J. Appl. Prob. **55,** 1113–1130 (2018)
*doi:*10.1017/jpr.2018.74

ON RANDOM QUADRATIC FORMS: SUPPORTS OF POTENTIAL LOCAL MAXIMA

BORIS PITTEL,* *The Ohio State University*

Abstract

The selection model in population genetics is a dynamic system on the set of the probability distributions $\boldsymbol{p} = (p_1, \dots, p_n)$ of the alleles $A_1, \dots, A_n$, with $p_i(t+1)$ proportional to $p_i(t)$ multiplied by $\sum_j f_{i,j} p_j(t)$, and $f_{i,j} = f_{j,i}$ interpreted as a fitness of the gene pair (A_i, A_j). It is known that $\hat{\boldsymbol{p}}$ is a locally stable equilibrium if and only if $\hat{\boldsymbol{p}}$ is a strict local maximum of the quadratic form $\boldsymbol{p}^T \boldsymbol{f} \boldsymbol{p}$. Usually, there are multiple local maxima and $\lim \boldsymbol{p}(t)$ depends on $\boldsymbol{p}(0)$. To address the question of a *typical* behavior of $\{\boldsymbol{p}(t)\}$, John Kingman considered the case when the $f_{i,j}$ are independent and $[0, 1]$-uniform. He proved that with high probability (w.h.p.) no local maximum may have more than $2.49n^{1/2}$ positive components, and reduced 2.49 to 2.14 for a nonbiological case of exponentials on $[0, \infty)$. We show that the constant 2.14 serves a broad class of smooth densities on $[0, 1]$ with the increasing hazard rate. As for a lower bound, we prove that w.h.p. for all $k \le 2n^{1/3}$, there are many k-element subsets of $[n]$ that pass a partial test to be a support of a local maximum. Still, it may well be that w.h.p. the actual supports are much smaller. In that direction, we prove that w.h.p. a support of a local maximum, which does not contain a support of a local equilibrium, is very unlikely to have size exceeding $\frac{2}{3}\log_2 n$ and, for the uniform fitnesses, there are super-polynomially many potential supports free of local equilibriums of size close to $\frac{1}{2}\log_2 n$.

Keywords: Stable polymorphism; random fitness; asymptotics

2010 Mathematics Subject Classification: Primary 34E05; 60C05

1. Introduction and main results

The classic selection model in population genetics is a dynamic system on the set of the probability distributions $\boldsymbol{p} = (p_1, \dots, p_n) \in \Delta_n := \{\boldsymbol{x} \ge \boldsymbol{0}, \sum_{i\in[n]} x_i = 1\}$ of the alleles $A_1, \dots, A_n$ at the single locus

$$p_i(t+1) = p_i(t)\frac{\sum_j f_{i,j} p_j(t)}{\sum_{r,s} f_{r,s} p_r(t) p_s(t)}, \qquad i \in [n]. \tag{1.1}$$

Here, each $f_{r,s} = f_{s,r} \in [0, 1]$ is interpreted as the fitness, i.e. the probability that the unordered gene pair (A_r, A_s) survives to an adult age. While the dynamic behavior of $\boldsymbol{p}(t)$ in this model certainly depends on the fitness matrix $\boldsymbol{f} = \{f_{r,s}\}$, it has long been known that the *average fitness* $V(\boldsymbol{p}(t)) := \sum_{r,s} f_{r,s} p_r(t) p_s(t)$ strictly increases with t unless $\boldsymbol{p}(t+1) = \boldsymbol{p}(t)$. Hofbauer and Sigmund [9] characterized this property as a consequence of Fisher's fundamental theorem of natural selection [6], provided a full proof following Kingman [11], and sketched the different proofs given by Scheuer and Mandel [19] and Baum and Eagon [2].

Received 11 August 2017; revision received 23 October 2018.
* Postal address: Department of Mathematics, The Ohio State University, Columbus, Ohio 43210, USA.
Email address: bgp@math.ohio-state.edu

Using the increase of the average fitness, it was later proven by various researchers that $\boldsymbol{p}(\infty) = \lim_{t\to\infty} \boldsymbol{p}(t)$ exists for every initial gene distribution $\boldsymbol{p}(0)$, and $\boldsymbol{p} := \boldsymbol{p}(\infty)$ is a fixed point of the mapping $\boldsymbol{\Phi}(\cdot)\colon \Delta_n \to \Delta_n$ defined by (1.1), with the property, for a nonempty $I \subseteq [n]$,

$$p_i = 0, \quad (i \notin I), \qquad \sum_{j\in I} f_{i,j} p_j \equiv V(\boldsymbol{p}), \quad (i \in I).$$

Remarkably, a fixed point $\boldsymbol{p}$ is a *locally* stable equilibrium if and only if $\boldsymbol{p}$ is a strict local maximum of $V(\cdot)$. There is no reason to expect that a local maximum is unique; so typically the limit $\boldsymbol{p}(\infty)$ depends on $\boldsymbol{p}(0)$.

For $\boldsymbol{p} \in \Delta_n$ to be a local maximum of $V(\cdot)$, the following three conditions are both necessary and sufficient; see [12]. If $I = I(\boldsymbol{p}) := \{i\colon p_i > 0\}$ then

$$\begin{gathered} \sum_{j\in I} f_{i,j} p_j \equiv V(\boldsymbol{p}), \quad (i \in I), \\ \sum_{i,j\in I} f_{i,j} x_i x_j \le 0 \quad \text{for all } \{x_i\}_{i\in I} \quad \text{with } \sum_{i\in I} x_i = 0, \\ \sum_{j\in I} f_{i,j} p_j \le V(\boldsymbol{p}), \quad (i \notin I). \end{gathered} \tag{1.2}$$

The second necessary condition applied to $\boldsymbol{x}$ such that $x_i = 1$, $x_j = -1$, with the remaining $x_k = 0$, easily yields

$$f_{i,j} \ge \tfrac{1}{2}(f_{i,i} + f_{j,j}), \qquad i, j \in I,\ i \ne j. \tag{1.3}$$

To quote from [12]: 'This condition uses internal stability alone, and takes no account of vulnerability to mutation'.

Inequality (1.3) was earlier obtained by Lewontin *et al.* [17] as a corollary of a determinantal criterion applied to the system of $(|I| - 1)$ linear equations for p_i, $i \in I \setminus \{i_0\}$, implicit in

$$\sum_{j\in I} f_{i,j} p_j \equiv V(\boldsymbol{P}), \quad (i \in I), \qquad \sum_{i\in I} p_i = 1.$$

It was also asserted in [17] that $f_{i,j} < \max_k(f_{i,k} + f_{k,j})$; the proof is valid under an additional condition $f_{i,j} > \max\{f_{i,i}, f_{j,j}\}$.

A subset I meeting condition (1.3) is a candidate to be a support set of a local maximum of $\boldsymbol{p}^T \boldsymbol{f}\boldsymbol{p}$. (We will use the term *K-set* for such sets I.) Kingman [12] posed a problem of analyzing these potential supports in a *typical* case, i.e. when $f_{i,j}$ are independent and identically distributed (i.i.d.) random variables with range [0, 1]. For the case when $f_{i,j}$ are [0, 1]-uniform, he proved that with high probability (w.h.p.), i.e. with probability approaching 1, $\max|I| \le 2.49n^{1/2}$: so 'the largest stable polymorphism will contain at most of the order of $n^{1/2}$ alleles'. The key tool was the bound $\mathbb{P}(D_I) \le 1/r!$, $r := |I|$, where D_I is the event in (1.3). He found that, for a (nonbiological) exponential distribution on $[0, \infty)$, $\mathbb{P}(D_I) = (2/(r+1))^r \ll 1/r!$ and the constant 2.49 was reduced to 2.14.

Haigh [7], [8] established the counterparts of some of Kingman's results for the case of a nonsymmetric payoff matrix. For instance, he proved that for the density $e^{-x}/\sqrt{\pi x}$, $(x > 0)$, of χ_1^2 distribution, w.h.p. no evolutionarily stable strategy has support of size exceeding $1.64n^{2/3}$. Kontogiannis and Spirakis [16] used the technique from [8] to resolve the cases of uniform distribution and of standard normal distribution left open there.

https://doi.org/10.1017/jpr.2018.74 Published online by Cambridge University Press

Recently, and independently of the work cited above, Chen and Peng [3] studied, in an operations research context of the random quadratic *minimization* problems, the probability of the event which is a multiplicative version of the Kingman event D_I, i.e. (1.3). Translated into the conditions in (1.3), their results include an upper bound $2^r/(r+1)!$ (for a general continuous distribution) (see [14]), the exact formula $2^r/(r+1)^r$ (for an exponential distribution) (see [12]), and a new lower bound $2^r/(r-1)^r$ (for a double exponential distribution).

In [12], Kingman suggested that it would be interesting 'to carry out a comparative analysis for other distributions of the $f_{i,j}$', and conjectured, in [14], that 'for every continuous distribution F of f, there is a finite $\beta(F) = \lim_{r\to\infty}\{r!\,P(D_I)\}^{1/r}$'. Whenever this limit exists, $\max|I| \le 2.49\beta(F)^{1/2}n^{1/2}$ w.h.p. In general, $\limsup_{r\to\infty}\{r!\,\mathbb{P}(D_I)\}^{1/r} \le 2$ and $\max|I| \le 2.49\sqrt{2}n^{1/2}$ w.h.p.

In this paper we consider a relatively broad class of the distributions F, meeting the following conditions:

(I) $F(x)$ has a differentiable positive density $g(x)$, $x \in [0, 1]$, such that $g'(x) \le 0$;

(II) the *hazard ratio* $\lambda(x) := g(x)/(1-F(x))$ is increasing with x. The nonincreasing linear density $g_c(x) = (1-cx)/(1-\frac{1}{2}c)$, $c \in [0, 1]$ ($g_0(x) \equiv 1$) meets these constraints, and so does $g(x) = ce^{-cx}/(1-e^{-c})$, the density of the negative exponential distribution *conditioned* on $[0, 1]$.

For F meeting conditions (I) and (II), we prove that

$$\left(\frac{2}{r+1}\right)^r \le \mathbb{P}(D_I) \le \frac{r^r}{\binom{r}{2}^{\overline{r}}} \le \left(\frac{2}{r-1}\right)^r, \tag{1.4}$$

$a^{\overline{b}} := a(a+1)\cdots(a+b-1)$. In combination with Kingman's analysis of the exponential distribution on $[0, \infty)$, it follows from (1.4) that for *every* F meeting the constraints above, we have $\max|I| \le 2.14n^{1/2}$ w.h.p. We see also that, for every F in question,

$$\lim_{r\to\infty}\{r!\,\mathbb{P}(D_I)\}^{1/r} =: \beta(F) = \frac{2}{e},$$

proving not only that $\beta(F)$ exists, but also that $\beta(F)$ does not depend on F in this class. This lends a certain support to Kingman's conjecture (see [14]) that $\lim_{r\to\infty}\{r!\,\mathbb{P}(D_I)\}^{1/r}$ exists for every continuous F.

Suppose we restrict our attention to the K-sets I such that there is no $J \subset I$, ($|J| \ge 2$), which supports a local equilibrium $\boldsymbol{p} = \{p_i\}_{i\in J}$, meeting the first two conditions in (1.2). We will call such K-sets minimal, hoping that the reader will accept our abuse of set minimality notion. Let $\mathcal{D}_I$ be the corresponding event. For the distributions F from the class described above, we prove that

$$\mathbb{P}(\mathcal{D}_I) \le 2^{-r^2/2}\left(\frac{4e}{r}\right)^{r/2}\exp(\Theta(r^{1/3})), \qquad r := |I|. \tag{1.5}$$

The fact that this probability is much smaller than $\mathbb{P}(D_I)$ may be a signal that the full power of condition (II) ($\boldsymbol{x}^T\boldsymbol{f}\boldsymbol{x} \ge 0$ on the hyperplane $\sum_i x_i = 0$) could possibly be marshaled to prove that this condition holds with a similarly small probability $\exp(-\Theta(r^2))$. Tellingly, having studied large samples of the $r \times r$ fitness matrices $\boldsymbol{f}$ with $[0, 1]$-uniform entries, Lewontin *et al.* [17] concluded that the empirical frequency of $\boldsymbol{f}$ with stable equilibrium support r decays as e^{-cr^2}.

https://doi.org/10.1017/jpr.2018.74 Published online by Cambridge University Press

Continuing with $\mathbb{P}(D_I)$, suppose that, in addition, $g^{(3)}(0)$ exists. Then

$$\mathbb{P}(D_I) = (1 + O(r^{-\sigma}))\left(\frac{2}{r}\right)^r \exp\left(\frac{g'(0)}{g^2(0)}\right) \qquad \text{for all } \sigma < \tfrac{1}{3}, \tag{1.6}$$

and if $|I_1| = |I_2| = r$, $|I_1 \cap I_2| = k$, then

$$\mathbb{P}(D_{I_1} \cap D_{I_2}) = O(\mathbb{P}(r,k)), \qquad \mathbb{P}(r,k) := r^{-6}\left(\frac{2}{r}\right)^{2(r-k-1)}\left(\frac{2}{2r-k}\right)^{k-1} \tag{1.7}$$

uniformly for $r \geq 2$ and $k \in [1, r-1]$.

Let $X_{n,r}$ be the total number of K-sets of $[n]$ of cardinality r. We already know that w.h.p. $X_{n,r} = 0$ for $r > 2.14n^{1/2}$, and the left-hand side of (1.4) implies that $\mathbb{E}[X_{n,r}] \to \infty$ for every $r \leq (\sqrt{2\mathrm{e}} - \varepsilon)n^{1/2}$. Note that $\sqrt{2\mathrm{e}} > 2.3 > 2.14$; so for r roughly from $2.14n^{1/2}$ to $2.3n^{1/2}$, $\mathbb{E}[X_{n,r}] \to \infty$, while $X_{n,r} = 0$ with probability approaching 1. We use estimates (1.4), (1.6), and (1.7) to show that

$$\frac{\mathrm{var}(X_{n,r})}{\mathbb{E}^2[X_{n,r}]} = O(n^{-2/3}), \qquad 2 \leq r \leq r(n) := \lceil 2n^{1/3} \rceil.$$

It follows that

$$\mathbb{P}\left(\bigcap_{\rho=2}^{r(n)} \left\{ \left| \frac{X_{n,\rho}}{\mathbb{E}[X_{n,\rho}]} - 1 \right| \leq n^{-1/6+\varepsilon} \right\}\right) = 1 - O(n^{-2\varepsilon}), \qquad \varepsilon < \tfrac{1}{6},$$

i.e. w.h.p. the counts of the K-sets of size r ranging from 2 to $r(n)$ are uniformly asymptotic to their expected values. In particular, setting $L_n = \max\{\rho\colon X_{n,\rho} > 0\}$, we have $\mathbb{P}(L_n > 2n^{1/3}) \to 1$, i.e. w.h.p. the size of the largest potential support of a local maximum is sandwiched between $2n^{1/3}$ and $2.14n^{1/2}$.

We cannot rule out the possibility that, w.h.p., the actual supports of local maxima are considerably smaller. In this direction, we use bound (1.5) to show that, with probability greater that $1 - n^{-a}$ (for all $a > 0$), there are no minimal K-sets of cardinality $> (\frac{2}{3})\log_2 n$. In other words, w.h.p. every K-set of this size, if any exists, has a subset of order exactly 2 satisfying the first two conditions in (1.2). Complementing this claim, we show that, w.h.p., the number of K-sets of size less than $\frac{1}{2}\log_2 n$ that do not contain the size 2 supports of local equilibriums is super-polynomially large. We emphasize that these special K-sets are just *potential* supports of the local maxima. Still, the fact that they are likely to be very numerous can be interpreted as supporting a conjecture that there exist (many) genuine local maxima with support sizes tending to ∞ in probability.

The already cited paper [3] was preceded by Chen *et al.* [4]; both papers addressed the likely behavior of an *absolute* minimum of a random quadratic form $\boldsymbol{x}^T Q\boldsymbol{x}$ for $\boldsymbol{x} \in \Delta_n$. Under the condition that the elements of Q are i.i.d. random variables with a cumulative distribution function F concave on its support, the support size of the absolute minimum point was shown to be *bounded* in probability, with its distribution tail decaying exponentially fast. In particular, it followed that, for $f_{i,j}$ uniform or *positive*-exponential on $[0,1]$, the absolute *maximum* of $\boldsymbol{p}^T\boldsymbol{f}\boldsymbol{p}$ is attained at a point of Δ_n with N, the number of positive components, satisfying $\mathbb{P}(N \geq k) = O(\rho^k)$, $k > 0$, as $n \to \infty$.

In view of all this information, we conjecture that—for $f_{i,j}$ meeting conditions (I) and (II)—the size of the largest support of a *local* maximum of $\boldsymbol{p}^T\boldsymbol{f}\boldsymbol{p}$ is, w.h.p., of (poly)logarithmic order.

https://doi.org/10.1017/jpr.2018.74 Published online by Cambridge University Press

If the Lewontin–Ginzburg–Tuljapurkar conjecture is confirmed, it will follow that, in fact, the largest such support w.h.p. is of logarithmic order, at most. Proving that this estimate is sharp would be the remaining challenge. We hope that the techniques/estimates developed in this paper will be useful for tackling these challenging problems. In this regard, a surprising appearance of Selberg's integral in the proof of Lemma 2.4 may be a sign that there are other multidimensional integrals yet to enter the stage.

2. Proofs

2.1. Estimate of $\mathbb{P}(D_I)$

Theorem 2.1. *Suppose that F*

(i) *has a positive, nonincreasing, differentiable density g;*

(ii) *has a nondecreasing hazard ratio $\lambda(x) = g(x)/(1 - F(x))$.*

Then, with $a^{\overline{b}} := a(a+1)\cdots(a+b-1)$, we have

$$\left(\frac{2}{r+1}\right)^r \le \mathbb{P}(D_I) \le \frac{r^r}{[\binom{r}{2}+1]^{\overline{r}}} \le \left(\frac{2}{r-1}\right)^r.$$

In the special case of the uniform density $g(x) \equiv 1$, this bound improves Kingman's bound $\mathbb{P}(D_I) \le 1/r!$. It also shows that, for all F meeting conditions (i) and (ii),

$$\lim_{r\to\infty} \{r!\,\mathbb{P}(D_I)\}^{1/r} = \frac{2}{\mathrm{e}}.$$

Proof of Theorem 2.1. As in [12], the probability of D_I, conditioned on $\{f_{i,i} = x_i,\ i \in I\}$, is

$$\prod_{(i,j)} \mathbb{P}\left(f \ge \frac{x_i + x_j}{2}\right) = \prod_{(i,j)} \left(1 - F\left(\frac{x_i + x_j}{2}\right)\right),$$

where $i \ne j$. The function $1 - F(x)$ is log-concave, since

$$\frac{\mathrm{d}}{\mathrm{d}x} \log(1 - F(x)) = -\frac{g(x)}{1 - F(x)} = -\lambda(x)$$

is decreasing with x.

Lower bound. By Jensen's inequality,

$$\prod_{(i,j)} \left(1 - F\left(\frac{x_i + x_j}{2}\right)\right) \ge \prod_{(i,j)} (1 - F(x_i))^{1/2}(1 - F(x_j))^{1/2} = \prod_{i=1}^{r} (1 - F(x_i))^{(r-1)/2}.$$

Consequently,

$$\mathbb{P}(D_I) \ge \int_{\boldsymbol{x}\in[0,1]^r} \cdots \int \prod_{i=1}^{r} (1 - F(x_i))^{(r-1)/2} \prod_{i=1}^{r} g(x_i)\,\mathrm{d}x_i,$$

and, switching to the variables $y_i = F(x_i)$,

$$\mathbb{P}(D_I) \ge \int_{\boldsymbol{y}\in[0.1]^n} \cdots \int \prod_{i=1}^{r} (1 - y_i)^{(r-1)/2}\,\mathrm{d}\boldsymbol{y} = \left(\int_0^1 (1 - y)^{(r-1)/2}\,\mathrm{d}y\right)^r = \left(\frac{2}{r+1}\right)^r.$$

https://doi.org/10.1017/jpr.2018.74 Published online by Cambridge University Press

Upper bound. Again by Jensen's inequality, denoting $s = \sum_i x_i$, we have

$$\begin{aligned}\prod_{(i,j)}\left(1 - F\left(\frac{x_i + x_j}{2}\right)\right) &= \exp\left[\binom{r}{2}\sum_{(i,j)}\frac{1}{\binom{r}{2}}\log\left(1 - F\left(\frac{x_i + x_j}{2}\right)\right)\right]\\ &\le \exp\left[\binom{r}{2}\log\left(1 - F\left(\frac{1}{r(r-1)}\sum_{(i,j)}(x_i + x_j)\right)\right)\right]\\ &= \exp\left[\binom{r}{2}\log\left(1 - F\left(\frac{s}{r}\right)\right)\right]\\ &= \left(1 - F\left(\frac{s}{r}\right)\right)^{\binom{r}{2}}.\end{aligned}$$

Consequently,

$$\mathbb{P}(D_I) \le \int_{\boldsymbol{x}\in[0,1]^r}\cdots\int\left(1 - F\left(\frac{s}{r}\right)\right)^{\binom{r}{2}}\prod_{i\in I} g(x_i)\,\mathrm{d}x_i. \tag{2.1}$$

Again change the variables of integration, setting $y_i = F(x_i)$, so that $x_i = F^{-1}(y_i)$ and $s = \sum_{i\in I} F^{-1}(y_i)$. Now

$$\frac{\mathrm{d}^2}{\mathrm{d}y^2}F^{-1}(y) = -\frac{g'(x)}{g(x)^3} \ge 0,$$

implying that $F^{-1}(y)$ is convex. Therefore, for each $t \le r$, we have

$$r^{-1}\min\left\{\sum_{i\in I}F^{-1}(y_i)\colon \sum_{i\in I} y_i = t\right\} = F^{-1}\left(\frac{t}{r}\right).$$

Hence,

$$\begin{aligned}\max\left\{1 - F\left(r^{-1}\sum_{i\in I}x_i\right)\colon \sum_{i\in I} y_i = t\right\} &= 1 - \min\left\{F\left(r^{-1}\sum_{i\in I}F^{-1}(y_i)\right)\colon \sum_{i\in I} y_i = t\right\}\\ &= 1 - F\left(r^{-1}\min\left\{\sum_{i\in I}F^{-1}(y_i)\colon \sum_{i\in I} y_i = t\right\}\right)\\ &= 1 - F\left(F^{-1}\left(\frac{t}{r}\right)\right)\\ &= 1 - \frac{t}{r}.\end{aligned}$$

So (2.1) yields

$$\mathbb{P}(D_I) \le \int_{\boldsymbol{y}\in[0,1]^r}\cdots\int\left(1 - \frac{t}{r}\right)^{\binom{r}{2}}\prod_{i\in I}\mathrm{d}y_i.$$

Since

$$\int_{\sum_i y_i\le t}\cdots\int\prod_{i\in I}\mathrm{d}y_i = \frac{t^r}{r!},$$

https://doi.org/10.1017/jpr.2018.74 Published online by Cambridge University Press

we conclude that

$$\begin{aligned}\mathbb{P}(D_I) &\le \int_0^r \left(1-\frac{t}{r}\right)^{\binom{r}{2}} \frac{t^{r-1}}{(r-1)!}\,\mathrm{d}t \\ &= \frac{r^r}{(r-1)!}\int_0^1 (1-\tau)^{\binom{r}{2}}\tau^{r-1}\,\mathrm{d}\tau \\ &= \frac{r^r}{(r-1)!}\,\frac{\binom{r}{2}!\,(r-1)!}{(\binom{r}{2}+r)!} \\ &\le \frac{r^r}{\binom{r}{2}^{\bar{r}}}.\end{aligned}$$

□

Theorem 2.2. *Suppose that, in addition to conditions (i) and (ii) in Theorem 2.1, we have the following additional condition:*

(iii) $g^{(3)}(0)$ *exists.*

Then

$$\mathbb{P}(D_I) = (1+O(r^{-\sigma}))\left(\frac{2}{r}\right)^r \exp\left(\frac{g'(0)}{g^2(0)}\right) \quad \textit{for every } \sigma < \tfrac{1}{3}.$$

To prove this claim, we shrink, in steps, the cube $[0,1]^n$ to a subset C^* in such a way that the integral of the product of $1-F(\frac{1}{2}(x_i+x_j))$ over C^* sharply approximates that over $[0,1]^n$, and the product itself admits a manageable approximation on C^*.

Given $C \subset [0,1]^n$, denote

$$\mathbb{P}_C(D_I) = \int_{\boldsymbol{x}\in C}\cdots\int \prod_{(i\ne j)}\left(1-F\left(\frac{x_i+x_j}{2}\right)\right)\mathrm{d}\boldsymbol{x}.$$

Lemma 2.1. *Let*

$$C_1 := \left\{\boldsymbol{x}\in[0,1]^n : \left|\sum_{i=1}^r F(x_i) - 2\right| \le r^{-1/3}\right\}.$$

Then

$$\mathbb{P}(D_I) - \mathbb{P}_{C_1}(D_I) \le \left(\frac{2}{r}\right)^r \exp\left(-\frac{r^{1/3}}{10}\right).$$

Proof. Let $\tau_{1,2} = 2/r \mp r^{-4/3}$. From the proof of Theorem 2.1, it follows that

$$\mathbb{P}(D_I) - \mathbb{P}_{C_1}(D_I) \le \frac{r^r}{(r-1)!}\int_{\tau\in[\tau_1,\tau_2]^c} (1-\tau)^{\binom{r}{2}}\tau^{r-1}\,\mathrm{d}\tau.$$

The (log-concave) integrand attains its maximum at $\tau_{\max} = 2/(2+r) \in [\tau_1,\tau_2]$, and

$$\max\left\{\frac{\mathrm{d}^2}{\mathrm{d}\tau^2}(\log(1-\tau)^{\binom{r}{2}}\tau^{r-1}) : \tau\in[\tau_1,\tau_2]\right\} \le -\frac{r^3}{4.1}.$$

Therefore, the integral is at most

$$\left(1-\frac{2}{2+r}\right)^{\binom{r}{2}}\left(\frac{2}{2+r}\right)^{r-1}\exp\left(-\frac{r^{1/3}}{9}\right),$$

https://doi.org/10.1017/jpr.2018.74 Published online by Cambridge University Press

so that

$$\mathbb{P}(D_I) - \mathbb{P}_{C_1}(D_I) \le \frac{r^r}{(r-1)!}\left(1-\frac{2}{2+r}\right)^{\binom{r}{2}}\left(\frac{2}{2+r}\right)^{r-1}\exp\left(-\frac{r^{1/3}}{9}\right)$$
$$\le \left(\frac{2}{r}\right)^r \exp\left(-\frac{r^{1/3}}{10}\right). \qquad \square$$

Lemma 2.2. *Let*

$$C_2 := \left\{\boldsymbol{x} \in C_1 \colon \max_i \frac{F(x_i)}{\sum_j F(x_j)} \le k\frac{\log r}{r}\right\}, \quad (k > 1).$$

Then

$$\mathbb{P}_{C_1}(D_I) - \mathbb{P}_{C_2}(D_I) \le \left(\frac{2}{r}\right)^r r^{-\alpha} \quad \textit{for all } \alpha < k-1.$$

Proof. Similarly to the proof of Lemma 2.1,

$$\mathbb{P}_{C_1}(D_I) - \mathbb{P}_{C_2}(D_I) \le \int\cdots\int_{\max y_i/t > k(\log r)/r} \left(1-\frac{t}{r}\right)^{\binom{r}{2}} \prod_{i\in I} \mathrm{d}y_i.$$

Introduce $L_1, \dots, L_r$ the lengths of the consecutive subintervals of $[0, 1]$ obtained by sampling uniformly at random $r-1$ points in $[0, 1]$. From [18, Lemma 1], the integral above is at most

$$\mathbb{P}\left(\max L_i \ge k\frac{\log r}{r}\right)\int_0^r \left(1-\frac{t}{r}\right)^{\binom{r}{2}} \frac{t^{r-1}}{(r-1)!}\,\mathrm{d}t = \mathbb{P}\left(\max L_i \ge k\frac{\log r}{r}\right)\frac{r^r}{\binom{r}{2}^{(r)}}.$$

And, introducing $U_1, \dots, U_{r-1}$ the independent $[0, 1]$-uniforms, the probability factor is at most

$$r\mathbb{P}\left(L_1 \ge k\frac{\log r}{r}\right) = r\mathbb{P}\left(\min_i U_i \ge k\frac{\log r}{r}\right) = r\left(1-k\frac{\log r}{r}\right)^{r-1} \le r\exp\left(-(r-1)k\frac{\log r}{r}\right),$$

concluding the proof. $\square$

One more reduction step defines the final

$$C^* = \left\{\boldsymbol{x} \in C_2 \colon \left|\frac{r}{2}\frac{\sum_i F^2(x_i)}{(\sum_j F(x_j))^2} - 1\right| \le r^{-\sigma}\right\}, \qquad \sigma < \tfrac{1}{3}. \tag{2.2}$$

Lemma 2.3. *It holds that*

$$\mathbb{P}_{C_2}(D_I) - \mathbb{P}_{C^*}(D_I) \le \left(\frac{2}{r}\right)^r \exp\left(-\frac{1}{2}r^{1/3-\sigma}\right).$$

Proof. Once again as in the proofs of Lemmas 2.1 and 2.2,

$$\mathbb{P}_{C_2}(D_I) - \mathbb{P}_{C^*}(D_I) \le \int\cdots\int_{|(r/2)(\sum_i y_i^2/(\sum_j y_j)^2)-1|>r^{-\sigma}} \left(1-\frac{t}{r}\right)^{\binom{r}{2}} \prod_{i\in I}\mathrm{d}y_i$$
$$\le \mathbb{P}\left(\left|\frac{r}{2}\sum_i L_i^2 - 1\right| > r^{-\sigma}\right)\frac{r^r}{\binom{r}{2}^{(r)}}$$
$$\le \left(\frac{2}{r}\right)^r \exp(-\Theta(r^{1/3-\sigma}))$$

since the probability is at most $\exp(-\Theta(r^{1/3-\sigma}))$; see [**?**, Lemma 3.2]). $\square$

https://doi.org/10.1017/jpr.2018.74 Published online by Cambridge University Press

Remark 2.1. A key to the proof of [**?**, Lemma 3.2] was the classic fact that $(L_1, \dots, L_r)$ and $(\sum_i W_i)^{-1}(W_1, \dots, W_r)$, ($W_j$ being i.i.d. Exponentials), are equidistributed; see [5]. While both of the distribution tails of $\sum_i W_j$ decay exponentially, for the right tail of $\sum_j W_j^2$ we could prove only the bound $\mathrm{e}^{-\Theta(r^\delta)}$, $\delta < \frac{1}{3}$. The obstacle here is that $\mathbb{E}[\mathrm{e}^{zW^2}] = \infty$ for every $z > 0$.

Combining Lemmas 2.1–2.3 brings us to our next result.

Corollary 2.1. *It holds that*

$$\mathbb{P}(D_I) - \mathbb{P}_{C^*}(D_I) \le \left(\frac{2}{r}\right)^r r^{-\alpha} \quad \textit{for all } \alpha < k - 1.$$

For $\boldsymbol{x} \in C^*$, we have $\max_i F(x_i) \le 3k \log r / r \to 0$, which implies that $\max_i x_i = O(r^{-1} \log r) \to 0$. For $x = O(r^{-1} \log r)$, we have

$$F(x) = g(0)x + \tfrac{1}{2} g^{(1)}(0)x^2 + O(x^3) = g(0)x + \tfrac{1}{2} g^{(1)}(0)x^2 + O(r^{-3} \log^3 r).$$

So

$$\log(1 - F(x)) = -g(0)x - \tfrac{1}{2}(g'(0) + g^2(0))x^2 + O(r^{-3} \log^3 r),$$

and with a bit of algebra

$$\begin{aligned} &\log\big(1 - F\big(\tfrac{1}{2}(x_i + x_j)\big)\big) - \tfrac{1}{2}(\log(1 - F(x_i)) + \log(1 - F(x_j))) \\ &\quad = \tfrac{1}{8}(g'(0) + g^2(0))(x_i - x_j)^2 + O(r^{-3} \log^3 r) \\ &\quad = \gamma(F(x_i) - F(x_j))^2 + O(r^{-3} \log^3 r), \qquad \gamma := \frac{g'(0) + g^2(0)}{8g^2(0)}. \end{aligned} \tag{2.3}$$

Therefore,

$$\begin{aligned} &\prod_{(i,j)} \log\left(1 - F\left(\frac{x_i + x_j}{2}\right)\right) \\ &\quad = \exp\left(\frac{r-1}{2} \sum_i \log(1 - F(x_i))\right) \exp\left(\gamma \sum_{(i,j)} (F(x_i) - F(x_j))^2 + O(r^{-1} \log r)\right), \end{aligned}$$

where

$$\begin{aligned} \frac{r-1}{2} \sum_i \log(1 - F(x_i) &= -\frac{r-1}{2} \sum_i \left(F(x_i) + \frac{F^2(x_i)}{2}\right) + O(r^{-1} \log r), \\ \sum_{(i,j)} (F(x_i) - F(x_j))^2 &= r \sum_i F^2(x_i) - \left(\sum_i F(x_i)\right)^2. \end{aligned}$$

Hence, on C^* (see (2.2)),

$$\begin{aligned} &\prod_{(i,j)} \log\left(1 - F\left(\frac{x_i + x_j}{2}\right)\right) \\ &\quad = \exp\left(-\frac{r-1}{2} \sum_i F(x_i) - \gamma \left(\sum_i F(x_i)\right)^2 \right. \\ &\qquad \left. + \left(-\frac{r-1}{4} + \gamma r\right) \sum_i F^2(x_i) + O(r^{-1} \log r)\right) \end{aligned}$$

https://doi.org/10.1017/jpr.2018.74 Published online by Cambridge University Press

$$= \exp\biggl(-\frac{r-1}{2}\sum_i F(x_i) + \Bigl(2\gamma - \frac{1}{2}\Bigr)\Bigl(\sum_i F(x_i)\Bigr)^2 + O(r^{-\sigma})\biggr)$$
$$= \exp\biggl(-\frac{r}{2}\sum_i F(x_i) + \frac{g'(0)}{g^2(0)} + O(r^{-\sigma})\biggr);$$

for the last equality we use the definition of γ in (2.3).

Switching to the variables $y_i = F(x_i)$ and denoting $t = \sum_i y_i$, we obtain

$$\mathbb{P}_{C^*}(D_I) = \int\limits_{\boldsymbol{y}\in\mathcal{C}^*}\cdots\int \exp\biggl(-\frac{r}{2}t + \frac{g'(0)}{g^2(0)} + O(r^{-\sigma})\biggr)\,\mathrm{d}\boldsymbol{y},$$
$$\mathcal{C}^* := \biggl\{\boldsymbol{y}\ge\boldsymbol{0}\colon |t-2|\le r^{-1/3}, \max_i \frac{y_i}{t} \le k\frac{\log r}{r}, \biggl|\frac{r}{2t^2}\sum_i y_i^2 - 1\biggr| \le r^{-\sigma}\biggr\}.$$

Note that on $\mathcal{C}^*$ we have $\max_i y_1 \to 0$, so that the omitted condition $\max_i y_i \le 1$ would have been superfluous. From [**?**, Lemma 3.1],

$$\int\limits_{y\in\mathcal{C}^*}\cdots\int \mathrm{e}^{-rt/2}\,\mathrm{d}\boldsymbol{y}$$
$$= \int\limits_{|t-2|\le 1/r^{1/3}} \frac{\mathrm{e}^{-rt/2}t^{r-1}}{(r-1)!}\mathbb{P}\biggl(\max L_i \le \min\Bigl(t^{-1}, \frac{k\log r}{r}\Bigr), \biggl|\frac{r}{2}\sum_i L_i^2 - 1\biggr| \le r^{-\sigma}\biggr)\,\mathrm{d}t$$
$$= \mathbb{P}\biggl(\max L_i \le \frac{k\log r}{r}, \biggl|\frac{r}{2}\sum_i L_i^2 - 1\biggr| \le r^{-\sigma}\biggr) \int\limits_{|t-2|\le 1/r^{1/3}} \frac{\mathrm{e}^{-rt/2}t^{r-1}}{(r-1)!}\,\mathrm{d}t.$$

From Lemmas 2.2 and 2.3 and their proofs, we know that the probability factor is at least $1 - r^{-\alpha}$ for all $\alpha < k-1$. Furthermore, the integral is equal to

$$\int_0^\infty \frac{\mathrm{e}^{-rt/2}t^{r-1}}{(r-1)!}\,\mathrm{d}t - \int\limits_{|t-2|>1/r^{1/3}} \frac{\mathrm{e}^{-rt/2}t^{r-1}}{(r-1)!}\,\mathrm{d}t = \Bigl(\frac{2}{r}\Bigr)^r - \frac{(2/r)^r}{(r-1)!}\int\limits_{|\tau-r|>r^{2/3}/2} \mathrm{e}^{-\tau}\tau^{r-1}\,\mathrm{d}\tau,$$

and, by Chebyshev's inequality,

$$\frac{1}{(r-1)!}\int\limits_{|\tau-r|>r^{2/3}/2} \mathrm{e}^{-\tau}\tau^{r-1}\,\mathrm{d}\tau \le \mathbb{P}\biggl(|\,\mathrm{Poisson}(r-1) - (r-1)| > \frac{r^{2/3}}{3}\biggr)$$
$$\le \frac{9(r-1)}{r^{4/3}}$$
$$\le 9r^{-1/3}.$$

So

$$\int\limits_{|t-2|\le 1/r^{1/3}} \frac{\mathrm{e}^{-rt/2}t^{r-1}}{(r-1)!}\,\mathrm{d}t = (1 + O(r^{-1/3}))\Bigl(\frac{2}{r}\Bigr)^r.$$

https://doi.org/10.1017/jpr.2018.74 Published online by Cambridge University Press

Consequently,

$$\mathbb{P}_{\mathcal{C}^*}(D_I) = (1 + O(r^{-\sigma}))\left(\frac{2}{r}\right)^r \exp\left(\frac{g'(0)}{g^2(0)}\right) \quad \text{for every } \sigma < \tfrac{1}{3}.$$

Combining this estimate with Corollary 2.1, we complete the proof of Theorem 2.2.

2.2. Estimate of $\mathbb{P}(D_{I_1} \cap D_{I_2})$

Let $I_1, I_2 \subset [n]$, $|I_j| = r$. If $I_1 \cap I_2 = \varnothing$ then the events D_{I_1} and D_{I_2} are independent and so, by Theorem 2.2,

$$\mathbb{P}(D_{I_1} \cap D_{I_2}) = \mathbb{P}(D_{I_1})\mathbb{P}(D_{I_2}) = (1 + O(r^{-\sigma}))\left(\frac{2}{r}\right)^{2r} \exp\left(2\frac{g'(0)}{g^2(0)}\right).$$

Consider the $|I_1 \cap I_2| = k \in [1, r-1]$ case. By symmetry, we can assume that $I_1 = \{1, \dots, r\}$, and $I_2 = \{r-k+1, \dots, 2r-k\}$. The probability of $D_{I_1} \cap D_{I_2}$, conditioned on the event $\{F_{i,i} = x_i : 1 \le i \le 2r-k\}$, is

$$\Psi(\boldsymbol{x}) = \prod_{i,j \le r\ (i \ne j)} \left(1 - F\left(\frac{x_i + x_j}{2}\right)\right) \prod_{r < i \le 2r-k,\, r-k+1 \le j < r} \left(1 - F\left(\frac{x_i + x_j}{2}\right)\right) \times \prod_{r \le i,\, j \le 2r-k\ (i \ne j)} \left(1 - F\left(\frac{x_i + x_j}{2}\right)\right).$$

The three products contain, respectively, $\binom{r}{2}$, $(r-k)(k-1)$, and $\binom{r-k+1}{2}$ factors. The total number of the factors is

$$N(r,k) = \binom{r}{2} + (r-k)(k-1) + \binom{r-k+1}{2}.$$

Now

$$\sum_{1 \le i,\, j \le r\ (i \ne j)} \frac{x_i + x_j}{2} = \frac{r-1}{2} \sum_{i=1}^{r} x_i, \quad \sum_{r \le i,\, j \le 2r-k\ (i \ne j)} \frac{x_i + x_j}{2} = \frac{r-k}{2} \sum_{i=r}^{2r-k} x_i,$$

$$\sum_{r < i \le 2r-k,\, r-k+1 \le j \le r} \frac{x_i + x_j}{2} = \frac{k-1}{2} \sum_{i=r+1}^{2r-k} x_i + \frac{r-k}{2} \sum_{j=r-k+1}^{r-1} x_j.$$

The total sum of the fractions $\frac{1}{2}(x_i + x_j)$ is

$$\tfrac{1}{2}(r-1)s_1 + \tfrac{1}{2}(2r-k-1)s_2 + \tfrac{1}{2}(r-1)s_3,$$

where

$$s_1 = \sum_{i=1}^{r-k} x_i, \qquad s_2 = \sum_{i=r-k+1}^{r} x_i, \qquad s_3 = \sum_{i=r+1}^{2r-k} x_i,$$

and the sum of the coefficients α_i by x_i in the sum of those fractions is $N(r,k)$. By the log-concavity of $1 - F(x)$,

$$\Psi(\boldsymbol{x}) \le \left(1 - F\left(\frac{((r-1)/2)s_1 + ((2r-k-1)/2)s_2 + ((r-1)/2)s_3}{N(r,k)}\right)\right)^{N(r,k)}.$$

https://doi.org/10.1017/jpr.2018.74 Published online by Cambridge University Press

As in the proof of Theorem 2.1, introduce $y_i = F(x_i)$, $1 \le i \le 2r-k$, so that

$$s_1 = \sum_{i=1}^{r-k} F^{-1}(y_i), \qquad s_2 = \sum_{i=r-k+1}^{r} F^{-1}(y_i), \qquad s_3 = \sum_{i=r+1}^{2r-k} F^{-1}(y_i).$$

Given t_1, t_2, and t_3, by the convexity of F^{-1}, we have

$$\min\left\{\sum_{i=1}^{2r-k} \frac{\alpha_i}{N(r,k)} F^{-1}(y_i)\colon \sum_{i=1}^{r-k} y_i = t_1, \sum_{i=r-k+1}^{r} y_i = t_2, \sum_{i=r+1}^{2r-k} y_i = t_3\right\}$$
$$\ge F^{-1}\left(\frac{((r-1)/2)t_1}{N(r,k)} + \frac{((2r-k-1)/2)t_2}{N(r,k)} + \frac{((r-k)/2)t_3}{N(r,k)}\right).$$

Consequently,

$$\Psi(\boldsymbol{x}) \le \Psi^*(\boldsymbol{t}) \coloneqq \left(1 - \frac{(r-1)t_1 + (2r-k-1)t_2 + (r-1)t_3}{2N(r,k)}\right)^{N(r,k)},$$
$$t_1 \coloneqq \sum_{i=1}^{r-k} F(x_i), \qquad t_2 \coloneqq \sum_{i=r-k+1}^{r} F(x_i), \qquad t_3 \coloneqq \sum_{i=r+1}^{2r-k} F(x_i).$$

Therefore,

$$\mathbb{P}(D_{I_1} \cap D_{I_2}) = \int\limits_{\boldsymbol{x}\in[0,1]^{2r-k}} \cdots \int \Psi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \le \int\limits_{t_1\le r-k,\, t_2\le k,\, t_3\le r-k} \cdots \int \Psi^*(\boldsymbol{t})\,\mathrm{d}\boldsymbol{y}$$
$$= \iiint\limits_{t_1,t_3\le r-k,\, t_2\le k} \Psi^*(\boldsymbol{t}) \frac{t_1^{r-k-1}}{(r-k-1)!} \frac{t_2^{k-1}}{(k-1)!} \frac{t_3^{r-k-1}}{(r-k-1)!}\,\mathrm{d}\boldsymbol{t}. \tag{2.4}$$

Introduce

$$\tau_1 = \frac{r-1}{2N(r,k)} t_1, \qquad \tau_2 = \frac{2r-k-1}{2N(r,k)} t_2, \qquad \tau_3 = \frac{r-1}{2N(r,k)} t_3.$$

Since $t_1 \le r-k$, $t_2 \le k$, $t_3 \le r-k$, and

$$\frac{(r-1)(r-k)}{2N(r,k)} + \frac{(2r-k-1)k}{2N(r,k)} + \frac{(r-1)(r-k)}{2N(r,k)} = 1,$$

we see that $\tau_1 + \tau_2 + \tau_3 \le 1$. Switching to τ_j, and denoting $N = N(r,k)$, we transform (2.4) into

$$\mathbb{P}(D_{I_1} \cap D_{I_2}) \le \frac{(2N/(r-1))^{r-k-1}}{(r-k-1)!} \frac{(2N/(2r-k-1))^{k-1}}{(k-1)!} \frac{(2N/(r-1))^{r-k-1}}{(r-k-1)!}$$
$$\times \iiint\limits_{\tau_1+\tau_2+\tau_3\le 1} \tau_1^{r-k-1}\tau_2^{k-1}\tau_3^{r-k-1}(1-\tau_1-\tau_2-\tau_3)^N\,\mathrm{d}\tau_1\,\mathrm{d}\tau_2\,\mathrm{d}\tau_3$$
$$= \frac{N!\,(2N/(r-1))^{2(r-k-1)}(2N/(2r-k-1))^{k-1}}{(N+2r-k)!}$$
$$\le N^{-3}\left(\frac{2}{r-1}\right)^{2(r-k-1)}\left(\frac{2}{2r-k-1}\right)^{k-1}.$$

https://doi.org/10.1017/jpr.2018.74 Published online by Cambridge University Press

(At the penultimate line we use the multidimensional extension of the beta integral; see [1, Theorem 1.8.6].) Since $N = \Theta(r^2)$, we have

$$\mathbb{P}(D_{I_1} \cap D_{I_2}) = O(\mathbb{P}(r,k)), \qquad \mathbb{P}(r,k) := r^{-6}\left(\frac{2}{r}\right)^{2(r-k-1)}\left(\frac{2}{2r-k}\right)^{k-1}. \tag{2.5}$$

2.3. Likely range of the maximum size of the K-set

We introduce $L_n = \{\max |I| \colon (1.3) \text{ holds}\}$. Kingman [12]–[14] proved that, for $F = \text{uniform}[0,1]$, w.h.p. $L_n \le n^{1/2}(\epsilon + o(1))$, where $\epsilon = \xi^{-1/2}(1-\xi)^{-1/2}$ and $\xi = 0.7968\ldots$ is a positive root of $1 - \xi = \mathrm{e}^{-2\xi}$, so $\epsilon = 2.485\ldots$. The proof consists of showing that $\mathbb{P}(D_I) \le 1/|I|!$, and that

$$\mathbb{P}(L_n \ge r) \le \frac{(n)_s}{(r)_s}\mathbb{P}(D_I), \qquad |I| = s \le r.$$

This inequality sharpens the (first-order moment) bound $\mathbb{P}(L_n \ge r) \le \binom{n}{r}\mathbb{P}(D_I)$, $|I| = r$, by using the fact that every subset of a K-set is a K-set as well. Kingman also demonstrated that his *exact* formula $\mathbb{P}(D_I) = (2/(r+1))^r$ for the negative exponential distribution on $[0,\infty)$ implied a better bound

$$L_n \le n^{1/2}[(2\mathrm{e}^{-1})^{1/2}\epsilon + o(1)], \qquad (2\mathrm{e}^{-1})^{1/2}\epsilon = 2.1317\ldots.$$

Now, by Theorem 2.1, we have $\mathbb{P}(D_I) \le \frac{1}{2}\mathrm{e}(2/r)^r$ for a wide class of densities on $[0,1]$ that includes the uniform density and the exponential density restricted to $[0,1]$. Combining this theorem and Kingman's proof for the exponential distribution, we have our next result.

Theorem 2.3. *Under conditions (i) and (ii) of Theorem 2.1, w.h.p.* $L_n \le 2.14n^{1/2}$.

Armed with bound (2.5) and the bounds in Theorem 2.1, we can prove a reasonably matching lower bound.

Theorem 2.4. *Let* $X_{n,r}$ *denote the total number of K-sets of cardinality* r. *Introduce* $r(n) = \lfloor 2n^{1/3} \rfloor$. *Then, under conditions (i)–(iii) of Theorem 2.2,*

$$\mathbb{P}\left(\bigcap_{\rho=2}^{r(n)}\left\{\left|\frac{X_{n,\rho}}{\mathbb{E}[X_{n,\rho}]} - 1\right| \le n^{-1/6+\varepsilon}\right\}\right) = 1 - O(n^{-2\varepsilon}) \quad \textit{for all } \varepsilon \in \left(0, \tfrac{1}{6}\right).$$

Consequently, $\min_{r\in[2,r(n)]} X_{n,r} \to \infty$ *in probability, and so*

$$\lim_{n\to\infty} \mathbb{P}(L_n \ge \lfloor 2n^{1/3} \rfloor) = 1.$$

Proof. This time we use the second-order moment approach. Using Theorem 2.1, for a generic set I of cardinality $r \in [2, r(n)]$ we have

$$\mathbb{E}[X_{n,r}] = \binom{n}{r}\mathbb{P}(D_I) \ge \frac{n^r}{r!}\left(\frac{2}{r+1}\right)^r \ge cn^2.$$

The total number of ordered pairs $\{I_1, I_2\}$, with $|I_1| = |I_2| = r$, $|I_1 \cap I_2| = k$, is

$$\mathcal{N}(r,k) = \binom{n}{r}\binom{r}{k}\binom{n-r}{r-k}.$$

https://doi.org/10.1017/jpr.2018.74 Published online by Cambridge University Press

Therefore, for a pair of generic sets I_1, I_2 meeting the conditions above,

$$\mathbb{E}[(X_{n,r})_2] = \sum_{k=0}^{r-1} \mathcal{N}(r,k)\mathbb{P}(D_{I_1} \cap D_{I_2}). \tag{2.6}$$

Here, $\mathbb{P}(D_{I_1} \cap D_{I_2}) = O(\mathbb{P}(r,k))$ with $\mathbb{P}(r,k)$ given in (2.5). After some elementary computation, we obtain

$$\begin{aligned}
&\max_{k\in[1,r-1]} \frac{\mathcal{N}(r,k+1)\mathbb{P}(r,k+1)}{\mathcal{N}(r,k)\mathbb{P}(r,k)} \\
&\quad\le \frac{r^2}{2} \max_{k\in[1,r-1]} \frac{(r-k)^2}{(k+1)(2r-k-1)(n-2r+k+1)} \exp\left(\frac{k-1}{2r-k-1}\right) \\
&\quad= \frac{r^2(r-1)}{8(n-2r+2)} \\
&\quad\le \frac{1-1/r}{1-2r/n} \\
&\quad\le 1-\frac{1}{2r}
\end{aligned}$$

(the second line maximum is attained at $k=1$). Consequently,

$$\begin{aligned}
\sum_{k=1}^{r-1} \mathcal{N}(r,k)\mathbb{P}(r,k) &\le 2r\mathcal{N}(r,1)\mathbb{P}(r,1) \\
&\le 2\binom{n}{r}\binom{n-r}{r-1}\left(\frac{2}{r}\right)^{2r} \\
&= O\left(\frac{r}{n}\mathcal{N}(r,0)\mathbb{P}^2(D_I)\right) \\
&= O\left(\frac{r}{n}\mathbb{E}^2[X_{n,r}]\right).
\end{aligned}$$

(For the last equality we use the lower bound for $\mathbb{P}(D_I)$ in Theorem 2.1.) Therefore, uniformly for $r \in [2, r(n)]$,

$$\frac{\sum_{k=1}^{r-1} \mathcal{N}(r,k)\mathbb{P}(r,k)}{\mathbb{E}^2[X_{n,r}]} = O(n^{-2/3}). \tag{2.7}$$

From (2.6), (2.7), and $\mathbb{E}[X_{n,r}] \ge c_1 n^2 \gg n^{2/3}$, we have

$$\frac{\mathbb{E}[(X_{n,r})_2]}{\mathbb{E}^2[X_{n,r}]} = 1 + O(n^{-2/3}) \quad\Longrightarrow\quad \frac{\mathrm{var}(X_{n,r})}{\mathbb{E}^2[X_{n,r}]} = O(n^{-2/3}).$$

By Chebyshev's inequality,

$$\mathbb{P}\left(\left|\frac{X_{n,r}}{\mathbb{E}[X_{n,r}]} - 1\right| \le \delta\right) \ge 1 - O(\delta^{-2}n^{-2/3}) \to 1$$

uniformly for all $\delta \gg n^{-1/3}$ and $r \in [2, r(n)]$. Therefore,

$$\sum_{r=2}^{r(n)} \mathbb{P}\left(\left|\frac{X_{n,r}}{\mathbb{E}[X_{n,r}]} - 1\right| \ge \delta\right) = O(\delta^{-2}n^{-1/3}) \to 0,$$

https://doi.org/10.1017/jpr.2018.74 Published online by Cambridge University Press

which implies that, for $\varepsilon \in (0, \frac{1}{6})$,

$$\mathbb{P}\Bigg(\bigcap_{r=2}^{r(n)}\Bigg\{\Bigg|\frac{X_{n,r}}{\mathbb{E}[X_{n,r}]}-1\Bigg| \le n^{-1/6+\varepsilon}\Bigg\}\Bigg) = 1 - O(n^{-2\varepsilon}). \qquad \square$$

2.4. Estimate of $\mathbb{P}(\mathcal{D}_I)$

Recall that the event $\mathcal{D}_I$ happens if and only if I is a K-set *and* no $J \subset I$, with $|J| \ge 2$, supports a local equilibrium $\boldsymbol{p} = \{p_i\}_{i\in J} > \mathbf{0}$ (with $\sum_{i\in J} p_i = 1$).

Let the event D_I hold, so that $f_{u,v} \ge \frac{1}{2}(f_{u,u} + f_{v,v})$ for all $u, v \in I$. So D_J holds for every $J \subseteq I$. Suppose that, for some $i \ne j$ in I, we have $f_{i,j} > \max\{f_{i,i}, f_{j,j}\}$. Set $J = \{i, j\}$ and

$$p_i := \frac{f_{i,j} - f_{j,j}}{2f_{i,j} - f_{i,i} - f_{j,j}} > 0, \qquad p_j = \frac{f_{i,j} - f_{i,i}}{2f_{i,j} - f_{i,i} - f_{j,j}} > 0.$$

Then $\boldsymbol{p} = (p_i, p_j)$ is a nontrivial local equilibrium, and this cannot happen on the event $\mathcal{D}_I$. Thus,

$$\mathcal{D}_I \subseteq \bigcap_{(i\ne j)\colon i,j\in I}\Bigg\{\frac{f_{i,i} + f_{j,j}}{2} \le f_{i,j} \le \max\{f_{i,i}, f_{j,j}\}\Bigg\}.$$

Consequently, we obtain

$$\mathbb{P}(\mathcal{D}_I \mid f_{i,i} = x_i,\ i \in I) \le \prod_{(i\ne j)\colon i,j\in I}\Bigg[F\Big(\max\{x_i, x_j\}\Big) - F\Bigg(\frac{x_i + x_j}{2}\Bigg)\Bigg]. \tag{2.8}$$

Introduce $y_i = F(x_i)$, i.e. $x_i = F^{-1}(y_i)$ (with $i \in I$). Then $F(\max\{x_i, x_j\}) = \max\{y_i, y_j\}$, and (since $F^{-1}(y)$ is convex),

$$F\big(\tfrac{1}{2}(x_i + x_j)\big) = F\big(\tfrac{1}{2}(F^{-1}(y_i) + F^{-1}(y_j))\big) \ge F\big(F^{-1}\big(\tfrac{1}{2}(y_i + y_j)\big)\big) = \tfrac{1}{2}(y_i + y_j).$$

Therefore,

$$\mathbb{P}(\mathcal{D}_I \mid f_{i,i} = x_i,\ i \in I) \le \prod_{(i\ne j)\colon i,j\in I}\frac{|y_i - y_j|}{2},$$

implying

$$\mathbb{P}(\mathcal{D}_I) \le 2^{-r(r-1)/2}\int\cdots\int_{\boldsymbol{y}\in[0,1]^r}\prod_{(i\ne j)\colon i,j\in I}|y_i - y_j|\,\mathrm{d}\boldsymbol{y}, \qquad r := |I|. \tag{2.9}$$

Since the integral is below 1, we see that

$$\mathbb{P}(\mathcal{D}_I) \le 2^{-r(r-1)/2}. \tag{2.10}$$

Hence, we arrive at our next corollary.

Corollary 2.2. *With probability greater than or equal to $1 - n^{-a}$ (for all $a > 0$), there is no K-set of cardinality greater than or equal to $r_n := \lceil 2\log_2 n\rceil$ that does not contain, properly, the support of a nontrivial local equilibrium.*

https://doi.org/10.1017/jpr.2018.74 Published online by Cambridge University Press

Proof. By (2.10), the expected number of K-sets in question is, at most, of order

$$\binom{n}{r_n}2^{-r_n^2/2} \le \frac{1}{r_n!} \le n^{-a} \quad \text{for all } a > 0. \qquad \square$$

We can do better though. The integral in (2.9) is a special case of Selberg's remarkable integral; see [1, Section 8.1]. In particular, for $\alpha > 0$, $\beta > 0$, $\gamma \ge 0$,

$$\int\cdots\int_{\boldsymbol{y}\in[0,1]^r} \prod_{i\in I}\{y_i^{\alpha-1}(1-y_i)^{\beta-1}\} \prod_{(i\ne j)\colon i,j\in I} |y_i - y_j|^{2\gamma}\,\mathrm{d}\boldsymbol{y}$$
$$= \prod_{j=1}^{r} \frac{\Gamma(\alpha+(j-1)\gamma)\Gamma((\beta+(j-1)\gamma)\Gamma(1+j\gamma)}{\Gamma(\alpha+\beta+(r+j-2)\gamma)\Gamma(1+\gamma)}. \tag{2.11}$$

So we have

$$\mathbb{P}(\mathcal{D}_I) \le 2^{-r(r-1)/2}\mathcal{S}(r), \qquad \mathcal{S}(r) := \prod_{j=1}^{r} \frac{\Gamma^2(1+(j-1)/2)\Gamma(1+j/2)}{\Gamma(1+(r+j)/2)\Gamma(3/2)}.$$

Using the Stirling formula,

$$\Gamma(1+z) = \sqrt{2\pi z}\left(\frac{z}{\mathrm{e}}\right)^z(1+O(z^{-1})), \qquad z\to\infty,$$

and applying the Euler summation formula to the logarithm of the resulting product, it follows that, for some constants η_1, η_2,

$$\mathcal{S}(r) = 2^{-r^2}\exp(\eta_1 r\log r + \eta_2 r + O(\log r)). \tag{2.12}$$

We have proved our next lemma.

Lemma 2.4. *There exist constants η_1^*, η_2^* such that*

$$\mathbb{P}(\mathcal{D}_I) \le 2^{-3r^2/2}\exp(\eta_1^* r\log r + \eta_2^* r + O(\log r)), \qquad r := |I|.$$

So $\mathbb{P}(\mathcal{D}_I)$ is of order $2^{-(3(1+o(1))/2)r^2}$, at most. This leads immediately to a better upper bound for the maximum size of a K-set free of supports of local equilibriums.

Theorem 2.5. *With probability greater than or equal to $1-\exp(-\Theta(\varepsilon\log^2 n))$, there is no K-set of cardinality $\ge r_n^* := \lceil(\frac{2}{3}+\varepsilon)\log_2 n\rceil$ that does not contain the support of a nontrivial local equilibrium.*

The sharp formula (2.12) allows us to show that w.h.p. there exist many K-sets of the logarithmic size that do not contain the size 2 supports of local equilibriums in the case when $f_{i,j}$ are uniform.

Given a set I, $|I|\ge 3$, let $\mathcal{D}_I^*$ be the event that I is a K-set meeting the above, less stringent, requirement. For brevity, we call such I a K*-set. Instead of inequality (2.8), here we have the equality

$$\mathbb{P}(\mathcal{D}_I^* \mid f_{i,i} = x_i,\ i\in I) = \prod_{(i\ne j)\colon i,j\in I}\left[F\Big(\max\{x_i,x_j\}\Big) - F\left(\frac{x_i+x_j}{2}\right)\right]. \tag{2.13}$$

https://doi.org/10.1017/jpr.2018.74 Published online by Cambridge University Press

For the uniform fitnesses, the right-hand side of (2.13) is the product of the $\frac{1}{2}|x_i - x_j|$. So, by (2.11) and (2.12),

$$\mathbb{P}(\mathcal{D}_I^*) = 2^{-3\rho^2/2} \exp(\eta_1^* \rho \log \rho + \eta_2^* \rho + O(\log \rho)), \qquad \rho := |I|.$$

Let $X_{n,r}^*$ denote the total number of the K*-sets of cardinality r. Then the expected number of the K*-sets of cardinality r is $\mathbb{E}[X_{n,r}^*] = \binom{n}{r}\mathbb{P}(\mathcal{D}_I^*)$ (with $|I| = r$). This expectation is easily shown to be of order greater than or equal to $\exp(\Theta(\varepsilon \log^2 n))$; thus it is super-polynomially large if $r = [\frac{2}{3}(1-\varepsilon)\log_2 n]$, $\varepsilon \in (0,1)$. In fact, we are about to prove that $X_{n,r}^*$ is likely to be this large if $r < \frac{1}{2}\log_2 n$.

Theorem 2.6. *For $r = [(\frac{1}{2} - \varepsilon)\log_2 n]$ (with $\varepsilon < \frac{1}{4}$), we have*

$$\mathbb{P}(X_{n,r}^* \geq \exp(\Theta(\varepsilon \log^2 n))) \geq 1 - O(n^{-2\varepsilon + O(\log\log n/(\log n))}).$$

Proof. We use the proof of Theorem 2.4 as a rough template. Given $0 \leq k \leq r-1$, let

$$I_1 = I_1(k) \equiv \{1, \ldots, r\}, \qquad I_2 = I_2(k) = \{r-k+1, \ldots, 2r-k\};$$

so $|I_1| = |I_2| = r$ and $|I_1 \cap I_2| = k$. Then, by symmetry,

$$\mathbb{E}[(X_{n,r}^*)_2] = \sum_{k=0}^{r-1} \mathcal{N}(r,k)\mathbb{P}(\mathcal{D}_{I_1(k)}^* \cap \mathcal{D}_{I_2(k)}^*), \qquad \mathcal{N}(r,k) = \binom{n}{r}\binom{r}{k}\binom{n-r}{r-k}.$$

To bound $\mathbb{P}(\mathcal{D}_{I_1(k)}^* \cap \mathcal{D}_{I_2(k)}^*)$, observe that, denoting by (i,j) a generic, unordered pair $(i \neq j)$, we have

$$\begin{aligned}
&\mathbb{P}(\mathcal{D}_{I_1(k)}^* \cap \mathcal{D}_{I_2(k)}^* \mid f_{i,i} = x_i, i \in I_1 \cup I_2) \\
&\quad = \prod_{(i,j):\, i,j \in [1,r] \cup [r-k+1,\, 2r-k]} \frac{|x_i - x_j|}{2} \\
&\quad \leq 2^{-(r)_2 + \binom{k}{2}} \prod_{(i,j):\, i,j \in [1,r-k]} |x_i - x_j| \prod_{(i,j):\, i,j \in [r-k+1,r]} |x_i - x_j| \\
&\qquad \times \prod_{(i,j):\, i,j \in [r+1,\, 2r-k]} |x_i - x_j|.
\end{aligned}$$

Unconditioning and using (2.13), we obtain $\mathbb{P}(\mathcal{D}_{I_1(k)}^* \cap \mathcal{D}_{I_2(k)}^*) = \mathcal{P}^*(r,k)\mathrm{e}^{O(\log r)}$, and

$$\begin{aligned}
\mathcal{P}^*(r,k) &:= 2^{-(r)_2 + \binom{k}{2}} \cdot 2^{-2(r-k)^2 - k^2} \\
&\quad \times \exp(2\eta_1^*(r-k)\log(r-k) + \eta_1^* k \log k + 2\eta_2^*(r-k) + \eta_2^* k).
\end{aligned}$$

It follows that

$$\frac{\mathcal{N}(r,k+1)\mathbb{P}(\mathcal{D}_{I_1(k+1)}^* \cap \mathcal{D}_{I_2(k+1)}^*)}{\mathcal{N}(r,k)\mathbb{P}(\mathcal{D}_{I_1(k)}^* \cap \mathcal{D}_{I_2(k)}^*)} \leq \frac{2^{2r}}{n}\exp(O(\log r)) \leq n^{-2\varepsilon + o(1)} \to 0,$$

since $r \leq (\frac{1}{2} - \varepsilon)\log_2 n$. Consequently,

$$\frac{\mathbb{E}[(X_{n,r}^*)_2]}{\mathcal{N}(r,0)\mathbb{P}^2(\mathcal{D}_{I_1(0)}^*)} \leq 1 + n^{-2\varepsilon + o(1)}.$$

https://doi.org/10.1017/jpr.2018.74 Published online by Cambridge University Press

Since

$$\mathcal{N}(r,0)=\left(1+O\left(\frac{r^2}{n}\right)\right)\binom{n}{r}^2, \qquad \mathbb{E}[X^*_{n,r}]=\mathbb{P}(\mathcal{D}_{I_1^*}(0))\binom{n}{r}\ge \exp(\Theta(\log^2 n)),$$

the Chebyshev inequality completes the proof. □

Acknowledgements

About thirty years ago, John Kingman gave a lecture on stable polymorphisms at Stanford University. The talk made a deep, lasting impression on me. At that time Don Knuth introduced me to his striking formula for the expected number of stable matchings via a highly dimensional integral (see [15]). Despite the world of difference between the stable polymorphisms and the stable matchings, the multidimensional integrals expressing the probability of respective stability conditions are of a similar kind. I am grateful to the referees for their meticulous reading of the manuscript and valuable suggestions, and I thank the Editor for advice on making the original version worthy of the editorial boards consideration.

References

[1] Andrews, G. E., Askey, R. and Roy, R. (1999). *Special Functions*. Cambridge University Press.
[2] Baum, L. E. and Eagon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bull. Amer. Math. Soc.* **73,** 360–363.
[3] Chen, X. and Peng, J. (2015). New analysis on sparse solutions to random standard quadratic optimization problems and extensions. *Math. Operat. Res.* **40,** 725–738.
[4] Chen, X., Peng, J. and Zhang, S. (2013). Sparse solutions to random standard quadratic optimization problems. *Math. Program.* (Ser. A **141**), 273–293.
[5] Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, 2nd edn. John Wiley, Oxford.
[6] Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford Clarendon Press.
[7] Haigh, J. (1988). The distribution of evolutionarily stable strategies. *J. Appl. Prob.* **25,** 113–125.
[8] Haigh, J. (1989). How large is the support of an ESS. *J. Appl. Prob.* **26,** 164–170.
[9] Hofbauer, J. and Sigmund, K. (1998). *Evolutionary Games and Replicator Dynamics*. Cambridge University Press.
[10] Kingman, J. F. C. (1961). A mathematical problem in population genetics. *Proc. Camb. Phil. Soc.* **57,** 574–582.
[11] Kingman, J. F. C. (1961). A matrix inequality. *Quart. J. Math.* **12,** 78–80.
[12] Kingman, J. F. C. (1988). Typical polymorphisms maintained by selection at a single locus. *J. Appl. Prob.* **25,** 113–125.
[13] Kingman, J. F. C. (1989). Maxima of random quadratic forms on a simplex. In *Probability, Statistics and Mathematics.* Academic Press, Boston, MA, pp. 123–140.
[14] Kingman, J. F. C. (1990). Some random collections of finite sets. In *Disorder in Physical Systems, A Volume in Honour of John M. Hammersley*, Oxford University Press, pp. 241–247.
[15] Knuth, D. E. (1996). Stable marriage and its relation to other combinatorial problems: an introduction to the mathematical analysis of algorithms. *CRM Proc. Lecture Notes* **10,** 74 pp.
[16] Kontogiannis, S. C. and Spirakis, P. G. (2009). On the support size of stable strategies in random games. *Theoret. Comput. Sci.* **410,** 933–942.
[17] Lewontin, R. C., Ginzburg, L. R. and Tuljapurkar, S. D. (1978). Heterosis as an explanation for large amounts of genic polymorphism. *Genetics* **88,** 149–170.
[18] Pittel, B. (1989). The average number of stable matchings. *SIAM J. Discrete Math.* **2,** 530–549.

Pittel, B. (2018). On random stable partitions. To appear in *Internat. J. Game Theory*. Available at https://doi.org/10.1007/s00182-018-0635-9.
[19] Scheuer, P. and Mandel, S. (1959). An inequality in population genetics. *Heredity* **13,** 519–524.