}\chi_{k,l}(G)} = e^{c_{r,s}\chi_{r,s}(G)}\ e^{C(G)}.$$

Clearly, $c_{r,s}\chi_{r,s}(G)$ is $c_{r,s}$ if $(r,s) \in E(G)$ and it is 0 if $(r,s) \notin E(G)$. Thus, (A.3) equals

$$\frac{e^{c_{r,s}} \sum_{\substack{G \in \mathcal{G}_N \\ (r,s) \in E(G)}} C(G)}{\sum_{\substack{G \in \mathcal{G}_N \\ (r,s) \in E(G)}} e^{C(G)} + e^{c_{r,s}} \sum_{\substack{G \in \mathcal{G}_N \\ (r,s) \notin E(G)}} e^{C(G)}}.$$

Let S be the graph with only one edge (r,s). The map $G \mapsto G \setminus S$ is a bijection between the graphs with the edge (r,s) and graphs without (r,s). In addition, $C(G) = C(G \setminus S)$, and so we get (4.2).

A.4. *Proof of Lemma 4.2.* By (4.1), each $G \in \mathcal{G}_N$ appears in $\mathcal{G}_{N,c}$ with probability $p_c(G) = e^{\phi_c(G)-\psi(\phi_c)}$. Then,

$$\mathbb{P}\left(\left(r_1,s_1\right),\left(r_2,s_2\right) \in E\left(\mathbb{G}\right)\right) = \sum_{\substack{G \in \mathcal{G}_N \\ (r_1,s_1),(r_2,s_2) \in E(G)}} \frac{e^{\phi_c(G)}}{\sum_{G \in \mathcal{G}_N} e^{\phi_c(G)}}.$$

As in the proof of Lemma 4.1, we observe that

$$\sum_{(k,l)\in\binom{V}{2}} c_{k,l}\chi_{k,l}(G) = c_{r_1,s_1}\chi_{r_1,s_1}(G) + c_{r_2,s_2}\chi_{r_2,s_2}(G) + \widetilde{C}(G),$$

where

$$\widetilde{C}(G) = \sum_{\substack{(k,l)\in\binom{V}{2}\\(k,l)\neq(r_1,s_1)\\(k,l)\neq(r_2,s_2)}} c_{k,l}\chi_{k,l}(G).$$

Thus,

$$e^{\sum_{(k,l)\in\binom{V}{2}} c_{k,l}\chi_{k,l}(G)} = e^{c_{r_1,s_1}\chi_{r_1,s_1}(G)}\ e^{c_{r_2,s_2}\chi_{r_2,s_2}(G)}\ e^{\widetilde{C}(G)}.$$

Let S_1 and S_2 be the graphs with only one edge, (r_1,s_1) and (r_2,s_2), respectively. Let

$$\begin{aligned}
\mathcal{G}_{N_{12}} &= \{G\in\mathcal{G}_N : S_1\subset G, S_2\subset G\},\\
\mathcal{G}_{N_{10}} &= \{G\in\mathcal{G}_N : S_1\subset G, S_2\not\subset G\},\\
\mathcal{G}_{N_{02}} &= \{G\in\mathcal{G}_N : S_1\not\subset G, S_2\subset G\},\\
\mathcal{G}_{N_{00}} &= \{G\in\mathcal{G}_N : S_1\not\subset G, S_2\not\subset G\}.
\end{aligned}$$

a partition of $\mathcal{G}_N$. We observe that the maps

$$G\mapsto G\setminus S_1,\quad G\mapsto G\setminus S_2,\quad G\mapsto G\setminus(S_1\cup S_2)$$

are bijections between $\mathcal{G}_{N_{10}}$, $\mathcal{G}_{N_{02}}$, $\mathcal{G}_{N_{12}}$, respectively, and $\mathcal{G}_{N_{00}}$. Also, for every $G\in\mathcal{G}_N$,

$$\widetilde{C}(G) = \widetilde{C}(G\setminus S_1) = \widetilde{C}(G\setminus S_2) = \widetilde{C}(G\setminus(S_1\cup S_2)).$$

Put differently, $\widetilde{C}(G)$ does not depend on the edges (r_1,s_1) and (r_2,s_2). As in the proof of Lemma 4.1, we obtain

$$\begin{aligned}
&\mathbb{P}\left((r_1,s_1),(r_2,s_2)\in E(\mathbb{G})\right)\\
&= \frac{e^{c_{r_1,s_1}}e^{c_{r_2,s_2}}}{1+e^{c_{r_1,s_1}}+e^{c_{r_2,s_2}}+e^{c_{r_1,s_1}}e^{c_{r_2,s_2}}} = p_{r_1,s_1}\ p_{r_2,s_2}.
\end{aligned}$$

A.5. *Proof of Lemma 5.1.*

Proof: Consider the positive half-cubes $H_1^+,\dots,H_k^+$. Let

$$\mathcal{B} = \mathrm{Lin}\left\{\prod_{i\in I_q}\mathbb{1}_{H_i^+} : I_q\subset\{0,\dots,k\} \text{ and } |I_q|\le q\right\}.$$

We have $\mathcal{B}=\mathcal{B}_q^k$, because $r_0=\mathbb{1}_{Q_k}$, $r_i = 2\mathbb{1}_{H_i^+}-\mathbb{1}_{Q_k}$ and by induction it is easy to see that for every $S\subset\{1,\dots,k\}$ and $|S|<q$, if Walsh function $w_S\in\mathcal{B}$ then their product with Rademacher function $w_S r_i\in\mathcal{B}$, for any $i=0,\dots,n$. Note that for any permutation σ of $\{1,2,\dots,q\}$,

$$\mathbb{1}_{H_{i_1}^+}\mathbb{1}_{H_{i_2}^+}\cdots\mathbb{1}_{H_{i_q}^+} = \mathbb{1}_{H_{i_{\sigma(1)}}^+}\mathbb{1}_{H_{i_{\sigma(2)}}^+}\cdots\mathbb{1}_{H_{i_{\sigma(q)}}^+}.$$

The functions $\mathbb{1}_{Q_k}$ and $\mathbb{1}_{H_{i_1}^+}\cdots\mathbb{1}_{H_{i_q}^+}$, $1\le i_1\le\ldots\le i_q\le k$, are linearly independent. Indeed, assume that

$$r := \alpha_0\mathbb{1}_{Q_k} + \sum_{i_1,\dots,i_q\in\{1,\dots,k\}}\alpha_{i_1\cdots i_q}\mathbb{1}_{H_{i_1}^+}\cdots\mathbb{1}_{H_{i_q}^+} = 0.$$

There are points $x_0 \in \bigcap_{i=1}^k H_i^-$, $x_{i_1} \dots x_{i_q} \in \bigcap_{l\in\{i_1,\dots,i_q\}} H_l^- \cap \bigcap_{l\neq i_1,\dots,i_q} H_l^-$ for each $1 \le i_1 \le i_2 \le \dots \le i_q \le k$. We obtain $\alpha_0 = r(x_0) = 0$ and $\alpha_{i_1\cdots i_q} = r(x_{i_1\cdots i_q}) = 0$ as needed. □

A.6. *Propagation of extrema, relative interior and the criterion of Barndorff-Nielsen.* In this section we give auxiliary results, but also explain connections to the criterion of Barndorff-Nielsen. Let $\mathcal{B}$ be an arbitrary linear subspace of $\mathbf{R}^{\mathcal{X}}$. In Corollary A.5 below we adapt the criterion in Theorem 2.2 to such $e(\mathcal{B})$. Let $\mathcal{B}'$ be the linear space spanned by $\mathcal{B}$ and $\mathbb{1}$.

Lemma A.2. *If $U \subset \mathcal{X}$, then $\phi = \min_{\mathcal{X}} \phi$ on U implies $\phi = \min_{\mathcal{X}} \phi$ on $\mathcal{X}$ for every $\phi \in \mathcal{B}$ if and only if $\phi = \max_{\mathcal{X}} \phi$ on U implies $\phi = \max_{\mathcal{X}} \phi$ on $\mathcal{X}$ for every $\phi \in \mathcal{B}$.*

Proof: The property with the minima is equivalent to the one with the maxima because $\mathcal{B}$ is closed upon multiplication by -1 and because $\max(-\phi) = -\min\phi$. □

Definition A.3. We say that $U \subset \mathcal{X}$ propagates extrema for $\mathcal{B}$ if $\phi = \inf_{\mathcal{X}} \phi$ on U implies that $\phi = \inf_{\mathcal{X}} \phi$ on $\mathcal{X}$ for every $\phi \in \mathcal{B}$.

Due to Lemma A.2, the property could be equivalently stated using maxima.

Lemma A.4. *A nonempty $U \subset \mathcal{X}$ propagates extrema for $\mathcal{B}$ if and only if U is of uniqueness for $\mathcal{B}'_+$.*

Proof: Assume that U is of uniqueness for $\mathcal{B}'_+$. Let $\phi \in \mathcal{B}$ and $\phi = \min_{\mathcal{X}} \phi$ on U. Then $\varphi = \phi - \min_{\mathcal{X}} \phi \in \mathcal{B}'_+$ and $\varphi = 0$ on U, so $\varphi = 0$ on $\mathcal{X}$ and $\phi = \min_{\mathcal{X}} \phi$ on $\mathcal{X}$. It follows that U propagates extrema for $\mathcal{B}$. Conversely, assume that U propagates extrema for $\mathcal{B}$. Let $\phi \in \mathcal{B}$. Then $\phi = \varphi + c$ for some $\varphi \in \mathcal{B}$ and $c \in \mathbf{R}$. If $\phi \ge 0$ and $\phi = 0$ on U, then $\varphi = \min_{\mathcal{X}} \varphi = -c$ on U, hence $\varphi = -c$ on $\mathcal{X}$, and so $\phi = 0$ on $\mathcal{X}$. Thus, U is of uniqueness for $\mathcal{B}'_+$. □

Theorem 2.2 yields the following.

Corollary A.5. *MLE for $e(\mathcal{B})$ and $x_1, ..., x_n \in \mathcal{X}$ exists if and only if $\{x_1, \dots, x_n\}$ propagates extrema for $\mathcal{B}$.*

Proof: The MLE for $e(\mathcal{B})$ and $e(\mathcal{B}')$ must be the same. Indeed, we have $e(\mathcal{B}) = e(\mathcal{B}')$ so the suprema of the likelihood functions are the same, see Section 1.1. Of course, if $\phi \in \mathcal{B}$ and $e(\phi)$ is the MLE for $e(\mathcal{B})$ then it is also the MLE for $e(\mathcal{B}')$. Conversely, if $\phi \in \mathcal{B}'$, then $\phi = \varphi + c$ for some $\varphi \in \mathcal{B}$ and $c \in \mathbf{R}$. If $e(\phi)$ is the MLE for $e(\mathcal{X}')$, then $e(\varphi)$ is the MLE for $e(\mathcal{B})$. Considering $\mathcal{B}'$, by Theorem 2.2 we see that MLE for $e(\mathcal{B}')$ and $x_1, ..., x_n \in \mathcal{X}$ exists if and only if $\{x_1, \dots, x_n\}$ is of uniqueness for $\mathcal{B}'_+$, and – by Lemma A.4 – if and only if $\{x_1, \dots, x_n\}$ propagates extrema for $\mathcal{B}$. □

The next lemma hinges on the trivial observation that if the sample mean equals the minimum, then the sample is constant.

Lemma A.6. $\{x_1, \dots, x_n\}$ *propagates extrema for $\mathcal{B}$ if and only if for every $\phi \in \mathcal{B}$,* $\min_{\mathcal{X}} \phi < \max_{\mathcal{X}} \phi$ *implies* $\min_{\mathcal{X}} \phi < \bar{\phi} < \max_{\mathcal{X}} \phi$.

Proof: Let $\{x_1, \dots, x_n\}$ propagate extrema for $\mathcal{B}$. If $\min_{\mathcal{X}} \phi = \bar{\phi}$, then $\phi = \min_{\mathcal{X}} \phi$ on $\{x_1, \dots, x_n\}$, hence $\phi = \min_{\mathcal{X}} \phi$ on $\mathcal{X}$ and so $\min_{\mathcal{X}} \phi = \max_{\mathcal{X}} \phi$. A similar argument works if $\bar{\phi} = \max_{\mathcal{X}} \phi$; see also Lemma A.2. Conversely, if $\{x_1, \dots, x_n\}$ does not propagate extrema for $\mathcal{B}$ then there is $\phi \in \mathcal{B}$ such that $\phi = \min_{\mathcal{X}} \phi$ on $\{x_1, \dots, x_n\}$, but $\max_{\mathcal{X}} \phi > \min_{\mathcal{X}} \phi$. Then $\min_{\mathcal{X}} \phi = \bar{\phi} < \max_{\mathcal{X}} \phi$. □

Recall the setting and notation of Section 1.2. The following theorem was essentially proved in Barndorff-Nielsen (1978, Theorem 9.13), except that it was stated for the minimal representation of exponential families. The formulation presented in Theorem A.7 below was given in Johansen (1979, Theorem 3.5), which covers the arbitrary canonical representation and does so with a more direct proof. Notably, Johansen (1979) uses the notion of relative interior of a convex set. Let C be the convex hull of S. We say that $t \in \mathbf{R}^d$ is in the relative interior of C if for every $\theta \in \mathbf{R}^d$, $\min_{y\in C} \theta \cdot y < \max_{y\in C} \theta \cdot y$ implies $\min_{y\in C} \theta \cdot y < \theta \cdot t < \max_{y\in C} \theta \cdot y$.

Theorem A.7. *Johansen (1979, Theorem 3.5.) MLE for $e(\mathcal{B})$ and $x_1, ..., x_n \in \mathcal{X}$ exists, if and only if $\bar{T}$ is in the relative interior of C.*

To close the circle of ideas, we give a self-contained proof of Theorem A.7, which may also be used to obtain Theorem 2.2 from Theorem A.7.

Proof of Theorem A.7: By the discussion in this section we know very well that MLE for $x_1, ..., x_n$ and $e(\mathcal{B})$ exists if and only if for every $\phi \in \mathcal{B}$, $\min_{\mathcal{X}} \phi < \max_{\mathcal{X}} \phi$ implies $\min_{\mathcal{X}} \phi < \bar{\phi} < \max_{\mathcal{X}} \phi$. Recall that $\phi \in \mathcal{B}$ if and only if there is $\theta \in \mathbf{R}^d$ such that $\phi = \theta \cdot T$. Then $\min_{x\in\mathcal{X}} \phi(x) = \min_{y\in\mathcal{S}} \theta \cdot y = \min_{y\in\mathcal{C}} \theta \cdot y$, $\max_{x\in\mathcal{X}} \phi(x) = \max_{y\in\mathcal{C}} \theta \cdot y$, and, of course, $\bar{\phi} = \theta \cdot \bar{T}$. Therefore the existence of MLE for $x_1, ..., x_n$ and $e(\mathcal{B})$ is equivalent to $\bar{T}$ being in the relative interior of C. □

For clarity, we recall that we agreed in Example 1.2 that the existence of MLE for $x_1, ..., x_n \in \mathcal{X}$ and $e(\mathcal{B})$ is the same as the existence of MLE for $x_1, ..., x_n$ and the exponential family given by the canonical statistics T and (1.10), and that it is equivalent to the existence of MLE for the sample $y_1 := T(x_1), ..., y_n = T(x_n) \in \mathbf{R}^d$ and the standard exponential family in (1.11). From the above discussion we also see that the convex hull C and the notion of relative interior are merely auxiliary objects to express the property in Lemma A.6, or the propagation of extrema property.

References

Albert, A. and Anderson, J. A. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, **71** (1), 1–10 (1984). MR738319.

Andersen, E. B. Sufficiency and exponential families for discrete sample spaces. *J. Amer. Statist. Assoc.*, **65**, 1248–1255 (1970). MR268992.

Barndorff-Nielsen, O. *Information and exponential families in statistical theory*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Ltd., Chichester (1978). ISBN 0-471-99545-2. MR489333.

Bezáková, I., Kalai, A., and Santhanam, R. Graph Model Selection Using Maximum Likelihood. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pp. 105–112. Association for Computing Machinery (2006). ISBN 1595933832. DOI: 10.1145/1143844.1143858.

Bogdan, K. and Bogdan, M. On existence of maximum likelihood estimators in exponential families. *Statistics*, **34** (2), 137–149 (2000). MR1768245.

Bogdan, M. and Ledwina, T. Testing uniformity via log-spline modeling. *Statistics*, **28** (2), 131–157 (1996). MR1405606.

Bollobás, B. *Modern graph theory*, volume 184 of *Graduate Texts in Mathematics*. Springer-Verlag, New York (1998). ISBN 0-387-98488-7. MR1633290.

Brown, L. D. *Fundamentals of statistical exponential families with applications in statistical decision theory*, volume 9 of *Institute of Mathematical Statistics Lecture Notes—Monograph Series*. Institute of Mathematical Statistics, Hayward, CA (1986). ISBN 0-940600-10-2. MR882001.

Chatterjee, S. and Diaconis, P. Estimating and understanding exponential random graph models. *Ann. Statist.*, **41** (5), 2428–2461 (2013). MR3127871.

Chatterjee, S., Diaconis, P., and Sly, A. Random graphs with a given degree sequence. *Ann. Appl. Probab.*, **21** (4), 1400–1435 (2011). MR2857452.

Crain, B. R. Estimation of distributions using orthogonal expansions. *Ann. Statist.*, **2**, 454–463 (1974). MR362678.

Crain, B. R. Exponential models, maximum likelihood estimation, and the Haar condition. *J. Amer. Statist. Assoc.*, **71** (355), 737–740 (1976). MR428559.

Darroch, J. N., Lauritzen, S. L., and Speed, T. P. Markov fields and log-linear interaction models for contingency tables. *Ann. Statist.*, **8** (3), 522–539 (1980). MR568718.

Diaconis, P. *Group representations in probability and statistics*, volume 11 of *Institute of Mathematical Statistics Lecture Notes—Monograph Series*. Institute of Mathematical Statistics, Hayward, CA (1988). ISBN 0-940600-14-5. MR964069.

Diaconis, P. and Freedman, D. Partial exchangeability and sufficiency. In *Statistics: applications and new directions (Calcutta, 1981)*, pp. 205–236. Indian Statist. Inst., Calcutta (1984). MR786142.

Eisenberg, B. On the expectation of the maximum of IID geometric random variables. *Statist. Probab. Lett.*, **78** (2), 135–143 (2008). MR2382066.

Erdős, P. and Rényi, A. On random graphs. I. *Publ. Math. Debrecen*, **6**, 290–297 (1959). MR120167.

Erdős, P. and Rényi, A. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, **5**, 17–61 (1960). MR125031.

Erdős, P. and Rényi, A. On a classical problem of probability theory. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, **6**, 215–220 (1961). MR150807.

Eriksson, N., Fienberg, S. E., Rinaldo, A., and Sullivant, S. Polyhedral conditions for the nonexistence of the MLE for hierarchical log-linear models. *J. Symbolic Comput.*, **41** (2), 222–233 (2006). MR2197157.

Fienberg, S. E. and Rinaldo, A. Three centuries of categorical data analysis: log-linear models and maximum likelihood estimation. *J. Statist. Plann. Inference*, **137** (11), 3430–3445 (2007). MR2363267.

Fienberg, S. E. and Rinaldo, A. Maximum likelihood estimation in log-linear models. *Ann. Statist.*, **40** (2), 996–1023 (2012). MR2985941.

Frieze, A. and Karoński, M. *Introduction to random graphs*. Cambridge University Press, Cambridge (2016). ISBN 978-1-107-11850-8. MR3675279.

Galvin, D. Three tutorial lectures on entropy and counting. *ArXiv Mathematics e-prints* (2014). arXiv: 1406.7872.

Geyer, C. J. *Likelihood and exponential families*. Ph.D. thesis, University of Washington (1990). MR2685353.

Gilbert, E. N. Random graphs. *Ann. Math. Statist.*, **30**, 1141–1144 (1959). MR108839.

Haberman, S. J. *The analysis of frequency data*. Statistical Research Monographs, Vol. IV. University of Chicago Press, Chicago, Ill.-London (1974). MR0408098.

Jacobsen, M. Existence and unicity of MLEs in discrete exponential family distributions. *Scand. J. Statist.*, **16** (4), 335–349 (1989). MR1039287.

Jendrej, J., Oleszkiewicz, K., and Wojtaszczyk, J. O. On some extensions of the FKN theorem. *Theory Comput.*, **11**, 445–469 (2015). MR3446023.

Johansen, S. *Introduction to the theory of regular exponential families.* University of Copenhagen, Institute of Mathematical Statistics, Copenhagen (1979). Lecture Notes, 3. MR558392.

Konis, K. *Linear programming algorithms for detecting separated data in binary logistic regression models.* Ph.D. thesis, University of Oxford (2007).

Lauritzen, S. L. Extreme point models in statistics. *Scand. J. Statist.*, **11** (2), 65–91 (1984). MR754971.

Lauritzen, S. L. *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York (1996). ISBN 0-19-852219-3. MR1419991.

Lehmann, E. L. and Casella, G. *Theory of point estimation.* Springer Texts in Statistics. Springer-Verlag, New York, second edition (1998). ISBN 0-387-98502-6. MR1639875.

Letac, G. *Lectures on natural exponential families and their variance functions*, volume 50 of *Monografías de Matemática*. Instituto de Matemática Pura e Aplicada (IMPA), Rio de Janeiro (1992). ISBN 85-244-0066-8. MR1182991.

Lovász, L. and Szegedy, B. Limits of dense graph sequences. *J. Combin. Theory Ser. B*, **96** (6), 933–957 (2006). MR2274085.

Mukherjee, R., Mukherjee, S., and Sen, S. Detection thresholds for the β-model on sparse graphs. *Ann. Statist.*, **46** (3), 1288–1317 (2018). MR3798004.

Perry, P. O. and Wolfe, P. J. Null models for network data. *ArXiv Mathematics e-prints* (2012). arXiv: 1201.5871.

Pitman, E. J. G. *Some basic theory for statistical inference.* Monographs on Applied Probability and Statistics. Chapman and Hall, London; John Wiley & Sons, Inc., New York (1979). ISBN 0-470-26554-X. MR549771.

Pósfai, A. *Approximation theorems related to the coupon collector's problem.* Ph.D. thesis, University of Szeged (2010). arXiv: 1006.3531.

Rinaldo, A., Fienberg, S. E., and Zhou, Y. On the geometry of discrete exponential families with application to exponential random graph models. *Electron. J. Stat.*, **3**, 446–484 (2009). MR2507456.

Rinaldo, A., Petrović, S., and Fienberg, S. E. Maximum likelihood estimation in the β-model. *Ann. Statist.*, **41** (3), 1085–1110 (2013). MR3113804.

Schrijver, A. *Theory of linear and integer programming.* Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, Ltd., Chichester (1986). ISBN 0-471-90854-1. MR874114.

Schweinberger, M., Krivitsky, P. N., Butts, C. T., and Stewart, J. R. Exponential-family models of random graphs: inference in finite, super and infinite population scenarios. *Statist. Sci.*, **35** (4), 627–662 (2020). MR4175389.

Stone, C. J. Large-sample inference for log-spline models. *Ann. Statist.*, **18** (2), 717–741 (1990). MR1056333.

Sullivant, S. *Algebraic statistics*, volume 194 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI (2018). ISBN 978-1-4704-3517-2. MR3838364.

Wainwright, M. J. and Jordan, M. I. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning*, **1** (1–2), 1–305 (2008). DOI: 10.1561/2200000001.

Wang, N., Rauh, J., and Massam, H. Approximating faces of marginal polytopes in discrete hierarchical models. *Ann. Statist.*, **47** (3), 1203–1233 (2019). MR3911110.