Threshold phenomena in k-dominant skylines of random samples

HSIEN-KUEI HWANG
Institute of Statistical Science
Academia Sinica
Taipei 115
Taiwan

TSUNG-HSI TSAI
Institute of Statistical Science
Academia Sinica
Taipei 115
Taiwan

WEI-MEI CHEN
Department of Electronic Engineering
National Taiwan University of Science and Technology
Taipei 106
Taiwan

November 7, 2018

Abstract

Skylines emerged as a useful notion in database queries for selecting representative groups in multivariate data samples for further decision making, multi-objective optimization or data processing, and the k-dominant skylines were naturally introduced to resolve the abundance of skylines when the dimensionality grows or when the coordinates are negatively correlated. We prove in this paper that the expected number of k-dominant skylines is asymptotically zero for large samples when $1 \leq k \leq d-1$ under two reasonable (continuous) probability assumptions of the input points, d being the (finite) dimensionality, in contrast to the asymptotic unboundedness when $k = d$. In addition to such an asymptotic zero-infinity property, we also establish a sharp threshold phenomenon for the expected $(d-1)$-dominant skylines when the dimensionality is allowed to grow with n. Several related issues such as the dominant cycle structures and numerical aspects, are also briefly studied.

Key words. Skyline, dominance, maxima, random samples, Pareto optimality, threshold phenomena, multi-objective optimization, computational geometry, asymptotic approximations, average-case analysis of algorithms.

1 Introduction

The last decade has undergone a drastic change of information dissemination from Web 1.0 to Web 2.0, the most notable representative products being YouTube and Facebook. Data have

been generated in an unprecedented pace and range, powerful search engines are indispensable, and screening useful or usable information (via "sort engines") from the vast is generally becoming more important than searching and gathering. Skylines of multivariate data sample were introduced for selecting representative groups in the database query literature by Börzsönyi et al. (see [7]) and had appeared in diverse areas under several different guises and names: *Pareto optimality*, *efficiency*, *maxima*, *admissibility*, *elite*, *sink*, etc.; see [11, 12] and the references therein for more information. These diverse terms reveal the importance of the use of skyline as an effective means of data summarization in theory and in practice. Many different notions and variants of skylines have been proposed in the literature, following the original paper [7]. In particular, the k-dominant skylines were introduced by Chan et al. (see [9]) in situations when the skylines are abundant and have received much attention since, although they had already been studied in the Russian literature (see for example [3, 23]). We focus in this paper on the asymptotic estimates of such skylines and prove several types of threshold phenomena under different probability assumptions of the input samples, which, in addition to their theoretical interests, are believed to be useful for practitioners.

Skylines and k-dominant skylines The definitions of skyline and many of its variants are based on the notion of dominance. Given a d-dimensional dataset $\mathscr{D}$, a point $\mathbf{p} \in \mathscr{D}$ is said to *dominate* another point $\mathbf{q} \in \mathscr{D}$ if $p_j \le q_j$ for $1 \le j \le d$, where $\mathbf{p} = (p_1, \dots, p_n)$ and $\mathbf{q} = (q_1, \dots, q_n)$, and is less than in at least one dimension. The non-dominated points in $\mathscr{D}$ are called the *skyline* (or *skyline points*) of $\mathscr{D}$. By relaxing the full dominance definition to partial dominance, we say that a point $\mathbf{p} \in \mathscr{D}$ *k-dominates* another point $\mathbf{q} \in \mathscr{D}$ if there are k dimensions in which p_j is not greater than q_j and is less than in at least one of these k dimensions[1]. The points in $\mathscr{D}$ that are not k-dominated by any other points are defined to be the *k-dominant skyline* of $\mathscr{D}$; see [9]. See also [3] for a different formulation.

The definition of k-dominant skyline implies that for a fixed dataset the number of k-dominant skylines decreases as k becomes smaller. Such a monotonicity property will be used later. To see this, consider any point $\mathbf{p}$ in the unit square. It is a skyline (or 2-dominant skyline) point if no other points have simultaneously smaller x- and smaller y-values; namely, no other points can lie in the shaded region ⊞ (where $\mathbf{p}$ is the dotted point in the middle of this figure). However, to be a 1-dominant skyline point requires that all other points must have simultaneously larger x- and larger y-values, or, equivalently, they cannot lie in the shaded region ⊞.

On the other hand, the transitivity property of skylines fails for k-dominant skylines when $1 \le k \le d-1$, meaning that their cardinality may be zero and there may be cycles.

The number of skyline points The number of skyline points is a key issue in their use and usefulness. This quantity under suitable random assumptions of the input is also important for practical modeling or reference purposes, as well as for the analysis of skyline-finding algorithms. The two major, simple, representative random models are *hypercubes* and *simplices*. Assuming that the input dataset $\mathscr{D} = \{\mathbf{p}_1, \dots, \mathbf{p}_n\}$ is taken uniformly and independently from the hypercube $[0,1]^d$, then it has been known since the 1960's (see [1]) that the expected number of skyline points of $\mathscr{D}$ is asymptotic to $\frac{(\log n)^{d-1}}{(d-1)!}$ for large n and finite d,

[1] If we change the definition of the k-dominant skyline to be "exactly k" (instead of $\ge k$) coordinates smaller than or equal to and at least 1 smaller than, then the same types of results in this paper also hold.

exhibiting the independence of the coordinates. (Intuitively, if one sorts according to one dimension, then each other dimension roughly contributes $\log n$ skyline points.) On the other hand, if we assume that the input points are uniformly sampled from the d-dimensional simplex $\{|x_1| + \cdots + |x_d| \leq 1, x_j \in (-1, 0]\}$, then the expected number of skyline points is asymptotic to $\Gamma\left(\frac{1}{d}\right) n^{1-\frac{1}{d}}$, reflecting obviously a stronger negative correlation of the coordinates; see [5] and the references cited there. Here Γ denotes Euler's Gamma function. For the number of skyline points under other models, see [2, 14, 15, 25] and the references therein.

On the other hand, in contrast to the recent growing trend of studying high dimensional datasets, not much is known for the expected number of skyline points when d is allowed to grow with n. Such a direction is especially useful as practical situations always deal with finite n and finite d (whose dependence on n is often not clear). The only exception along this direction is the uniform estimates given in [18] (see also [5]) for the expected number of skyline points in a random uniform samples of n points from the hypercube $[0, 1]^d$. While the order $\frac{(\log n)^{d-1}}{(d-1)!}$ may seem slowly growing as d increases, it soon reaches the order n when d is around $\log n$, which is relatively small for moderate values of n. Consequently, the skyline points become too numerous to be of direct use. The growth of skyline points in the random d-dimensional simplex model is even faster and we can show that almost all points are skylines when d roughly exceeds $\frac{\log n}{\log \log n}$, again small for n not too large.

The cardinality of k-dominant skyline Since k-dominant skyline were proposed (see [9]) to resolve the skyline-abundance problem, it is of interest to know their quantity under suitable random models. A critical step in applying k-dominant skyline is to identify an appropriate k such that the size of the k-dominant skyline is within the acceptable ranges. But this may not be always feasible. Consider the 5-dimensional dataset $\mathscr{D}$ given in Table 1. The six points are all skyline points, one ($\mathbf{p}_6$) is the 4-dominant skyline point and no point is in the 3-dominant skyline. Clearly, $\mathbf{p}_6$ is to some extent better than the other points since it contains two components with the lowest value 1. However, it was already mentioned in [9] that some k-dominant skylines may be empty. For example, if we drop $\mathbf{p}_6$ from $\mathscr{D}$, then the five points are all skyline points but all k-dominant skylines are empty for $1 \leq k \leq 4$. In this example, other alternatives to k-dominant skylines have to be used. Unfortunately, such a property of *excessive skylines but few k-dominant skylines* is not uncommon, and we show in this paper that, under the hypercube and the simplex random models, the expected number of k-dominant skylines both tends to zero for large n and $1 \leq k \leq d - 1$.

point	skyline	4-dominant skyline	3-dominant skyline
$\mathbf{p}_1\ (1, 2, 2, 3, 3)$	✔	-	-
$\mathbf{p}_2\ (3, 1, 2, 2, 3)$	✔	-	-
$\mathbf{p}_3\ (3, 3, 1, 2, 2)$	✔	-	-
$\mathbf{p}_4\ (2, 3, 3, 1, 2)$	✔	-	-
$\mathbf{p}_5\ (2, 2, 3, 3, 1)$	✔	-	-
$\mathbf{p}_6\ (2, 3, 1, 1, 3)$	✔	✔	-

Table 1: An example showing the property of many skylines but few k-dominant skylines.

Threshold phenomena We clarify two types of threshold phenomena for the expected number of k-dominant skylines in random samples.

1. *Large sample, bounded dimension*:

$$\text{Expected number of } k\text{-dominant skylines} \to \begin{cases} 0, & \text{if } 1 \le k \le d-1; \\ \infty, & \text{if } k = d, \end{cases}$$

as the sample size $n \to \infty$. While such a result is not new and contained as a special case of the general theory developed in [3] for finite dimensional skylines, we will give an independent, transparent, self-contained proof, which, in addition to being more precise, can be extended to the case when the dimensionality goes unbounded with the sample size.

2. *Large sample, moderate dimension*: There exists an integer $d_0 = d_0(n) \approx \sqrt{\frac{2\log n}{\log \frac{\log n}{\log\log n}}} + 1$ such that (see (23))

$$\text{Expected number of } (d-1)\text{-dominant skylines} \to \begin{cases} 0, & \text{if } d \le d_0 - 1; \\ \infty, & \text{if } d \ge d_0 + 2, \end{cases}$$

as $n \to \infty$, and the two cases $d = d_0$ and $d = d_0 + 1$ lead to two different oscillating functions, the first ($d = d_0$) fluctuating between 0 and $\frac{e^{-\gamma}}{2-e^{-e^{-1}}}$ and the second between $\frac{e^{-\gamma}}{2-e^{-e^{-1}}}$ and $O\left(\frac{\log n}{\log\log n}\right)$, where γ is Euler's constant; see (24) and (25). We consider only random samples from hypercubes. Other regions and other values of k, $k < d-1$ are expected to exhibit similar threshold phenomena with different d_0, but the analysis becomes excessively long and involved. More details will be discussed elsewhere.

We see from these phenomena that the usual "curse of high dimensionality" has thus another form here which one may term "curse of constant dimensionality," which refers to the situation when no k-dominant skyline point at all exists. Also the model where dimensionality can vary with the sample size is, at least from a practical point of view, more reasonable; see Sections 6 and 7 for more discussions and details.

Related works In addition to the partial dominance used in defining k-dominant skylines (see [9]), there are also several other skyline variants for retrieving more representative points; these include skybands [24], top-k dominating queries [20, 24, 27], strong skylines [28], skyline frequency [10], approximately dominating representatives [21], ε-skylines [26], and top-k skylines [8, 22]. See also the survey paper [20] for more information.

Organization of the paper This paper presents a systematic study on the asymptotic estimates of the number of k-dominant skyline points under random models. It is organized as follows. We derive in the next section (§ 2) an asymptotic vanishing property for the number of k-dominant skyline points under a common hypercube model when the dimensionality is bounded. The extension to include more points in the partial dominant skyline is showed to suffer from a similar drawback in Section 3. We then prove in Section 4 that changing the underlying model from hypercube to simplex does not improve either the asymptotic vanishing property. Section 5 deals with a categorical model for which the results have a very different

nature. Roughly, as the total number of sample points are finite in this model, the expected number of k-dominant skylines will be asymptotically linear, meaning too many choices for ranking or selection purposes. All these results point to the negative side for the use of k-dominant skylines under similar data situations. We then address the positive side in the last few sections by considering again the hypercubes but with growing dimensionality. A sharp threshold phenomenon is discovered in Section 7 when $d \to \infty$ with n, the asymptotic approximations needed being derived in Section 6. Another new threshold result is given in Section 8 of the expected number of dominant cycles. Section 9 provides a uniform lower-bound estimate for the expected number of skyline points for $1 \le k \le d-1$. We conclude in Section 10 with some numerical aspects of the estimates we derived.

2 Random samples from hypercubes

The simplest random model is the hypercube $[0,1]^d$, which is also the most natural and most studied one. They can also be used when data are discrete in nature but span uniformly over a sufficiently large interval.

In this section, we derive asymptotic estimates for the expected number of k-dominant skyline points in a random sample of n points $\mathscr{D} := \{\mathbf{p}_1, \ldots, \mathbf{p}_n\}$ uniformly and independently drawn from $[0,1]^d$, $d \ge 2$. Let $M_{d,k}(n)$ denote the number of k-dominant skyline points of $\mathscr{D}$. We first derive a crude upper bound for the expected number $\mathbb{E}[M_{d,k}(n)]$, which implies that $\mathbb{E}[M_{d,k}(n)]$ is asymptotically zero as n grows unbounded and $1 \le k \le d-1$. More precise estimates are possible and will be derived in Section 6. For a point $\mathbf{p} \in [0,1]^d$, denoted by $B_k(\mathbf{p})$ the region of the points in $[0,1]^d$ that k-dominates $\mathbf{p}$. Also, $|A|$ denotes the volume of the region A.

Theorem 1 (Asymptotic zero-infinity property for large n and bounded d)**.** *For fixed* $d \ge 2$

$$\mathbb{E}[M_{d,k}(n)] \to \begin{cases} 0, & \text{if } 1 \le k \le d-1; \\ \infty, & \text{if } k = d, \end{cases} \tag{1}$$

as $n \to \infty$.

Proof. The case $k = d$ has been known since the 1960's (see [1]) and were re-derived several times in the literature. We assume $1 \le k \le d-1$. Since $M_{d,k}(n) \le M_{d,d-1}(n)$ for fixed d and for $1 \le k \le d-1$, we only prove that $\mathbb{E}[M_{d,d-1}(n)] \to 0$.

We start from the integral representation

$$\begin{aligned} \mathbb{E}[M_{d,d-1}(n)] &= n\mathbb{P}\left(\mathbf{p}_1 \text{ is a } (d-1)\text{-dominant skyline point}\right) \\ &= n \int_{[0,1]^d} \left(1 - |B_{d-1}(\mathbf{x})|\right)^{n-1} \mathrm{d}\mathbf{x}, \end{aligned} \tag{2}$$

because if $\mathbf{x}$ is not k-dominated by any of the other $n-1$ points, they all have to lie in the region $[0,1]^d \setminus B_k(\mathbf{x})$. Here and throughout this paper, $\mathrm{d}\mathbf{x}$ is the abbreviation of $\mathrm{d}x_1 \cdots \mathrm{d}x_d$.

To estimate the integral in (2), we split it into two parts, one part having sufficiently small volume (corresponding roughly to small $x_1 \cdots x_d$) and the other with $|B_{d-1}(\mathbf{x})|$ bounded away from zero, rendering the term $(1 - |B_{d-1}(\mathbf{x})|)^{n-1}$ also small.

For a fixed number t satisfying $1 < t < \frac{d}{d-1}$, define the region

$$Q_n := \bigcup_{1\le \ell \le d} \left\{ \mathbf{x} \in [0,1]^d : x_\ell \le n^{-\frac{t}{d}} \text{ and } \prod_{j\neq \ell} x_j \le n^{-\frac{d-1}{d}\,t} \right\}. \tag{3}$$

Then

$$\mathbb{E}[M_{d,d-1}(n)] \le n\,|Q_n| + n \int_{[0,1]^d \setminus Q_n} (1 - |B_{d-1}(\mathbf{x})|)^{n-1}\,\mathrm{d}\mathbf{x}.$$

The volume of Q_n is bounded above by

$$|Q_n| \le d n^{-\frac{t}{d}} \int_{\substack{x_1\cdots x_{d-1} \le n^{-\frac{d-1}{d}\,t} \\ \mathbf{x}\in[0,1]^d}} \mathrm{d}\mathbf{x}.$$

To estimate the last integral, let

$$A_d(\delta) := \int_{\substack{x_1\cdots x_{d-1}\le \delta \\ \mathbf{x}\in[0,1]^d}} \mathrm{d}\mathbf{x} \qquad (d \ge 2),$$

where $0 < \delta < 1$. Then $A_2(\delta) = \delta$, and

$$A_d(\delta) = \int_\delta^1 A_{d-1}\left(\frac{\delta}{t}\right) \mathrm{d}t \qquad (d \ge 3).$$

A simple induction gives

$$A_d(\delta) = \delta \frac{|\log \delta|^{d-2}}{(d-2)!} \qquad (d \ge 2),$$

and we obtain, by taking $\delta = n^{-\frac{d-1}{d}\,t}$,

$$|Q_n| = O\left(n^{-t}(\log n)^{d-2}\right),$$

On the other hand, by an inclusion-exclusion argument, we have

$$|B_{d-1}(\mathbf{x})| = \sum_{1\le \ell \le d} \prod_{j\neq \ell} x_j - (d-1) \prod_{1\le j\le d} x_j. \tag{4}$$

Now if $\mathbf{x} \in [0,1]^d \setminus Q_n$, then

$$|B_{d-1}(\mathbf{x})| \ge \max_{1\le \ell \le d} \prod_{i\neq \ell} x_i \ge n^{-\frac{d-1}{d}\,t}.$$

Thus, we have

$$\mathbb{E}[M_{d,d-1}(n)] = O\left(n^{1-t}(\log n)^{d-2}\right) + O\left(n \exp\left(-(n-1)n^{-\frac{d-1}{d}\,t}\right)\right), \tag{5}$$

and we see easily that the right-hand side tends to zero by our choice of t. More precisely, if we take

$$t = \frac{d}{d-1}\left(1 - \frac{\log\left(\frac{d}{d-1}\log n\right)}{\log n}\right),$$

so as to balance the two O-terms in (5), then

$$\mathbb{E}[M_{d,d-1}(n)] = O\left(n^{-\frac{1}{d-1}}(\log n)^d\right).$$

This and the monotonicity of $M_{d,k}(n)$ (in k) proves (1). ■

The fact that $\mathbb{E}[M_{d,k}(n)] \to 0$ implies that there are many cycles formed by the k-dominant relation, but the corresponding cycle structures are very difficult to quantify; see Section 10 for some preliminary results.

3 "Clouds" of k-dominant skylines

The asymptotic vanishing property (Theorem 1) for the expected number of k-dominant skylines limits their usefulness if the input data are known to be in similar randomness conditions. In particular, if one is interested in finding the top-K representative points, then the probability of getting enough number of candidates tends to zero. A simple remedy to this situation (and still following the same notion of partial dominance between points) is to consider the number of points that are k-dominated by a specified number, say j of other points, which we refer to as the "cloud" of k-dominant skylines. But we show that this also suffers from similar vanishing drawback under the random hypercube model, unless j is chosen to be large enough.

Let $L_{d,k}(n,j)$ denote the number of points in the random sample $\{\mathbf{p}_1,\ldots,\mathbf{p}_n\}$ that are k-dominated by exactly j points, where the n points are uniformly and independently selected from $[0,1]^d$. Note that $L_{d,k}(n,0)$ is nothing but $M_{d,k}(n)$.

Theorem 2 (Asymptotic zero-infinity property for clouds of k-dominant skylines)**.** *For fixed $d \geq 2$ and $1 \leq k \leq d-1$,*

$$\mathbb{E}[L_{d,k}(n,j)] \to \begin{cases} 0, & \text{if } 1 \leq k \leq d-1; \\ \infty, & \text{if } k = d, \end{cases}$$

uniformly for $0 \leq j = o(n^{(1-\varepsilon)/d})$, as $n \to \infty$, where $\varepsilon > 0$ is an arbitrarily small constant.

The theorem roughly says that even allowing more flexible partial dominance relation, the expected number of the skylines so constructed still approaches zero as long as the dimensionality is fixed.

Proof. The case when $k = d$ is also derived in [1] (under the name of "$(j+1)^{\text{st}}$ layer, 1-st quadrant-admissible points"), where it is showed that

$$\mathbb{E}[L_{d,d}(n,j)] = \sum_{j<i_1\leq\cdots\leq i_{d-1}\leq n} \frac{1}{i_1\cdots i_{d-1}},$$

from which we obtain

$$\mathbb{E}[L_{d,d}(n,j)] \sim \frac{\left(\log\frac{n}{j+1}\right)^{d-1}}{(d-1)!}, \tag{6}$$

if $\log(n/(j+1)) \to \infty$, where the symbol "$\sim$" means that the ratio of both sides tends to 1 as n goes unbounded. Alternatively, we can use the integral representation (see [4])

$$\begin{aligned}\mathbb{E}[L_{d,d}(n,j)] &= n\binom{n-1}{j}\int_{[0,1]^d}(x_1\cdots x_d)^j\,(1-x_1\cdots x_d)^{n-1-j}\,\mathrm{d}\mathbf{x}\\ &= \frac{n}{(d-1)!}\binom{n-1}{j}\int_0^1 t^j(1-t)^{n-1-j}\log\left(\tfrac{1}{t}\right)^{d-1}\mathrm{d}t, \qquad (7)\end{aligned}$$

by the change of variables $t \mapsto x_1\cdots x_d$. A straightforward evaluation then gives (6).

Note that $\frac{\mathbb{E}[L_{d,d}(n,j)]}{n}$ equals the probability that the first-quadrant subtree of the root has size j in random quadtrees; see [16, Appendix]. This connection also provides several other expressions for $\mathbb{E}[L_{d,d}(n,j)]$. For example,

$$\mathbb{E}[L_{d,d}(n,j)] = \binom{n-1}{j}\sum_{0\le\ell\le n-1-j}\binom{n-1-j}{\ell}\frac{(-1)^\ell}{(j+1+\ell)^d};$$

see also [5].

For the remaining cases, we consider only $k = d-1$ and prove that $\mathbb{E}[L_{d,d-1}(n,j)] \to 0$. The reason is that

$$\sum_{0\le\ell\le j} L_{d,k}(n,\ell) \le \sum_{0\le\ell\le j} L_{d,d-1}(n,\ell) \qquad (1\le k\le d-1).$$

To see this, observe that if a point $\mathbf{p}$ $(d-1)$-dominates another point $\mathbf{q}$, then $\mathbf{p}$ also k-dominates $\mathbf{q}$ for $1 \le k \le d-2$. Thus, the sum on the left-hand side, which stands for the set that is k-dominated by at most j points, is less than the sum on the right-hand side, the set that is $(d-1)$-dominated by at most j points.

To prove $\mathbb{E}[L_{d,d-1}(n,j)] \to 0$, we apply the same argument used in the proof of Theorem 1 starting from the integral representation

$$\begin{aligned}\mathbb{E}[L_{d,d-1}(n,j)] &= n\int_{[0,1]^d}\mathbb{P}(\text{exactly } j \text{ points in } \{\mathbf{p}_2,\dots,\mathbf{p}_n\} \text{ that } k\text{-dominate } \mathbf{p}_1)\\ &= n\binom{n-1}{j}\int_{[0,1]^d} B_{d-1}(\mathbf{x})^j\,(1-B_{d-1}(\mathbf{x}))^{n-1-j}\,\mathrm{d}\mathbf{x}.\end{aligned}$$

Now we fix a constant t satisfying $1 < t < \frac{d}{d-1}$, and then choose Q_n as in (3). Then we have

$$|Q_n| = O\left(n^{-t}(\log n)^{d-2}\right),$$

and

$$n^{-\frac{d-1}{d}t} \le |B_{d-1}(\mathbf{x})| \le 1 \qquad (\mathbf{x}\in[0,1]^d\setminus Q_n).$$

It follows that

$$\begin{aligned}\mathbb{E}[L_{d,d-1}(n,j)] &\le n|Q_n| + n\binom{n-1}{j}\int_{[0,1]\setminus Q_n} B_{d-1}(\mathbf{x})^j\,(1-B_{d-1}(\mathbf{x}))^{n-1-j}\,\mathrm{d}\mathbf{x}\\ &= O\left(n^{1-t}(\log n)^{d-2}\right) + O\left(n\binom{n-1}{j}\exp\left(-(n-1-j)n^{-\frac{d-1}{d}t}\right)\right).\end{aligned}$$

Now choose

$$t = \frac{d}{d-1}\left(1 - \frac{\log((j+\frac{d}{d-1})\log n)}{\log n}\right).$$

So that

$$n\binom{n-1}{j}\exp\left(-(n-1-j)n^{-\frac{d-1}{d}t}\right) = O\left(n^{1+j}n^{-j-\frac{d}{d-1}}\right) = O(n^{-\frac{1}{d-1}}),$$

and

$$n^{1-t} = n^{-\frac{1}{d-1}}\left(j+\tfrac{d}{d-1}\right)^{\frac{d}{d-1}}(\log n)^{\frac{d}{d-1}} = O\left(n^{-\frac{\varepsilon}{d-1}}(\log n)^{\frac{d}{d-1}}\right),$$

uniformly for $j = O(n^{\frac{1-\varepsilon}{d}})$. Thus

$$\mathbb{E}[L_{d,d-1}(n,j)] = O\left(n^{-\frac{\varepsilon}{d-1}}(\log n)^{d-2+\frac{d}{d-1}} + n^{-\frac{1}{d-1}}\right) \to 0.$$

This proves the theorem. ■

A more precise asymptotic estimate for $\mathbb{E}[L_{d,d-1}(n,j)]$ will be derived in Section 6; see (21). Another easy special case is $k=1$, which is dual to the case $k=d$ because we have

$$\mathbb{E}[L_{d,1}(n,j)] = \mathbb{E}[L_{d,d}(n,n-1-j)].$$

Thus, by (7), we have

$$\begin{aligned}\mathbb{E}[L_{d,1}(n,j)] &= \frac{n}{(d-1)!}\binom{n-1}{j}\int_0^1 t^{n-1-j}(1-t)^j(-\log t)^{d-1}\mathrm{d}t\\ &\sim \frac{n^{j+1}}{(d-1)!j!}\int_0^\infty e^{-nt}t^{j+d-1}\mathrm{d}t\\ &\sim \binom{j+d-1}{j}n^{-d+1},\end{aligned}$$

for large n and $0 \le j = o(\sqrt{n})$.

In general, if we are to select the top K representatives using such clusters of partial dominant skylines, then how large should j be? That is, what is the minimum m such that $\sum_{0\le j\le m} L_{d,k}(n,j) > K$? Some simulation results are given in Figure 1.

4 Random samples from simplices

We show in this section that the asymptotic vanishing property of k-dominant skylines occurs not only in the case of the d-dimensional hypercube distribution, but also in the d-dimensional simplex distribution

$$S_d = \left\{\mathbf{x} : -1 \le x_j \le 0 \text{ and } \|\mathbf{x}\| := \sum_{1\le j\le d}|x_j| \le 1\right\}.$$

In particular, S_2 is the right triangle ◹. Such a shape implies a negative dependence of the two coordinates and thus a larger number of skyline points.

Let $M_k^{[s]}(n)$ denote the cardinality of the k-dominant skyline of the set $\mathscr{D} := \{\mathbf{p}_1,\dots,\mathbf{p}_n\}$, where these n points are uniformly and independently distributed over S_d. For a point $\mathbf{p}\in S_d$, denote by $B_k^{[s]}(\mathbf{p})$ the region of points in S_d that k-dominate $\mathbf{p}$.

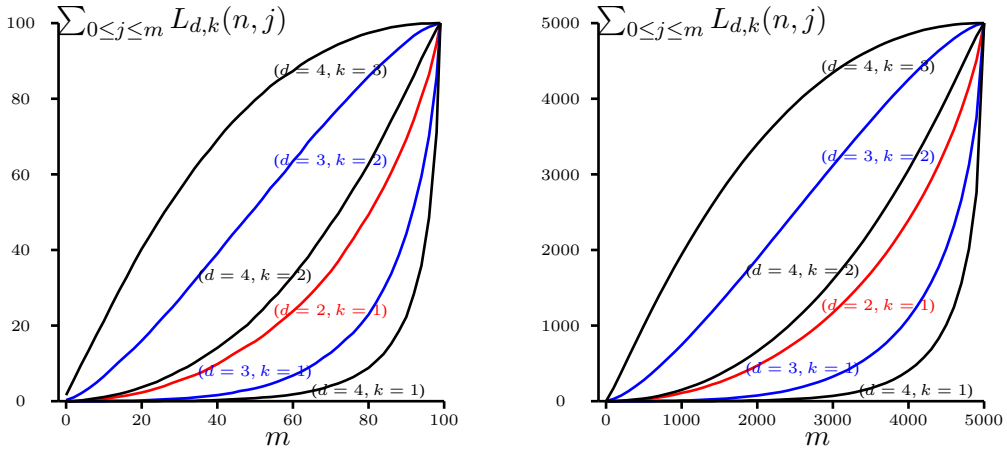

Figure 1: *Simulated values of* $\sum_{0\le j\le m} L_{d,k}(n,j)$ *for* $n = 100$ *(left) and* 5000 *(right). Interestingly, the simulations suggest some general pattern that seems independent of the size of the samples and they are consistent with our analysis since* m *has to be very large (compared with* n*).*

Theorem 3 (Asymptotic vanishing property for finite-dimensional simplex)**.** *For* $1 \le k \le d-1$,

$$\mathbb{E}[M_{d,k}^{[s]}(n)] \to \begin{cases} 0, & \text{if } 1 \le k \le d-1; \\ \infty, & \text{if } k = d, \end{cases}$$

as $n \to \infty$.

Proof. For $k = d$, it is known (see [12]) that

$$\begin{aligned} \mathbb{E}[M_{d,d}^{[s]}(n)] &= d!n \int_D \left(1 - \left(1 - \textstyle\sum_{1\le i\le d} x_i\right)^d\right)^{n-1} \mathbf{dx} \\ &= n \sum_{0\le j<d} \binom{d-1}{j} (-1)^j \frac{\Gamma(n)\Gamma\left(\frac{j+1}{d}\right)}{\Gamma\left(n+\frac{j+1}{d}\right)} \\ &= \Gamma\left(\tfrac{1}{d}\right) n^{1-\frac{1}{d}} \left(1 + O\left(dn^{-\frac{1}{d}}\right)\right), \end{aligned}$$

where Γ denotes the Gamma function. Thus the expected number of skylines tends to infinity as n goes unbounded.

Consider now $1 \le k < d$. It suffices to examine the case $k = d-1$. For a point $\mathbf{x} \in S_d$ ($\mathbf{x} \ne \mathbf{0}$), let $\boldsymbol{\xi} := \frac{\mathbf{x}}{\|\mathbf{x}\|}$. Then $B_{d-1}^{[s]}(\boldsymbol{\xi}) \subset B_{d-1}^{[s]}(\mathbf{x})$. We now prove that

$$\left|B_{d-1}^{[s]}(\boldsymbol{\xi})\right| \ge \frac{1}{d!d^d} \qquad (\boldsymbol{\xi} \in S_d, \|\boldsymbol{\xi}\| = 1). \tag{8}$$

Since $\|\boldsymbol{\xi}\| = 1$, there is at least one coordinate $|\xi_j| \ge \frac{1}{d}$. Without loss of generality, assume $|\xi_d| \ge \frac{1}{d}$. Then $\sum_{1\le j<d} |\xi_j| \le \frac{d-1}{d}$. Let

$$T := \{\mathbf{y} \in S_d : y_j \le \xi_j \text{ for } 1 \le j \le d-1 \text{ and } y_d \le 0\}.$$

We have $T \subset B_{d-1}^{[s]}(\boldsymbol{\xi})$ and

$$|T| = |S_d||\xi_d| \ge \frac{1}{d!d^d},$$

since T is itself a simplex. Thus (8) holds and we have

$$\begin{aligned}\mathbb{E}[M^{[s]}_{d,d-1}(n)] &= nd! \int_{S_d} \left(1 - d! \left|B^{[s]}_{d-1}(\mathbf{x})\right|\right)^{n-1} \mathrm{d}\mathbf{x} \\ &= O\left(n\left(1 - d^{-d}\right)^n\right) \\ &\to 0,\end{aligned}$$

as $n \to \infty$. ■

We see in such a simplex model that the expected number of k-dominant tends to zero at an *exponential* rate (in n), in contrast to the *polynomial* rate in the hypercube model. Does the expected number of k-dominant skyline points always tend to zero? Here is a simple, artificial counterexample.

Example 1. Assume $d = 4, k = 3$. Let

$$A := \{(-t, -2t, 3t, 4t) : 1 \le t \le 2\}.$$

Then any two points in A are incomparable (none dominating the other) by the relation of k-dominance. Thus, the number of k-dominant skyline points is equal to n almost surely if $\mathbf{p}_1, \ldots, \mathbf{p}_n$ are uniformly and independently distributed in A.

5 A categorical model

The preceding negative results are based on assuming that the points are generated from some *continuous models*, which are often a good approximation to situations where the input can assume a sufficiently large range of different values. What if we assume instead that the inputs are sampled from some *discrete space*, which is also often encountered in practical applications? We show in this section that *the expected number of k-dominant skylines is always linear for $1 \le k \le d$*, in contrast to the asymptotic zero-infinity property we derived above.

Assume that n points $\mathscr{D} := \{\mathbf{p}_1, \ldots, \mathbf{p}_n\}$ are chosen uniformly and independently from the product space

$$\mathscr{P} := \bigotimes_{1 \le j \le d} S_j,$$

where

$$S_j = \{1, 2, \ldots, u_j\} \qquad (u_j \ge 2).$$

Let $M^{[c]}_{d,k}(n)$ denote the number of k-dominant skylines in $\mathscr{D}$. Unlike the continuous cases, the variation of the random variables $M^{[c]}_{d,k}(n)$ is easier to predict as the number of possible points in $\mathscr{P}$ is finite. Interestingly, the first-order asymptotic estimate for the expected value of $M^{[c]}_{d,k}(n)$ is independent of k for $1 \le k \le d$, where the case $k = d$ gives the expected skyline count.

Theorem 4 (Asymptotic linearity for finite-dimensional categorical model)**.** *The expected number of k-dominant skylines satisfies*

$$\frac{\mathbb{E}[M^{[c]}_{d,k}(n)]}{n} \to \frac{1}{u} \qquad (1 \le k \le d; d \ge 2), \tag{9}$$

as $n \to \infty$, *where*

$$u := \prod_{1 \le j \le d} u_j.$$

Now the problem is again the excessive number of skyline points. Such a discrete model exhibits another interesting phenomenon, not present for continuous model, namely, for fixed n, the expected number of k-dominant skyline points is not monotonically increasing as d grows.

Proof. Let $\mathbf{x} = (x_1, x_2, \ldots, x_d) \in \mathscr{P}$. Denote by $B_k^{[c]}(\mathbf{x})$ the set of points in $\mathscr{P}$ that k-dominate $\mathbf{x}$. Then

$$\begin{aligned} \mathbb{E}[M_{d,k}^{[c]}(n)] &= n\mathbb{P}(\mathbf{p}_1 \text{ is a } k\text{-dominant skyline point}) \\ &= \frac{n}{u} \sum_{\mathbf{x} \in \mathscr{P}} \left(1 - \frac{\left|B_k^{[c]}(\mathbf{x})\right|}{u}\right)^{n-1}. \end{aligned} \tag{10}$$

If $\mathbf{y} \in B_k^{[c]}(\mathbf{x})$, then $\mathbf{y}$ is better than or equal to $\mathbf{x}$ in all coordinates (at least one better) except for the coordinates, say $j_1, \ldots, j_\ell$ for $0 \le \ell \le d-k$. Thus

$$\left|B_d^{[c]}(\mathbf{x})\right| = \prod_{1 \le j \le d} x_j - 1,$$

and for $1 \le k < d$

$$\left|B_k^{[c]}(\mathbf{x})\right| = \sum_{0 \le \ell \le d-k} \sum_{1 \le j_1 < j_2 < \cdots < j_\ell \le d} \left(\frac{\prod_{1 \le i \le d} x_i}{\prod_{1 \le i \le \ell} x_{j_i}} - 1\right) \prod_{1 \le i \le \ell} (u_{j_i} - x_{j_i}). \tag{11}$$

Here the product

$$\frac{\prod_{1 \le i \le d} x_i}{\prod_{1 \le i \le \ell} x_{j_i}} = \prod_{i \ne j_r; r = 1, \ldots, \ell} x_i,$$

enumerates all possible locations in the $d - \ell$ ($\ge k$) coordinates that k-dominant skyline point can assume, and the factor "-1" removes the possibility that all $d - \ell$ coordinates are equal to the corresponding x_i. The last product in (11) describes all possible locations for the other ℓ coordinates.

Since there is a unique point $\mathbf{1} := (\overbrace{1, \ldots, 1}^{d})$ in $\mathscr{P}$ with $\left|B_k^{[c]}(\mathbf{1})\right| = 0$, all other terms in the sum on the right-hand side of (10) being exponentially small, we obtain (9). ■

In the special case when all $u_j = 2$ for $1 \le j \le d$, then

$$\left|B_k^{[c]}(\mathbf{x})\right| = \left(2^\ell - 1\right) \sum_{0 \le j \le d-k} \binom{d-\ell}{j},$$

where $\mathbf{x} \in \{1, 2\}^d$ and ℓ denotes the number of times "2" occurs in $\mathbf{x}$ (and "1" occurring $d - \ell$ times). The closed-form expression (10) simplifies

$$\mathbb{E}[M_{d,k}^{[c]}(n)] = \frac{n}{2^d} \sum_{0 \le \ell \le d} \binom{d}{\ell} \left(1 - \frac{2^\ell - 1}{2^d} \sum_{0 \le j \le d-k} \binom{d-\ell}{j}\right)^{n-1},$$

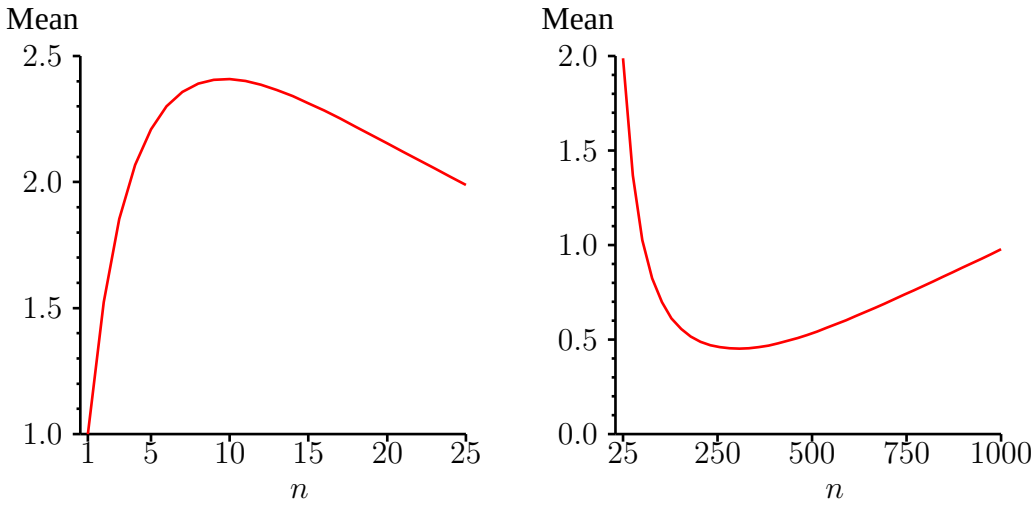

Figure 2: *A graphical rendering of* $\mathbb{E}[M_{d,k}^{[c]}(n)]$ *in the discrete space* $\{0,1\}^d$ *for* $d = 10$, $k = 9$ *and* $n = 1, \ldots, 25$ *(left) and* $n = 25, \ldots, 1000$ *(right).*

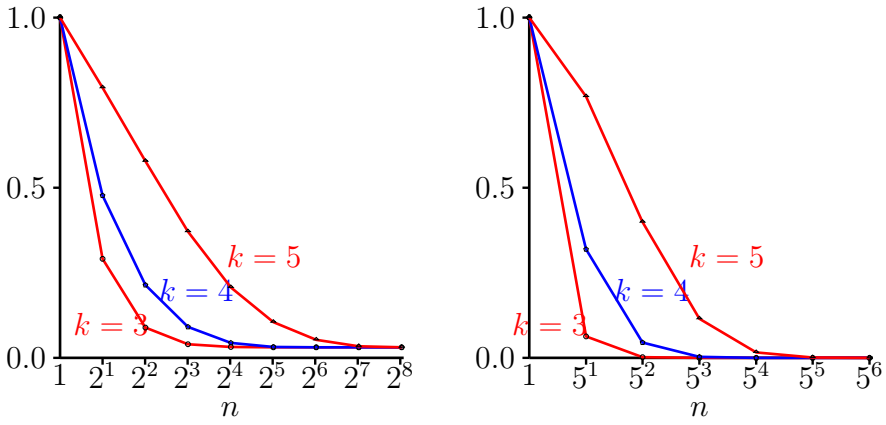

Figure 3: *Two plots of the ratio* $\mathbb{E}[M_{d,k}^{[c]}(n)]/n$ *when* $d = 5$, $k = 3, 4, 5$ *(here the case* $k = 5$ *corresponds to the skyline),* $u_i \equiv 2$ *(left) and* $u_i \equiv 5$ *(right). All curves in the left figure tend to the limit* $2^{-5} = 0.03125$ *while those in the right to* $5^{-5} = 0.00032$*, which is almost zero.*

from which it follows that

$$\frac{\mathbb{E}[M_{d,k}^{[c]}(n)]}{n} \to \frac{1}{2^d} \quad \text{as } n \to \infty.$$

Since the product space $\mathscr{P}$ is finite, we can indeed fully characterize the asymptotic distribution of $M_{d,k}^{[c]}(n)$.

Theorem 5 (Asymptotic binomial distribution for finite-dimensional categorical model)**.** *The distribution of $M_{d,k}^{[c]}(n)$ is asymptotically equivalent to a binomial distribution with parameters n and $1/u$.*

Proof. Let X_n denote the number of j's for which $\mathbf{p}_j = (1, \ldots, 1)$, $1 \le j \le n$. Then, obviously, X_n is binomially distributed with parameters n and $1/u$, namely,

$$\mathbb{P}(X_n = \ell) = \binom{n}{\ell} \frac{1}{u^\ell} \left(1 - \frac{1}{u}\right)^{n-\ell} \qquad (0 \le \ell \le n).$$

Now if one of the points $\mathbf{p}_j$ equals $(1, \ldots, 1)$, then $M_{d,k}^{[c]}(n) = X_n$. Thus

$$\mathbb{P}\left(M_{d,k}^{[c]}(n) \neq X_n\right) \le \mathbb{P}\left(\mathbf{p}_j \neq (1, \ldots, 1)\right) = \left(1 - \frac{1}{u}\right)^n \to 0,$$

and thus the distribution of $M_{d,k}^{[c]}(n)$ is asymptotic to the distribution of X_n. ■

In particular, we see that the variance of $M_{d,k}^{[c]}(n)$ is also asymptotically linear

$$\frac{\mathbb{V}[M_{d,k}^{[c]}(n)]}{n} \to \frac{1}{u}\left(1 - \frac{1}{u}\right) \qquad (1 \le k \le d).$$

The consideration can be easily extended to the case of non-uniform discrete distributions. More generally, assume that the data set is sampled from the set $\{\mathbf{a}_1, \ldots, \mathbf{a}_m\} \subset \mathscr{P}$ and each point is endowed with the probability $\mathbb{P}(\mathbf{a}_j)$. Let $p_k(\mathbf{a}_j)$ be the probability that $\mathbf{a}_j$ is k-dominated, that is, $p_k(\mathbf{a}_j)$ is equal to the sum of $\mathbb{P}(\mathbf{a}_i)$ such that $\mathbf{a}_i$ k-dominates $\mathbf{a}_j$. Then the expected number of k-dominant skyline points satisfies

$$\mathbb{E}[M_{d,k}^{[c]}(n)] = n \sum_{1 \le j \le m} \mathbb{P}(\mathbf{a}_j) \left(1 - p_k(\mathbf{a}_j)\right)^{n-1}.$$

Let

$$q_k := \sum_{\substack{p_k(\mathbf{a}_j) = 0 \\ 1 \le j \le m}} \mathbb{P}(\mathbf{a}_j)$$

be the probability of points in $\{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ that are not k-dominated. Then since the expected number of k-dominant is expressed as a finite sum, we have

$$\frac{\mathbb{E}[M_{d,k}^{[c]}(n)]}{n} \to q_k, \quad \text{as } n \to \infty.$$

Note that p_k may range from zero to one.

6 Uniform asymptotic estimates for $\mathbb{E}[M_{d,d-1}(n)]$

We derive in this section two uniform asymptotic estimates for $\mathbb{E}[M_{d,d-1}(n)]$ in two overlapping ranges. To state our results, we need to introduce the Lambert W-function (see [13]), which is implicitly defined by the equation

$$W(z)e^{W(z)} = z. \tag{12}$$

For our purpose, we take W to be the principal branch that is positive for positive z and satisfies the asymptotic approximation

$$W(x) = \log x - \log\log x + \frac{\log\log x}{\log x} + O\left(\frac{(\log\log x)^2}{(\log x)^2}\right), \tag{13}$$

for large x.

Our first asymptotic estimate covers d in the range

$$3 \le d \le \sqrt{\frac{2\log n}{W(2\log n) + K}},$$

where $K \to \infty$ with n, and the second the range

$$(\log n)^{1/3} \ll d \le 2\sqrt{\frac{\log n}{W(\log n) - C}},$$

for some constant $C > 0$. The upper bounds of the two ranges do not differ significantly but are sufficient for our purposes of proving the threshold phenomenon, which we discuss in the next section.

Very roughly, the expected number of $(d-1)$-dominant skylines is asymptotically negligible in the first range, and undergoes the phase transition from being almost zero to unbounded in the second.

Theorem 6 (Uniform estimate for large n and moderate d)**.** *If $d \ge 3$ and*

$$\frac{2\log n}{d^2} - W(2\log n) \to \infty, \tag{14}$$

then

$$\mathbb{E}[M_{d,d-1}(n)] = \frac{n^{-\frac{1}{d-1}}}{d-1}\,\Gamma\left(\frac{1}{d-1}\right)^d \left(1 + O\left(dn^{-\frac{1}{(d-1)(d-2)}}\right)\right), \tag{15}$$

uniformly in d for large n.

Note that if d is of the form

$$d = \left\lfloor \sqrt{\frac{2\log n}{W(2\log n) + 2v}} \right\rfloor,$$

then

$$dn^{-\frac{1}{(d-1)(d-2)}} = e^{-v}\left(1 + O\left(\frac{(1+|v|)W(2\log n)^{3/2}}{\sqrt{\log n}}\right)\right),$$

which becomes $o(1)$ if $v \to \infty$.

On the other hand, when $d = 2$, we have, by (2),

$$\mathbb{E}[M_{d,d-1}(n)] = n\int_0^1\int_0^1 (1-x-y+xy)^{n-1}\,\mathrm{d}x\mathrm{d}y = \frac{1}{n}.$$

Proof. We again begin with the integral representation (2), where $B_{d-1}(\mathbf{x})$ is given in (4).

By the elementary inequalities (see [6])

$$e^{-nt}(1-nt^2) \le (1-t)^n \le e^{-nt} \qquad (n \ge 1; t \in [0,1]),$$

we have

$$E_{n,d} - E'_{n,d} \le \mathbb{E}[M_{d,d-1}(n+1)] \le E_{n,d},$$

where

$$E_{n,d} := n\int_{[0,1]^d} e^{-n|B_{d-1}(\mathbf{x})|}\mathrm{d}\mathbf{x},$$
$$E'_{n,d} := n^2\int_{[0,1]^d} |B_{d-1}(\mathbf{x})|^2 e^{-n|B_{d-1}(\mathbf{x})|}\mathrm{d}\mathbf{x}.$$

We will see that $E'_{n,d}$ is asymptotically of smaller order than $E_{n,d}$. The intuition here is that most contribution to the integral comes from $\mathbf{x}$ for which $|B_{d-1}(\mathbf{x})|$ is small, implying that $(1-|B_{d-1}(\mathbf{x})|)^n$ is close to $e^{-n|B_{d-1}(\mathbf{x})|}$. Also replacing $n+1$ by n in the resulting asymptotic approximation gives rise only to smaller order errors. However, the uniform error bound represents the most delicate part of our proof.

We start with the asymptotic evaluation of $E_{n,d}$. By making the change of variables $x_j \mapsto \frac{y_j}{N}$, where $N := n^{\frac{1}{d-1}}$,

$$\begin{aligned} E_{n,d} &= N^{-1}\int_{[0,N]^d} e^{-y_1\cdots y_d\left(\frac{1}{y_1}+\cdots+\frac{1}{y_d}\right)+\frac{d-1}{N}y_1\cdots y_d}\mathrm{d}\mathbf{y} \\ &= N^{-1}\left(\phi_d(n) - f_d(n) + R_d(n)\right), \end{aligned} \tag{16}$$

where

$$\phi_d(n) := \int_{\mathbb{R}_+^d} e^{-y_1\cdots y_d\left(\frac{1}{y_1}+\cdots+\frac{1}{y_d}\right)}\mathrm{d}\mathbf{y},$$
$$f_d(n) := \left(\int_{\mathbb{R}_+^d} - \int_{[0,N]^d}\right) e^{-y_1\cdots y_d\left(\frac{1}{y_1}+\cdots+\frac{1}{y_d}\right)}\mathrm{d}\mathbf{y},$$
$$R_d(n) := \int_{[0,N]^d} e^{-y_1\cdots y_d\left(\frac{1}{y_1}+\cdots+\frac{1}{y_d}\right)}\left(e^{\frac{d-1}{N}y_1\cdots y_d}-1\right)\mathrm{d}\mathbf{y}.$$

We focus on the evaluation of the integral $\phi_d(n)$, leaving the lengthier estimation of the two error terms $f_d(n)$ and $R_d(n)$ to Appendix A.

We now carry out the change of variables $t_j := \prod_{\ell \neq j} y_\ell$ for $1 \leq j \leq d$, the Jacobian being

$$\frac{\partial(y_1, \ldots, y_d)}{\partial(t_1, \cdots, t_d)} := \begin{bmatrix} \frac{\partial y_1}{\partial t_1} & \cdots & \frac{\partial y_1}{\partial t_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_d}{\partial t_1} & \cdots & \frac{\partial y_d}{\partial t_d}, \end{bmatrix}$$

whose determinant is equal to $1/\det J$, where

$$J := \frac{\partial(t_1, \ldots, t_d)}{\partial(y_1, \cdots, y_d)}.$$

Note that the entries of J satisfy

$$J_{i,j} = \begin{cases} 0, & \text{if } i = j; \\ \dfrac{y_1 \cdots y_d}{y_i y_j}, & \text{if } i \neq j. \end{cases}$$

It follows that

$$\det J = (y_1 \cdots y_d)^{d-2} \det T,$$

where T is a $d \times d$ matrix with $T_{i,i} = 0$ and $T_{i,j} = 1$ for $i \neq j$. The determinant of T is seen to be $(-1)^{d-1}(d-1)$ by adding all rows of T to the first, by taking the factor $d-1$ out, and then by subtracting the first row from all other rows. Thus we have

$$\begin{aligned} \det J &= (-1)^{d-1}(d-1)(y_1 \cdots y_d)^{d-2} \\ &= (-1)^{d-1}(d-1)(t_1 \cdots t_d)^{\frac{d-2}{d-1}}. \end{aligned}$$

Thus, by the integral representation of the Gamma function

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \mathrm{d}t \qquad (x > 0),$$

we obtain

$$\begin{aligned} \phi_d(n) &= \frac{1}{d-1} \int_{\mathbb{R}_+^d} e^{-(t_1 + \cdots + t_d)} (t_1 \cdots t_d)^{-\frac{d-2}{d-1}} \mathrm{d}\mathbf{t} \\ &= \frac{1}{d-1} \left(\int_0^\infty e^{-u} u^{-\frac{d-2}{d-1}} \mathrm{d}u \right)^d \\ &= \frac{1}{d-1} \Gamma\left(\frac{1}{d-1} \right)^d. \end{aligned}$$

We will prove in Appendix A that

$$\begin{aligned} \frac{f_d(n)}{\phi_d(n)} &= O\left(d n^{-\frac{1}{(d-1)(d-2)}} \right), \\ \frac{R_d(n)}{\phi_d(n)} &= O\left(d 2^{-d} n^{-\frac{1}{d-1}} \right). \end{aligned} \tag{17}$$

In a similar manner, we have

$$E'_{n,d} = O\left(n^2 \int_{\mathbb{R}_+^d} \left(x_1 \cdots x_d \sum_{1\le j\le d} \frac{1}{x_j}\right)^2 e^{-nx_1\cdots x_d \sum_{1\le j\le d} \frac{1}{x_j}} \mathrm{d}\mathbf{x}\right)$$
$$= O\left(\frac{n^{-\frac{2}{d-1}}}{d-1} \int_{\mathbb{R}_+^d} (t_1 + \cdots + t_d)^2 \, e^{-(t_1+\cdots+t_d)} (t_1 \cdots t_d)^{-\frac{d-2}{d-1}} \mathrm{d}\mathbf{t}\right).$$

The last integral in a more general form can be evaluated as follows. Let $[z^n]f(z)$ denote the coefficient of z^n in the Taylor expansion of f.

$$\begin{aligned}
&\int_{\mathbb{R}_+^d} (t_1 + \cdots + t_d)^j \, e^{-(t_1+\cdots+t_d)} (t_1 \cdots t_d)^{-\frac{d-2}{d-1}} \mathrm{d}\mathbf{t} \\
&\quad = j![z^j] \int_{\mathbb{R}_+^d} e^{-(1-z)(t_1+\cdots+t_d)} (t_1 \cdots t_d)^{-\frac{d-2}{d-1}} \mathrm{d}\mathbf{t} \\
&\quad = j![z^j] \frac{\Gamma(\frac{1}{d-1})^d}{(1-z)^{\frac{d}{d-1}}} \\
&\quad = j!\Gamma\left(\frac{1}{d-1}\right)^d \binom{\frac{1}{d-1}+j}{j},
\end{aligned}$$

for $j \ge 0$. Thus

$$\frac{E'_{n,d}}{\phi_d(n)} = O\left(n^{-\frac{2}{d-1}}\right).$$

Collecting these estimates proves the theorem. ■

When d increases beyond the range (14), the error term $f_d(n)$ (see (16)) is no more negligible, and a more delicate analysis is needed.

Theorem 7 (Uniform asymptotic estimate in the critical range)**.** *If*

$$\frac{d}{(\log n)^{1/3}} \to \infty \quad \textit{and} \quad d \le 2\sqrt{\frac{\log n}{W\left(\frac{4\log n}{(e\log 2)^2}\right)}}, \tag{18}$$

then, with $\rho := \frac{d}{en^{1/d^2}}$,

$$\mathbb{E}[M_{d,d-1}(n)] = \frac{n^{-\frac{1}{d-1}}}{d-1}\Gamma\left(\frac{1}{d-1}\right)^d \left(\frac{1}{2-e^{-\rho}} + O\left(\frac{\rho(\rho+1)e^{-\rho}}{(2-e^{-\rho})^3}\left(\frac{1}{d} + \frac{\log n}{d^3}\right)\right)\right), \tag{19}$$

uniformly in d *for large* n.

The proof of this theorem is very long and is thus relegated in Appendix B. The crucial step is to prove an asymptotic estimate for $f_d(n)$ by an inductive argument by deriving first a recurrence of the form

$$f_d(n) = g_d(n) + \Phi[f_d](n) + \text{smaller order terms},$$

where

$$g_d(n) := \sum_{1\le j\le d-2} \binom{d}{j}(-1)^{j-1}(d-1-j)^{j-1}\Gamma\left(\tfrac{1}{d-1-j}\right)^{d-j} n^{\frac{1}{d-1}-\frac{1}{d-1-j}},$$

and Φ is an operator defined by

$$\Phi[f_d](n) := \sum_{1\le j\le d-2} \binom{d}{j}(-1)^j n^{\frac{1}{d-1}-\frac{1}{d-1-j}} \int_{(1,\infty)^j} (v_1\cdots v_j)^{-1-\frac{1}{d-1-j}}\, f_{d-j}(nv_1\cdots v_j)\mathbf{dv}.$$

Then (19) follows from iterating the operator and a careful analysis of the resulting sums.

Corollary 1. *If d is of the form*

$$d = \left\lfloor \sqrt{\frac{2\log n}{W(2\log n) - 2v - 2}} \right\rfloor,$$

then

$$\frac{\mathbb{E}[M_{d,d-1}(n)]}{\frac{n^{-\frac{1}{d-1}}}{d-1}\Gamma\left(\frac{1}{d-1}\right)^d} \sim \begin{cases} 1, & \text{if } v\to-\infty; \\ \frac{1}{2-e^{-e^v}}, & \text{if } v = O(1); \\ \frac{1}{2}, & \text{if } v\to\infty. \end{cases} \tag{20}$$

Proof. Observe that

$$\rho = \frac{d}{en^{1/d^2}} = e^v\left(1 + O\left(\frac{1+|v|}{W(2\log n)}\right)\right).$$

Thus (20) follows from this and (19). ■

Combining the ranges (14) and (18) of the two estimates (15) and (19), we see that

Corollary 2. *If*

$$3 \le d \le 2\sqrt{\frac{\log n}{W(4e^{-2}\log n)}},$$

then

$$\mathbb{E}[M_{d,d-1}(n)] \sim \frac{1}{2-e^{-\rho}}\cdot\frac{n^{-\frac{1}{d-1}}}{d-1}\Gamma\left(\frac{1}{d-1}\right)^d,$$

uniformly in d.

We conclude from these estimates that $\mathbb{E}[M_{d,d-1}(n)]$ is, modulo a constant term, very well approximated by $\frac{n^{-\frac{1}{d-1}}}{d-1}\Gamma\left(\frac{1}{d-1}\right)^d$.

Remark. A similar analysis as that for (15) leads to ($L_{d,k}(n,j)$ is defined in Section 3)

$$\mathbb{E}[L_{d,d-1}(n,j)] \sim c_{d,j}n^{-\frac{1}{d-1}}, \tag{21}$$

for each finite integer $j\ge 0$, where

$$\begin{aligned} c_{d,j} &:= \frac{1}{(d-1)j!}\int_{\mathbb{R}_+^d} (v_1+\cdots+v_d)^j e^{-(v_1+\cdots+v_d)}(v_1\cdots v_d)^{-\frac{d-2}{d-1}}\mathbf{dv} \\ &= \frac{1}{d-1}\Gamma\left(\frac{1}{d-1}\right)^d \binom{j+\frac{1}{d-1}}{j}, \end{aligned}$$

uniformly when $\frac{2\log n}{d^2} - W(2\log n)\to\infty$ and $j = o\left(n^{\frac{1-\varepsilon}{d}}\right)$, $\varepsilon\in(0,1)$. The consideration for larger d as for (19) is similar.

7 Threshold phenomenon for $\mathbb{E}[M_{d,d-1}(n)]$ when $d \to \infty$

With the asymptotic estimates (15) and (19) we derived in the previous section, we prove in this section a less expected threshold phenomenon for the expected number of $(d-1)$-dominant skylines $\mathbb{E}[M_{d,d-1}(n)]$ (in random samples from d-dimensional hypercube) when $d-1$ is near $\sqrt{\frac{2\log n}{W(2\log n)}}$.

Theorem 8 (Threshold phenomenon)**.** *Let*

$$d_0 = d_0(n) := \left\lfloor \sqrt{\frac{2\log n}{W(2\log n)}} \right\rfloor + 1, \tag{22}$$

where W denotes the Lambert-W function. Then the expected number of $(d-1)$-dominant skyline points satisfies

$$\lim_{n\to\infty} \mathbb{E}[M_{d,d-1}(n)] \to \begin{cases} 0, & \text{if } d < d_0; \\ \infty, & \text{if } d > d_0 + 1. \end{cases} \tag{23}$$

If $d = d_0$, then $\lim_{n\to\infty} \mathbb{E}[M_{d,d-1}(n)]$ does not exist and is oscillating between 0 and $\frac{e^{-\gamma}}{2-e^{-e^{-1}}}$

$$\mathbb{E}[M_{d,d-1}(n)] \sim \frac{e^{-\gamma}}{2-e^{-e^{-1}}}\, \varphi_0\left(\sqrt{\frac{2\log n}{W(2\log n)}}\right), \tag{24}$$

where $\varphi_0(x)$ is a bounded oscillating function of x defined by

$$\varphi_0(x) := e^{-\{x\}} x^{-2\{x\}}.$$

If $d = d_0 + 1$, then $\lim_{n\to\infty} \mathbb{E}[M_{d,d-1}(n)]$ does not exist and is oscillating between $\frac{e^{-\gamma}}{2-e^{-e^{-1}}}$ and $O\left(\frac{\log n}{\log\log n}\right)$

$$\mathbb{E}[M_{d,d-1}(n)] \sim \frac{e^{-\gamma}}{2-e^{-e^{-1}}}\, \varphi_1\left(\sqrt{\frac{2\log n}{W(2\log n)}}\right), \tag{25}$$

where $\varphi_1(x)$ is an oscillating function of x defined by

$$\varphi_1(x) := e^{1-\{x\}} x^{2-2\{x\}}.$$

Proof. By monotonicity, it suffices to examine the asymptotic behavior of $\mathbb{E}[M_{d,d-1}(n)]$ for d near d_0. Observe that if

$$d = d_0 + m = \sqrt{\frac{2\log n}{W_n}} - \tau_n + m + 1,$$

where m is an integer and τ denotes the fractional part of $\sqrt{\frac{2\log n}{W(2\log n)}}$, namely,

$$\tau_n := \left\{ \sqrt{\frac{2\log n}{W_n}} \right\} = \sqrt{\frac{2\log n}{W_n}} - \left\lfloor \sqrt{\frac{2\log n}{W_n}} \right\rfloor,$$

then

$$\rho = \frac{d}{en^{1/d^2}} = e^{-1}\left(1 + O\left(\frac{W_n^{\frac{3}{2}}|m+\tau_n|}{\sqrt{\log n}}\right)\right) \to e^{-1},$$

where, here and throughout the proof, $W_n := W(2\log n)$. Thus for bounded m

$$\frac{1}{2-e^{-\rho}} \to \frac{1}{2-e^{-e^{-1}}}.$$

On the other hand, by (19) and the asymptotic estimate $\Gamma(x) = x^{-1} - \gamma + O(x)$ as $x \to 0$, where γ denotes the Euler constant, we see that

$$\frac{n^{-\frac{1}{d-1}}}{d-1}\Gamma\left(\frac{1}{d-1}\right)^d = e^{-\gamma+m-\tau_n}\left(\frac{2\log n}{W_n}\right)^{m-\tau_n}\left(1 + O\left(\frac{W_n^{\frac{3}{2}}(m+\tau_n+1)^2}{\sqrt{\log n}}\right)\right)$$

$$\begin{cases} \to 0, & \text{if } m \le -1; \\ \sim e^{-\gamma}\varphi_0\left(\sqrt{\frac{2\log n}{W_n}}\right), & \text{if } m = 0; \\ \sim e^{-\gamma}\varphi_1\left(\sqrt{\frac{2\log n}{W_n}}\right), & \text{if } m = 1; \\ \to \infty, & \text{if } m \ge 2. \end{cases}$$

This proves (23), (24) and (25). It remains to consider more precisely the behavior of $\varphi_0(x)$ and $\varphi_1(x)$.

Obviously, by definition, $\varphi_0(x) \in (0,1]$ and $\varphi_1(x) \in [1,\infty)$ because $\{x\} \in [0,1)$ for $x \in \mathbb{R}_+$. If $\{x\} = 0$, then $\varphi_0(x) = 1$; more generally,

$$\varphi_0(x) \to \begin{cases} 1, & \text{if } \{x\}\log x = o(1); \\ 0, & \text{if } \{x\}\log x \to \infty. \end{cases}$$

On the other hand,

$$\varphi_1(x) \to \begin{cases} 1, & \text{if } (1-\{x\})\log x = o(1); \\ \infty, & \text{if } (1-\{x\})\log x \to \infty. \end{cases}$$

We now prove that

$$\tau_n = 0 \text{ if and only if } n = i^{i^2}\ (i \ge 2). \tag{26}$$

First, if $n = i^{i^2}$, then $2\log n = 2i^2\log i$ and the positive solution to the equation (see (12))

$$W_n e^{W_n} = 2i^2\log i,$$

is given by $W_n = 2\log i$, as can be easily checked. Thus

$$\sqrt{\frac{2\log n}{W_n}} = i \qquad (i \ge 2). \tag{27}$$

Conversely, if the relation (27) holds, then the positive solution to the equations

$$\frac{2\log n}{W_n} = i^2, \text{ and } W_n e^{W_n} = 2\log n,$$

is given by $n = i^{i^2}$. This proves (26).

It follows particularly, by (19), that

$$\lim_{i\to\infty} \mathbb{E}[M_{i,i-1}]\left(i^{i^2}\right) = \frac{e^{-\gamma}}{2 - e^{-e^{-1}}}.$$

This completes the proof of the theorem. ■

The function d_0 of n on the right-hand side of (22) grows extremely slowly. Let $a_i := i^{i^2}$ with $a_1 := 2$. Then $d = i + 1$ for $a_i \le n < a_{i+1}$, which is small for almost all practical sizes of n

$$d_0 = \begin{cases} 2, & \text{if } 2 \le n \le 15; \\ 3, & \text{if } 16 \le n \le 19682; \\ 4, & \text{if } 19683 \le n \le 42949\,67295; \\ 5, & \text{if } 42949\,67296 \le n \le 2.98\cdots \times 10^{17}; \\ 6, & \text{if } 2.98\cdots \times 10^{17} \le n \le 1.03\cdots \times 10^{28}. \end{cases}$$

This partly explains why the asymptotic vanishing property of $\mathbb{E}[M_{d,k}(n)]$ for large n and fixed d is "invisible" for moderate values of n.

Note that we did not replace the Lambert-W function in (22) by its asymptotic expansion (13) so as to make the expression more transparent, the reason being that no matter how many terms of the asymptotic expansion of W we use, the resulting expression is never $o(1)$. This is because all terms in the expansion are of orders in powers of $\log\log n$ and $\log\log\log n$, and they are all much smaller than $\log n$ in the numerator of the first term on the right-hand side of (22).

Extending the same analysis to other values of k becomes more difficult and messy except for $k = 1$ for which we have

$$\mathbb{E}[M_{d,1}(n)] = n \int_{[0,1]^d} (x_1 \cdots x_d)^{n-1} \mathbf{dx} = n^{1-d}.$$

Note that this always tends to zero no matter how large the value of d is.

On the other hand, for $1 \le k \le d-1$, we can derive the more precise estimate

$$\begin{aligned} \mathbb{E}[M_{d,k}(n)] &= O\left(n \int_{[0,1]^d} \exp\left(-n \sum_{1\le j_1<\cdots<j_k\le d} x_{j_1}\cdots x_{j_k}\right) \mathbf{dx}\right) \\ &= O\left(n^{1-\frac{d}{k}}\right). \end{aligned}$$

However, a more precise uniform asymptotic approximation (in n, d, and k) is less obvious and describing the corresponding threshold phenomena if any for other values of k also remains unclear. Intuitively, the asymptotic vanishing property is expected to hold as long as $k \ge d/2$ no matter d is finite or growing with n because the probability of a k-dominance for a random pair of points is larger than one half, meaning that it is less likely to find k-dominant skyline in such a case.

8 Expected number of dominant cycles

The asymptotic zero-infinity property can be viewed from another different angle by examining the *number of dominant cycles*.

Definition. We say that m points $\{\mathbf{p}_1, \ldots, \mathbf{p}_m\}$ form a k-dominant cycle (of length m) if $\mathbf{p}_i$ k-dominates $\mathbf{p}_{i+1}$ for $i = 1, \ldots, m-1$ and $\mathbf{p}_m$ k-dominates $\mathbf{p}_1$.

Roughly, the number of k-dominant cycles is inversely proportional to the number of k-dominant skylines. Note that by transitivity there is no cycle when $k = d$. Thus the number of cycles seems a better measure to clarify the structure of k-dominant skylines. However, the general configuration of the cycle structure is very complicated. We contend ourselves in this section with the consideration of cycles of length d when $k = d - 1$.

Lemma 1. *Let $C_{n,d}$ denote the number of $(d-1)$-dominant cycles of length d in a random sample of n points uniformly and independently chosen from $[0,1]^d$. Then the expected value of $C_{n,d}$ satisfies*

$$\mathbb{E}[C_{n,d}] = \binom{n}{d} \frac{d!^{2-d}}{d}. \tag{28}$$

Proof. Since the total number of cycles of length d is given by $\binom{n}{d}\frac{d!}{d}$, we see that

$$\mathbb{E}[C_{n,d}] = \binom{n}{d} \frac{d!}{d} \mathbb{P}\left(\{\mathbf{p}_1, \ldots, \mathbf{p}_d\} \text{ form a } (d-1)\text{-dominant cycle of length } d\right).$$

Assume that $\{\mathbf{p}_1, \ldots, \mathbf{p}_d\}$ form a $(d-1)$-dominant cycle of length d. Let

$$\mathbf{p}_i = (p_{i,1}, \ldots, p_{i,d}) \qquad (i = 1, \ldots, d).$$

Then for each coordinate j, there exists an ℓ such that

$$p_{1,j} > p_{2,j} > \cdots > p_{\ell,j}, \quad p_{\ell,j} < p_{\ell+1,j}, \quad p_{\ell+1,j} > \cdots > p_{d,j} > p_{1,j},$$

and the ℓ's are all distinct ($d!$ cases). Thus the probability of the event that $\{\mathbf{p}_1, \ldots, \mathbf{p}_d\}$ form a $(d-1)$-dominant cycle is given by

$$\frac{d!}{d!^d},$$

from which (28) follows. ∎

In particular, we see that

$$\mathbb{E}[C_{n,2}] = \frac{n(n-1)}{4},$$

which means that half of the pairs are cycles, rendering the 1-dominant skylines less likely to occur. The first few other $\mathbb{E}[C_{n,d}]$ are given by

$$\{\mathbb{E}[C_{n,d}]\}_{d\geq 3} = \left\{ \frac{n(n-1)(n-2)}{108}, \frac{n(n-1)(n-2)(n-3)}{55296}, \frac{n(n-1)(n-2)(n-3)(n-4)}{1036800000}, \right.$$
$$\left. \frac{n(n-1)(n-2)(n-3)(n-4)(n-5)}{1160950579200000}, \ldots \right\}.$$

We see that the denominator grows very fast and we expect another type of threshold phenomenon.

Let

$$d_1 := \left\lfloor \frac{\log n}{W(e^{-1}\log n)} + \tfrac{1}{2} \right\rfloor,$$

and τ_n denote the fractional part of $\frac{\log n}{W(e^{-1}\log n)} + \frac{1}{2}$. Also let

$$\upsilon(t) := \frac{1+\frac{1}{2}\log 2\pi}{W+1} + \frac{W}{(\log n)(W+1)}\left(t - \frac{12W^3 + (35-12\log 2\pi)W^2 + (34-24\log 2\pi)W + 23 + (\log 2\pi)^2}{24(W+1)^3}\right),$$

where $t \in \mathbb{R}$ and W represents $W(e^{-1}\log n)$. Note that W is of order $\log\log n$.

Theorem 9. *The expected number of $(d-1)$-dominant cycles of length d satisfies*

$$\lim_{n\to\infty} \mathbb{E}[C_{n,d}] \to \begin{cases} \infty, & \text{if } 2 \le d < d_1; \\ 0, & \text{if } d > d_1. \end{cases}$$

When $d = d_1$, we can write $\tau_n = \upsilon(t)$; then

$$\lim_{n\to\infty} \mathbb{E}[C_{n,d}] \begin{cases} \to 0 & \text{if } t \to -\infty; \\ \sim e^t, & \text{if } t = O(1); \\ \to \infty, & \text{if } t \to \infty. \end{cases} \tag{29}$$

Proof. Write

$$d = d_1 - m = \frac{\log n}{W(e^{-1}\log n)} + \tfrac{1}{2} - \upsilon,$$

where $\upsilon = m + \tau_n$. Then a straightforward calculation using (28) and Stirling's formula gives

$$\frac{1}{d}\log \mathbb{E}[C_{n,d}] = \upsilon\left(W(e^{-1}\log n) + 1\right) - 1 - \tfrac{1}{2}\log 2\pi + O\left(\frac{W(e^{-1}\log n)^2 + (\upsilon^2+1)W(e^{-1}\log n)}{\log n}\right).$$

Thus $\mathbb{E}[C_{n,d}] \to \infty$ if $m \ge 1$ and $\mathbb{E}[C_{n,d}] \to -\infty$ if $m \le -1$. When $m = 0$ ($\upsilon = \tau_n$), this asymptotic expansion is insufficient and we need more terms. If $\upsilon = \tau_n = \upsilon(t)$, then the same calculation as above gives

$$\mathbb{E}[C_{n,d}] = e^t\left(1 + O\left(\frac{W^2+1}{\log n}\right)\right).$$

This implies (29). ■

Let

$$a_i := \left\lfloor \left(\frac{i-\frac{1}{2}}{e}\right)^{i-\frac{1}{2}} \right\rfloor + 1 \qquad (i \ge 1).$$

Then

$$d_1 = d_1(n) = i \text{ if } a_i \le n < a_{i+1}.$$

The first few values of a_i are given as follows.

i	4	5	6	7	8	9	10	11	12
a_i	3	10	49	290	2022	16165	145405	1453435	15982276

9 A uniform lower bound for $\mathbb{E}[M_{d,k}(n)]$

The convergence rate in (1) is very slow if d is large and k is close to d. It is interesting to characterize the transition of $M_{d,k}(n)$ from zero to n as k increases under the condition that d and n are fixed. However, the exact characterization is not easy, so we derive instead a lower bound that provides a good approximation to the real transition.

Theorem 10 (Uniform lower bound in d, k and n)**.** *Define*

$$\beta_{d,k} := \sum_{0\le j\le d-k} \binom{d}{j} 2^{-d}.$$

Then, for $n \ge 1$ *and* $1 \le k \le d-1$*,*

$$\mathbb{E}[M_{d,k}(n)] \ge n I_n(\beta_{d,k}), \tag{30}$$

where

$$I_n(x) := x \int_x^1 t^{-2}\,(1-t)^{n-1}\,\mathrm{d}t.$$

Proof. Select two random points $\mathbf{x}, \mathbf{y}$ uniformly and independently in $[0,1]^d$. Obviously,

$$\mathbb{P}\left(\mathbf{x}\ k\text{-dominates }\mathbf{y}\right) = \beta_{d,k}.$$

On the other hand, by definition, $\mathbb{P}\left(\mathbf{x}\ k\text{-dominates }\mathbf{y}\right) = \int_{[0,1]^d} |B_k(\mathbf{x})|\,\mathrm{d}\mathbf{x}$. Thus

$$\int_{[0,1]^d} |B_k(\mathbf{x})|\,\mathrm{d}\mathbf{x} = \beta_{d,k}.$$

Let

$$F(t) = \left|\left\{\mathbf{x}\in[0,1]^d : |B_k(\mathbf{x})| \le t\right\}\right|,$$

be the distribution function of $|B_k(\mathbf{x})|$. By Markov inequality

$$t\,(1-F(t)) \le \int_{[0,1]^d} |B_k(\mathbf{x})|\,\mathrm{d}\mathbf{x} \qquad (t\in(0,1)).$$

Thus

$$F(t) \ge 1 - \frac{\int_{[0,1]^d} |B_k(\mathbf{x})|\,\mathrm{d}\mathbf{x}}{t} = 1 - \frac{\beta_{d,k}}{t}.$$

Define

$$G(t) := \max\left\{1 - \frac{\beta_{d,k}}{t}, 0\right\}.$$

Then $F(t) \ge G(t)$. Now

$$\mathbb{E}[M_{d,k}(n)] = n\int_{[0,1]^d} (1-|B_k(\mathbf{x})|)^{n-1}\,\mathrm{d}\mathbf{x} = n\int_0^1 (1-t)^{n-1}\,\frac{\mathrm{d}F(t)}{\mathrm{d}t}. \tag{31}$$

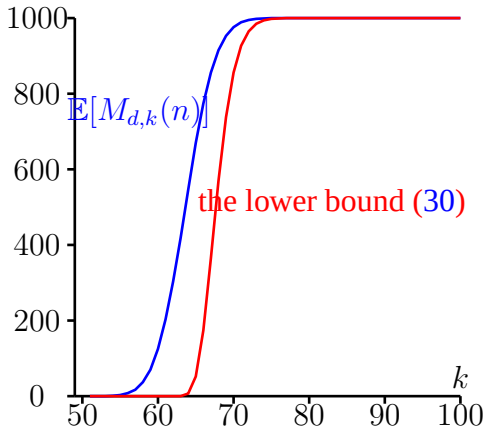

Figure 4: *Simulation result of* $\mathbb{E}[M_{d,k}(n)]$ *and the lower bound* (30) *for* $n = 1000, d = 100$ *and* k *from* 50 *to* 100.

Since the integral on the right-hand side of (31) becomes smaller if the distribution function $F(t)$ is replaced by $G(t)$, we have

$$\mathbb{E}[M_{d,k}(n)] \geq n \int_0^1 (1-t)^{n-1} \frac{\mathrm{d}G(t)}{\mathrm{d}t},$$

from which (30) follows. ■

A useful, convergent asymptotic expansion for $I_n(x)$, derived by successive integration by parts, is as follows.

$$\begin{aligned} I_n(x) &= \sum_{j\geq 0} \frac{(-1)^j (j+1)!}{n(n+1)\cdots(n+j)} x^{-j-1}(1-x)^{n+j} \\ &= \frac{(1-x)^n}{nx} - \frac{2(1-x)^{n+1}}{n(n+1)x^2} + \cdots, \end{aligned}$$

as long as $x \gg 1/n$. In particular, $I_n(x) \to 0$ in this range of x. If $xn \to c > 0$, then

$$I_n(x) \to c \int_c^\infty u^{-2} e^{-u} \mathrm{d}u,$$

the latter tending to 1 as c approaches zero.

We see that the transition of $I_n(x)$ from zero to one occurs at $x \asymp n^{-1}$ (meaning that x is of order proportional to n^{-1}). In terms of d and k, this arises when $d \to \infty$ and $\beta_{d,k} \asymp n^{-1}$. Now, by known estimate for binomial distribution (see [17] and the references cited there)

$$\beta_{d,k} \asymp (2\alpha - 1)^{-1} d^{-1/2} 2^{-d} \alpha^{-\alpha d} (1-\alpha)^{-(1-\alpha)d},$$

when $k \geq d/2 + K\sqrt{d}$, where $\alpha := k/d$ and $K > 1$ is a constant. We deduce from this that the transition of $I_n(\beta_{d,k})$ from zero to one occurs at $c \log n$ for some $c \in (0, 1)$. The exact location of this c matters less since I_n is simply a lower bound; see Figure 4.

10 Conclusions

While the notion of k-dominant skyline appeared as a natural means of solving the abundance of skyline, its use in diverse contexts has to be carefully considered, in view of the results we derived in this paper. We summarize our findings and highlight suggestions for possible practical uses.

The asymptotic results we derived in this paper are either of a vanishing type or of a blow-up nature; briefly, they are either zero or infinity when the sample size goes unbounded, making the selection of representative points more subtle. The expected number of k-dominant skyline points approaches zero under either of the following situations.

- Hypercube: both d and $k < d$ bounded;
- Simplex: both d and $k < d$ bounded;
- Hypercube: extending the k-dominant skyline to the dominance by a cluster of j points with both d and k bounded.

In all cases, zero appears as the limit when $n \to \infty$. However, for practical purposes, n is always finite, and thus the above limit results become less useful from a computational point of view. One needs asymptotic estimates that are uniform in d, k and n. But such results are often very difficult. The uniform asymptotic approximation (15) we obtained leads to several interesting consequences, including particularly the threshold phenomenon (23).

We conclude this paper by showing how the asymptotic results we derived above can be applied in more practical situations. Assume that our sample is of size, say $n = 10^4$ or $n = 10^5$, and the dimensionality d is in the range $\{4, 5, 6, 7, 8\}$ (smaller d may result in more biased inferences while larger d will yield too many skyline points). We also assume that our data set is sufficiently random and can be modeled by the hypercube model. If our aim is to choose a reasonably small number of candidates for further decision making, then how can our asymptotic estimates help?

First, for this range of n and d, the expected numbers of skyline points can be easily computed by the recurrence relation (see [5])

$$\mu_{n,d} = \frac{1}{d-1} \sum_{1 \le j \le d-1} H_n^{(d-j)} \mu_{n,j} \qquad (d \ge 2),$$

where $\mu_{n,d} := \mathbb{E}[M_{d,d}(n)]$, $H_n^{(a)} := \sum_{1 \le j \le n} j^{-a}$ are the harmonic numbers and $\mu_{n,1} := 1$, and are given approximately by

$$\{164.7, 426.3, 902.7, 1633.1, 2603\} \qquad (n = 10^4; d = 4, 5, 6, 7, 8),$$

and

$$\{304.9, 955.8, 2432.1, 5239.4, 9845\} \qquad (n = 10^5; d = 4, 5, 6, 7, 8),$$

which are often too many for further consideration. So we turn to $(d-1)$-dominant skyline and estimate their numbers by our asymptotic approximations. However, both Theorems 6 and 7 have poor error terms, and a better numerical approximation to $\mathbb{E}[M_{d,d-1}(n)$ for most moderately values of n and d is given by

$$\phi_d(n) - g_d(n) = \sum_{0 \le j \le d-2} \binom{d}{j} (-1)^j (d-1-j)^{j-1} \Gamma\left(\tfrac{1}{d-1-j}\right)^{d-j} n^{\frac{1}{d-1} - \frac{1}{d-1-j}}.$$

We thus obtain, for example, the following numerical values

$\mathbb{E}[M_{d,d-1}(10^4)] \approx$

d	4	5	6	7	8
$\phi_d(n) - g_d(n)$	0.61	5.06	24.85	88.90	243.96
Monte Carlo	0.57	4.82	23.98	83.89	226.65

and

$\mathbb{E}[M_{d,d-1}(10^5)] \approx$

d	4	5	6	7	8
$\phi_d(n) - g_d(n)$	0.31	3.69	24.94	115.31	404.7
Monte Carlo	0.29	3.61	24.38	111.79	386.08

From these tables, one can choose a suitable d according to the need of practical uses. Here we also see the characteristic property of the skylines, either very few or very many points.

Our Monte Carlo simulations are carried out by a three-phase algorithm (extending our two-phase maxima-finding one in [12]) for finding the k-dominant skylines. Briefly, the first two phases are modified from the algorithms presented in [12] and the last phase removes all cycles.

Acknowledgements

We thank Yuliy Baryshnikov for pointing out the references [3] and [23].

References

[1] O. Barndorff-Nielsen and M. Sobel (1966), On the distribution of the number of admissible points in a vector random sample, *Theor. Probability Appl.*, **11** 249–269.

[2] Y. Baryshnikov, On expected number of maximal points in polytopes. *2007 Conference on Analysis of Algorithms*, AofA 07, pp. 227–236, *Discrete Math. Theor. Comput. Sci. Proc.*, Nancy, 2007.

[3] Y. M. Baryshnikov and E. S. Orlova, Determination of maxima for arbitrary orders, *Avtomat. i Telemekh.* 1996, no. 1, 139–148; translation in *Automat. Remote Control*, **57** (1996), 112–119.

[4] Z.-D. Bai, C.-C. Chao, H.-K. Hwang, W.-Q. Liang, On the variance of the number of maxima in random vectors and its applications, *Ann. Appl. Probab.* **8** (1998), 886–895.

[5] Z.-D. Bai, L. Devroye, H.-K. Hwang and T.-H. Tsai, Maxima in hypercubes, *Random Structures Algorithms*, **27** (2005), 290–309.

[6] Z.-D. Bai, H.-K. Hwang, W.-Q. Liang, and T.-H. Tsai, Limit theorems for the number of maxima in random samples from planar regions, *Electron. J. Probab.*, **6** (2001) paper no. 3. 41 pp.

[7] S. Börzsönyi, D. Kossmann, and K. Stocker. The skyline operator, *Proceedings of the 17th International Conference on Data Engineering*, 421–430, 2001.

[8] C. Brando, M. Goncalves, and V. González, Evaluating top-k skyline queries over relational databases, *Lecture Notes in Computer Science*, **4653**, 254–263, 2007.

[9] C. Y. Chan, H. V. Jagadish, K.-L. Tan, A. K. H. Tung, and Z. Zhang, Finding k-dominant skylines in high dimensional space, *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, 503–514, 2006.

[10] C. Y. Chan, H. V. Jagadish, K.-L. Tan, A. K. H. Tung, and Z. Zhang, On high dimensional skylines, *Lecture Notes in Computer Science*, **3896**, 478–495, 2006.

[11] W.-M. Chen, H.-K. Hwang, and T.-H. Tsai, Efficient maxima-finding algorithms for random planar samples, *Discrete Math. Theor. Comput. Sci.*, **6**:1 (2003), 107–122.

[12] W.-M. Chen, H.-K. Hwang, and T.-H. Tsai, Maxima-finding algorithms for multidimensional samples: A two-phase approach, *Comput. Geom. Theor. Appl.*, **45**:1–2 (2012), 33–53.

[13] R. M. Corless, G. H. Gonnet, D. E. G. Hare and D. E. Knuth, On the Lambert W function, *Adv. Comput. Math.*, **5** (1996), 329–359.

[14] L. Devroye, *Lecture Notes on Bucket Algorithms*, Birkhäuser Boston, Inc., Boston, MA, 1986.

[15] L. Devroye, Records, the maximal layer, and uniform distributions in monotone sets. *Comput. Math. Appl.* **25** (1993), 19–31.

[16] P. Flajolet, G. Labelle, L. Laforest and B. Salvy, Hypergeometrics and the cost structure of quadtrees, *Random Structures Algorithms* **7** (1995), 117–144.

[17] H.-K. Hwang, Asymptotic estimates of elementary probability distributions, *Stud. Appl. Math.* **99** (1997), 393–417.

[18] H.-K. Hwang, Phase changes in random recursive structures and algorithms, in *Probability, Finance and Insurance*, pp. 82–97, World Sci. Publ., River Edge, NJ, 2004,

[19] H.-K. Hwang and T.-H. Tsai, Multivariate records based on dominance, *Electron. J. Probab.* 15 (2010), 1863–1892.

[20] I. F. Ilyas, G. Beskales and M. A. Soliman, A survey of top-k query processing techniques in relational database systems, *ACM Comput. Surveys*, **40** (2008), 1–58.

[21] V. Koltun and C. Papadimitriou, Approximately dominating representatives, *Theoret. Comput. Sci.*, **371**:3 (2007), 148–154.

[22] J. Lee, G.-W. You, and S.-W. Hwang, Personalized top-k skyline queries in high-dimensional space, *Inform. Sci.*, **34**:1 (2009), 45–61.

[23] E. S. Orlova, Asymptotics of the mean number of nondominated variants for binary relations, (translation) *Automat. Remote Control* **52** (1991), 1312–1316.

[24] D. Papadias, Y. Tao, G. Fu and B. Seeger, Progressive skyline computation in database systems, *ACM Trans. Database Systems*, **30** (2005), 41–82.

[25] T. Schreiber and J. E. Yukich, Variance asymptotics and central limit theorems for generalized growth processes with applications to convex hulls and maximal points. *Ann. Probab.* **36** (2008), 363–396.

[26] T. Xia, D. Zhang, and Y. Tao, On skylining with flexible dominance relation, *Proceedings of the 27th International Conference on Data Engineering*, 1397–1399, 2008.

[27] M. L. Yiu and N. Mamoulis, Multi-dimensional top-k dominating queries, *VLDB Journal*, **18**:3 (2009), 695–718.

[28] Z. Zhang, X. Guo, H. Lu, A. K. H. Tung, and N. Wang, Discovering strong skyline points in high dimensional spaces, in *ACM Fourteenth Conference on Information and Knowledge Management*, 247–248, 2005.

Appendix A. Error analysis: $d \le \sqrt{\frac{2\log n}{W(2\log n)+K}}$

Recall that $N := n^{\frac{1}{d-1}}$ and consider the integral

$$f_d(n) = \left(\int_{\mathbb{R}_+^d} - \int_{[0,N]^d} \right) e^{-y_1 \cdots y_d \left(\frac{1}{y_1}+\cdots+\frac{1}{y_d}\right)} \mathrm{d}\mathbf{y} = \sum_{1\le j\le d} \binom{d}{j} (-1)^{j-1} \phi_{d,j}(n),$$

where

$$\phi_{d,j}(n) := \int_{[0,N]^{d-j}\times(N,\infty)^j} e^{-y_1 \cdots y_d \left(\frac{1}{y_1}+\cdots+\frac{1}{y_d}\right)} \mathrm{d}\mathbf{y}. \tag{32}$$

So our $\phi_d(n) = \frac{1}{d-1}\Gamma\left(\frac{1}{d-1}\right)^d$ corresponds to $\phi_{d,0}(n)$; see (16).

Proposition 1. *Let $d \ge 3$ satisfies $\frac{2\log n}{d^2} - W(2\log n) \to \infty$. Then*

$$f_d(n) = O\left(\phi_d(n) d N^{-\frac{1}{d-2}}\right), \tag{33}$$

uniformly in d.

Proof. We first prove that uniformly for $1 \le j \le d$,

$$\phi_{d,j}(n) = O\left(\Gamma\left(\tfrac{1}{d-2}\right)^{d-1} N^{-\frac{j}{d-2}}\right). \tag{34}$$

Consider first the range $1 \le j \le d-2$. By extending the integration ranges and then carrying out the changes of variables $y_\ell \mapsto N v_{d-\ell+1}$ for $d-j+1 \le \ell \le d$, we obtain the bounds

$$\begin{aligned}\phi_{d,j}(n) &= N^j \int_{(1,\infty)^j}\int_{[0,N]^{d-j}} e^{-N^j v_1\cdots v_j y_1\cdots y_{d-j}\left(\frac{1}{y_1}+\cdots+\frac{1}{y_{d-j}}+\frac{1}{Nv_1}+\cdots+\frac{1}{Nv_j}\right)} \mathrm{d}\mathbf{y}\mathrm{d}\mathbf{v} \\ &\le N^j \int_{(1,\infty)^j}\int_{\mathbb{R}_+^{d-j}} e^{-N^j v_1\cdots v_j y_1\cdots y_{d-j}\left(\frac{1}{y_1}+\cdots+\frac{1}{y_{d-j}}\right)} \mathrm{d}\mathbf{y}\mathrm{d}\mathbf{v}.\end{aligned}$$

By the change of variables $y_j \mapsto \lambda^{-\frac{1}{d-1}} x_j$ for $1 \le j \le d$, we have, for $\lambda > 0$,

$$\int_{\mathbb{R}_+^d} e^{-\lambda y_1 \cdots y_d \left(\frac{1}{y_1}+\cdots+\frac{1}{y_d}\right)} \mathbf{dy} = \frac{\Gamma\left(\frac{1}{d-1}\right)^d}{d-1} \lambda^{-\frac{d}{d-1}} \qquad (d \ge 2).$$

It follows that

$$\begin{aligned}
\phi_{d,j}(n) &\le \frac{\Gamma\left(\frac{1}{d-1-j}\right)^{d-j}}{d-1-j} N^{-\frac{j}{d-1-j}} \int_{(1,\infty)^j} (v_1 \cdots v_j)^{-1-\frac{1}{d-1-j}} \mathbf{dv} \\
&= (d-1-j)^{j-1} \Gamma\left(\tfrac{1}{d-1-j}\right)^{d-j} N^{-\frac{j}{d-1-j}} \\
&= O\left(\Gamma\left(\tfrac{1}{d-2}\right)^{d-1} N^{-\frac{j}{d-2}}\right),
\end{aligned}$$

uniformly for $1 \le j \le d-2$. The remaining two cases $j = d-1, d$ are much smaller; we start with $\phi_{d,d}(n)$. By the same analysis used above, we have

$$\begin{aligned}
\phi_{d,d}(n) &= \int_{(N,\infty)^d} e^{-x_1 \cdots x_d \left(\frac{1}{x_1}+\cdots+\frac{1}{x_d}\right)} \mathbf{dx} \\
&\le \int_{(N,\infty)^d} e^{-x_1 \cdots x_d \left(\frac{1}{x_1}+\cdots+\frac{1}{x_{d-1}}\right)} \mathbf{dx} \\
&\le \int_{(N,\infty)^{d-1}} \frac{e^{-N x_1 \cdots x_{d-1} \left(\frac{1}{x_1}+\cdots+\frac{1}{x_{d-2}}\right)}}{x_1 \cdots x_{d-1} \left(\frac{1}{x_1}+\cdots+\frac{1}{x_{d-2}}\right)} \mathbf{dx}.
\end{aligned}$$

By the inequality

$$\int_N^\infty t^{-\alpha} e^{-\lambda t} \mathrm{d}t \le \lambda^{-1} N^{-\alpha} e^{-\lambda N} \qquad (\alpha \ge 0, \lambda > 0), \tag{35}$$

we obtain

$$\begin{aligned}
\phi_{d,d}(n) &\le N^{-2} \int_{(N,\infty)^{d-2}} \frac{e^{-N^2 x_1 \cdots x_{d-2} \left(\frac{1}{x_1}+\cdots+\frac{1}{x_{d-2}}\right)}}{x_1 \cdots x_{d-2} \left(\frac{1}{x_1}+\cdots+\frac{1}{x_{d-2}}\right)} \mathbf{dx} \\
&\le \cdots \\
&\le N^{-2-4-\cdots-2(d-3)} \int_{(N,\infty)^2} \frac{e^{-N^{d-2}(x_1+x_2)}}{(x_1+x_2)^{d-2}} \mathbf{dx} \\
&= N^{-(d-2)(d-3)} \int_{2N}^\infty \frac{e^{-N^{d-2} w}}{w^{d-2}} (w-2N) \mathrm{d}w \\
&\le 2^{3-d} N^{-d^2+3d-1} e^{-2N^{d-1}}.
\end{aligned}$$

Thus

$$\phi_{d,d}(n) = O\left(2^{-d} n^{-d+2+\frac{1}{d-1}} e^{-2n}\right). \tag{36}$$

Finally,

$$\begin{aligned}\phi_{d,d-1}(n) &\leq \int_{(N,\infty)^{d-1}} \frac{e^{-x_1\cdots x_{d-1}}}{x_1\cdots x_{d-1}\left(\frac{1}{x_1}+\cdots+\frac{1}{x_{d-1}}\right)}\mathrm{dx}\\ &\leq \frac{1}{d-1}\int_{(N,\infty)^{d-1}} \frac{e^{-x_1\cdots x_{d-1}}}{(x_1\cdots x_{d-1})^{1+\frac{1}{d-1}}}\mathrm{dx},\end{aligned}$$

by the inequality of arithmetic and geometric means

$$\frac{1}{d-1}\left(\frac{1}{x_1}+\cdots+\frac{1}{x_{d-1}}\right)\geq (x_1\cdots x_{d-1})^{\frac{1}{d-1}}.$$

Applying successively the inequality (35), we obtain

$$\begin{aligned}\phi_{d,d-1}(n) &\leq \frac{N^{-1-\frac{1}{d-1}}}{d-1}\int_{(N,\infty)^{d-2}} \frac{e^{-Nx_1\cdots x_{d-1}}}{(x_1\cdots x_{d-2})^{2+\frac{1}{d-1}}}\mathrm{dx}\\ &\leq \cdots\\ &\leq \frac{N^{-(d^2-2d+2)}}{d-1}e^{-N^{d-1}}.\end{aligned}$$

It follows that

$$\phi_{d,d-1}(n) = O\left(d^{-1}n^{-d+1-\frac{1}{d-1}}e^{-n}\right). \tag{37}$$

We see that both $\phi_{d,d}(n)$ and $\phi_{d,d}(n)$ are much smaller than the right-hand side of (34).

The remaining case is when $d=2$. Obviously,

$$\phi_{2,1}(n) < \int_0^\infty\int_N^\infty e^{-y_1-y_2}\mathrm{d}y_2\mathrm{d}y_1 = e^{-N}.$$

The upper bound (33) then follows from summing $\phi_{d,j}(n)$ for j from 1 to d using (34)

$$\begin{aligned}\sum_{1\leq j\leq d}\binom{d}{j}(-1)^{j-1}\phi_{d,j}(n) &= O\left(\Gamma\left(\tfrac{1}{d-2}\right)^{d-1}\sum_{j\geq 1}\frac{d^j}{j!}N^{-\frac{j}{d-2}}\right)\\ &= O\left(\Gamma\left(\tfrac{1}{d-2}\right)^{d-1}dN^{-\frac{1}{d-2}}\right),\end{aligned}$$

since $dN^{-\frac{1}{d-2}}\to 0$ for d in the range (14).

It remains to estimate $R_d(n)$, which can be proved to be bounded above by

$$\begin{aligned}R_d(n) &= O\left(\frac{d}{N}\int_{\mathbb{R}_+^d} y_1\cdots y_d e^{-y_1\cdots y_d\left(\frac{1}{y_1}+\cdots+\frac{1}{y_d}\right)}\mathrm{dy}\right)\\ &= O\left(\frac{1}{N}\Gamma\left(\frac{2}{d-1}\right)^d\right);\end{aligned}$$

this proves (17). ■

Appendix B. Proof of Theorem 7

We prove Theorem 7 in this Appendix. Our method of proof consists in a finer evaluation of the integrals $\phi_{d,j}(n)$, leading to a more precise asymptotic approximation to $f_d(n)$.

Proposition 2. *Uniformly for d in the range* (18)

$$f_d(n) \sim \frac{1-e^{-\rho}}{2-e^{-\rho}} \cdot \frac{1}{d-1}\Gamma\left(\frac{1}{d-1}\right)^d, \tag{38}$$

where $\rho := \frac{d}{en^{1/d^2}}$.

Proof. Consider again (32) and start with the changes of variables $y_\ell \mapsto Nv_{d-\ell+1}$ for $d-j+1 \le \ell \le d$,

$$\phi_{d,j}(n) = N^j \int_{(1,\infty)^j}\int_{[0,N]^{d-j}} e^{-\lambda_{N,j}(\mathbf{v})y_1\cdots y_{d-j}\left(\frac{1}{y_1}+\cdots+\frac{1}{y_{d-j}}+\frac{1}{Nv_1}+\cdots+\frac{1}{Nv_j}\right)}\mathbf{dydv},$$

where $\lambda_{N,j}(\mathbf{v}) := N^j v_1\cdots v_j$. Then we carry out the change of variables

$$y_\ell \mapsto \lambda_{N,j}(\mathbf{v})^{-\frac{1}{d-1-j}}x_\ell \qquad (1 \le \ell \le d-j),$$

and obtain

$$\phi_{d,j}(n) = \psi_{d,j}(n) + \omega_{d,j}(n),$$

where

$$\psi_{d,j}(n) = N^{-\frac{j}{d-1-j}}\int_{(1,\infty)^j}(v_1\cdots v_j)^{-1-\frac{1}{d-1-j}}\int_{[0,N_0]^{d-j}} e^{-x_1\cdots x_{d-j}\left(\frac{1}{x_1}+\cdots+\frac{1}{x_{d-j}}\right)}\mathbf{dxdv},$$

with

$$N_0 := N^{\frac{d-1}{d-1-j}}(v_1\cdots v_j)^{\frac{1}{d-1-j}} = (nv_1\cdots v_j)^{\frac{1}{d-1-j}},$$

and the error introduced is bounded above by

$$\begin{aligned}
\omega_{d,j}(n) &:= N^{-\frac{j}{d-1-j}}\int_{(1,\infty)^j}(v_1\cdots v_j)^{-1-\frac{1}{d-1-j}} \\
&\quad\times\int_{[0,N_0]^{d-j}} e^{-x_1\cdots x_{d-j}\left(\frac{1}{x_1}+\cdots+\frac{1}{x_{d-j}}\right)}\left(e^{-\frac{x_1\cdots x_{d-j}}{N_0}\left(\frac{1}{v_1}+\cdots+\frac{1}{v_j}\right)}-1\right)\mathbf{dxdv} \\
&= O\left(N^{-1-\frac{2j}{d-1-j}}\int_{(1,\infty)^j}(v_1\cdots v_j)^{-1-\frac{2}{d-1-j}}\left(\frac{1}{v_1}+\cdots+\frac{1}{v_j}\right)\right. \\
&\qquad\left.\times\int_{\mathbb{R}_+^{d-j}} e^{-x_1\cdots x_{d-j}\left(\frac{1}{x_1}+\cdots+\frac{1}{x_{d-j}}\right)}x_1\cdots x_{d-j}\mathbf{dxdv}\right) \\
&= O\left(j2^{-j}(d-1-j)^{j-2}\Gamma\left(\tfrac{2}{d-1-j}\right)^{d-j}N^{-1-\frac{2j}{d-1-j}}\right).
\end{aligned}$$

Thus the total contribution of $\omega_{d,j}(n)$ to $f_d(n)$ is bounded above by

$$\begin{aligned} h_d(n) &:= \sum_{1\le j\le d-2} \binom{d}{j} (-1)^{j-1} \omega_{d,j}(n) \\ &\le \sum_{1\le j\le d-2} \binom{d}{j} j 2^{-j} (d-1-j)^{j-2} \Gamma\left(\tfrac{2}{d-1-j}\right)^{d-j} n^{\frac{1}{d-1}-\frac{2}{d-1-j}}, \end{aligned} \tag{39}$$

which will be seen to be of a smaller order.

The recurrence relation Now

$$\begin{aligned} \psi_{d,j}(n) &= N^{-\frac{j}{d-1-j}} \int_{(1,\infty)^j} (v_1\cdots v_j)^{-1-\frac{1}{d-1-j}} \int_{\mathbb{R}_+^{d-j}} e^{-x_1\cdots x_{d-j}\left(\frac{1}{x_1}+\cdots+\frac{1}{x_{d-j}}\right)} \mathrm{dx}\mathrm{dv} \\ &\quad - N^{-\frac{j}{d-1-j}} \int_{(1,\infty)^j} (v_1\cdots v_j)^{-1-\frac{1}{d-1-j}} f_{d-j}(nv_1\cdots v_j)\mathrm{dv} \\ &= (d-1-j)^{j-1}\Gamma\left(\tfrac{1}{d-1-j}\right)^{d-j} N^{-\frac{j}{d-1-j}} \\ &\quad - N^{-\frac{j}{d-1-j}} \int_{(1,\infty)^j} (v_1\cdots v_j)^{-1-\frac{1}{d-1-j}} f_{d-j}(nv_1\cdots v_j)\mathrm{dv}. \end{aligned}$$

So we get the following recurrence relation.

Lemma 2. *The integrals $f_d(n)$ satisfy*

$$\begin{aligned} f_d(n) &= g_d(n) + h_d(n) + \eta_d(n) \\ &\quad + \sum_{1\le j\le d-2} \binom{d}{j} (-1)^j n^{\frac{1}{d-1}-\frac{1}{d-1-j}} \int_{(1,\infty)^j} (v_1\cdots v_j)^{-1-\frac{1}{d-1-j}} f_{d-j}(nv_1\cdots v_j)\mathrm{dv}, \end{aligned} \tag{40}$$

for $d \ge 3$, with the initial condition

$$f_2(n) = 2e^{-n} - e^{-2n},$$

where $h_d(n)$ is given in (39)*,*

$$g_d(n) := \sum_{1\le j\le d-2} \binom{d}{j} (-1)^{j-1} (d-1-j)^{j-1} \Gamma\left(\tfrac{1}{d-1-j}\right)^{d-j} n^{\frac{1}{d-1}-\frac{1}{d-1-j}},$$

and $\eta_d(n) := \phi_{d,d-1}(n) + \phi_{d,d}(n)$.

Note that, by (36) and (37),

$$\begin{aligned} \eta_d(n) &= O\left(d^{-1} n^{-d+1-\frac{1}{d-1}} e^{-n} + 2^{-d} n^{-d+2+\frac{1}{d-1}} e^{-2n}\right) \\ &= O\left(n^{-d+2} e^{-n}\right). \end{aligned}$$

Also, by the change of variables $t \mapsto v_1 \cdots v_j$, we have

$$f_d(n) = g_d(n) + h_d(n) + \eta_d(n) + \sum_{1\le j\le d-2} \binom{d}{j} \frac{(-1)^j n^{\frac{1}{d-1}-\frac{1}{d-1-j}}}{(j-1)!} \int_1^\infty t^{-1-\frac{1}{d-1-j}} (\log t)^{j-1} f_{d-j}(nt)\mathrm{d}t,$$

which is easier to use for symbolic computation softwares.

We then obtain, for example,

$$\begin{aligned} f_3(n) &= 3n^{-\frac{1}{2}} + O\left(n^{-\frac{3}{2}}\right), \\ f_4(n) &= 4\pi^{\frac{3}{2}} n^{-\frac{1}{6}} + O\left(n^{-\frac{2}{3}}\right), \\ f_5(n) &= \frac{80\pi^4}{9\Gamma\left(\frac{2}{3}\right)^4} n^{-\frac{1}{12}} - 60\pi^{\frac{3}{2}} n^{-\frac{1}{4}} + O\left(n^{-\frac{5}{12}}\right). \end{aligned}$$

But the expressions soon become too messy.

Asymptotic estimate for $g_d(n)$ We derive first a uniform asymptotic approximation to $g_d(n)$, which will be needed later. We focus on the case when d tends to infinity with n.

Lemma 3. *If d satisfies (18), then*

$$g_d(n) = \frac{1}{d-1}\Gamma\left(\frac{1}{d-1}\right)^d \left\{1 - e^{-\rho} + \rho e^{-\rho}\left(\frac{2\rho-1}{2d} + \frac{\rho-3}{d^3}\log n\right) + O\left(\frac{\rho e^{-\rho}(\rho^3+1)}{d^2}\left(1 + \frac{\log^2 n}{d^4}\right)\right)\right\}, \tag{41}$$

uniformly in d.

Proof. First, we have

$$\frac{\binom{d}{j}(-1)^{j-1}(d-1-j)^{j-1}\Gamma\left(\frac{1}{d-1-j}\right)^{d-j} n^{-\frac{j}{(d-1)(d-1-j)}}}{\frac{1}{d-1}\Gamma\left(\frac{1}{d-1}\right)^d} = \frac{d^j}{j!}(-1)^{j-1} n^{-\frac{j}{d^2}} \exp\left(-j - \frac{2j^2-j}{2d} - \frac{j(j+2)}{d^3}\log n + O\left(\frac{j^3}{d^2} + \frac{j^3}{d^4}\log n\right)\right),$$

uniformly for $j = o(d^{\frac{2}{3}})$. Summing over all j gives (41). Here the errors omitted are estimated by the inequalities

$$\begin{cases} \binom{d}{j} = O\left(\frac{d^j}{j!}\, e^{-\frac{j^2}{2d}}\right), \\ \Gamma\left(\frac{1}{x}\right) \le x, \qquad (x \ge 1) \\ (d-1-j)^{d-1} \le d^{d-1} e^{-j-\frac{j^2}{2d}}, \end{cases}$$

for $1 \le j \le d-2$, and we see that the contribution of terms in $g_d(n)$ with indices larger than, say $j_0 := \lfloor d^{\frac{3}{5}} \rfloor$ are bounded above by

$$\begin{aligned}
&\sum_{j \ge j_0} \binom{d}{j} (-1)^{j-1} (d-1-j)^{j-1} \Gamma\left(\tfrac{1}{d-1-j}\right)^{d-j} \; n^{-\frac{j}{(d-1)(d-1-j)}} \\
&\qquad = O\left(\frac{1}{d-1} \Gamma\left(\frac{1}{d-1}\right)^d \sum_{j \ge j_0} \frac{\rho^j}{j!} \right) \\
&\qquad = O\left(\frac{1}{d-1} \Gamma\left(\frac{1}{d-1}\right)^d \frac{\rho^{j_0}}{j_0!} \right).
\end{aligned}$$

Thus for d in the range (18)

$$\begin{aligned}
j_0 \log \rho - \log j_0! &= \tfrac{2}{5} d^{\frac{3}{5}} \log d - d^{-\frac{7}{5}} \log n + d^{\frac{3}{5}} + O(\log d) \\
&\le -\left(2^{-\frac{7}{5}} - \tfrac{1}{5} 2^{\frac{3}{5}}\right) (\log n)^{\frac{3}{10}} (\log\log n)^{\frac{7}{10}} (1+o(1)) \\
&\le -\tfrac{3}{40} (\log n)^{\frac{3}{10}} (\log\log n)^{\frac{7}{10}} (1+o(1)),
\end{aligned}$$

so that

$$\frac{\rho^{j_0}}{j_0!} = O\left(e^{-\frac{3}{40}(\log n)^{\frac{3}{10}} (\log\log n)^{\frac{7}{10}} (1+o(1))} \right),$$

and the sum of these terms is asymptotically negligible. The errors $\sum_{j \ge j_0} \frac{\rho^j}{j!}$ are estimated similarly. ■

Iteration of the Φ-operator To derive a similar estimate for $f_d(n)$, we define the operator

$$\Phi[f_d](n) := \sum_{1 \le j \le d-2} \binom{d}{j} (-1)^j n^{\frac{1}{d-1} - \frac{1}{d-1-j}} \int_{(1,\infty)^j} (v_1 \cdots v_j)^{-1-\frac{1}{d-1-j}} \, f_{d-j}(n v_1 \cdots v_j) \mathbf{dv}.$$

By iterating the recurrence (40), we obtain

$$f_d = g_d + h_d + \eta_d + \sum_{1 \le j \le d-2} \Phi^j[g_d + h_d + \eta_d],$$

where $\Phi^j[f_d] = \Phi[\Phi^{j-1}[f_d]]$ denotes the j-th iterate of the Φ-operator.

Surprisingly, despite of the complicated forms of the partial sums, each $\Phi^m[g_d]$ can be explicitly evaluated and differs from g_d only by a single term.

Lemma 4. *For any $m \ge 0$*

$$\Phi^m[g_d](n) = \sum_{m < \ell \le d-2} \binom{d}{\ell} (-1)^{\ell-1} (d-1-\ell)^{\ell-1} \Gamma\left(\tfrac{1}{d-1-\ell}\right)^{d-\ell} n^{\frac{1}{d-1} - \frac{1}{d-1-\ell}} \sigma_m(\ell), \qquad (42)$$

where $\sigma_m(\ell)$ is always positive and defined by

$$\sigma_m(\ell) := \sum_{\substack{j_1 + \cdots + j_{m+1} = \ell \\ j_1, \ldots, j_{m+1} \ge 1}} \binom{\ell}{j_1, \cdots, j_{m+1}}.$$

Note that

$$\begin{aligned}\sigma_m(\ell) &= \ell![z^\ell]\,(e^z-1)^{m+1}\\ &= \sum_{1\le r\le m+1}\binom{m+1}{r}(-1)^{m+1-r}r^\ell.\end{aligned}$$

Proof. By definition and by rearranging the terms

$$g_d(n) = \sum_{1\le j\le d-2}\binom{d}{j+1}(-1)^{d-j}j^{d-2-j}\Gamma\left(\tfrac{1}{j}\right)^{j+1}n^{\frac{1}{d-1}-\frac{1}{j}}.$$

Substituting this expression into the Φ-operator, we see that

$$\begin{aligned}\Phi[g_d](n) &= \sum_{1\le j\le d-2}\binom{d}{j}(-1)^j n^{\frac{1}{d-1}-\frac{1}{d-1-j}}\int_{(1,\infty)^j}(v_1\cdots v_j)^{-1-\frac{1}{d-1-j}}\,g_{d-j}(nv_1\cdots v_j)\mathbf{dv}\\ &= \sum_{1\le j\le d-2}\binom{d}{j}(-1)^j n^{\frac{1}{d-1}}\\ &\quad\times\sum_{1\le \ell\le d-j-2}\binom{d-j}{\ell+1}(-1)^{d-j-\ell}\ell^{d-2-j-\ell}\Gamma\left(\tfrac{1}{\ell}\right)^{\ell+1}n^{-\frac{1}{\ell}}\int_{(1,\infty)^j}(v_1\cdots v_j)^{-1-\frac{1}{\ell}}\,\mathbf{dv}.\end{aligned}$$

Then

$$\begin{aligned}\Phi[g_d](n) &= \sum_{1\le j\le d-2}\binom{d}{j}(-1)^j n^{\frac{1}{d-1}}\sum_{1\le \ell\le d-j-2}\binom{d-j}{\ell+1}(-1)^{d-j-\ell}\ell^{d-2-\ell}\Gamma\left(\tfrac{1}{\ell}\right)^{\ell+1}n^{-\frac{1}{\ell}}\\ &= \sum_{1\le \ell\le d-2}\binom{d}{\ell+1}(-1)^{d-\ell}\ell^{d-2-\ell}\Gamma\left(\tfrac{1}{\ell}\right)^{\ell+1}n^{\frac{1}{d-1}-\frac{1}{\ell}}\sum_{1\le j\le d-2-\ell}\binom{d-1-\ell}{j}\\ &= \sum_{1\le \ell\le d-2}\binom{d}{\ell+1}(-1)^{d-\ell}\ell^{d-2-\ell}\Gamma\left(\tfrac{1}{\ell}\right)^{\ell+1}n^{\frac{1}{d-1}-\frac{1}{\ell}}\left(2^{d-1-\ell}-2\right)\\ &= \sum_{1\le \ell\le d-2}\binom{d}{\ell}(-1)^{\ell-1}(d-1-\ell)^{\ell-1}\Gamma\left(\tfrac{1}{d-1-\ell}\right)^{d-\ell}n^{\frac{1}{d-1}-\frac{1}{d-1-\ell}}\left(2^\ell-2\right).\end{aligned}$$

By repeating the same analysis and induction, we prove (42). ■

Corollary 3. *If d satisfies (18), then*

$$\Phi^m[g_d](n)\sim(-1)^m\tfrac{1}{d-1}\Gamma\left(\tfrac{1}{d-1}\right)^d\left(1-e^{-\rho}\right)^{m+1}\qquad(m=0,1,\dots).$$

Summing over all $0\le m\le d-2$, we deduce (38) and it remains only the error estimates.

Error analysis The consideration of $\Phi^m[h_d]$ is similar and we obtain

$$\Phi^m[h_d](n)\le\sum_{m<\ell\le d-2}\binom{d}{\ell}2^{-\ell}(d-1-\ell)^{\ell-2}\Gamma\left(\tfrac{2}{d-1-\ell}\right)^{d-\ell}n^{\frac{1}{d-1}-\frac{2}{d-1-\ell}}\sigma'_m(\ell)$$

where

$$\begin{aligned}\sigma'_m(\ell) &:= \sum_{\substack{j_1+\cdots+j_{m+1}=\ell\\ j_1,\ldots,j_{m+1}\geq 1}} \binom{\ell}{j_1,\cdots,j_{m+1}} j_{m+1}\\ &= \ell![z^\ell]ze^z\left(e^z-1\right)^m\\ &= \ell\sum_{0\leq r\leq m}\binom{m}{r}(-1)^{m-r}(r+1)^\ell \qquad (m\geq 0).\end{aligned}$$

Thus, with

$$\rho_0 := \frac{d}{en^{2/d^2}}$$

which is always $\leq \log 2$ when d satisfies (18), we then have

$$\begin{aligned}\frac{\Phi^m[h_d](n)}{\frac{1}{d(d-1)2^d}\Gamma\left(\frac{1}{d-1}\right)^d n^{-\frac{1}{d-1}}} &= O\left(\sum_{0\leq r\leq m}\binom{m}{r}(-1)^{m-r}\sum_{\ell\geq 0}\frac{\rho_0^\ell}{(\ell-1)!}(r+1)^\ell\right)\\ &= O\left(\rho_0 e^{\rho_0}\sum_{0\leq r\leq m}\binom{m}{r}(-1)^{m-r}(r+1)e^{r\rho_0}\right)\\ &= O\left(\rho_0 e^{\rho_0}\left((e^{\rho_0}-1)^{m-1}\left((m+1)e^{\rho_0}-1\right)\right)\right).\end{aligned}$$

Now

$$\sum_{0\leq m\leq d-2}\left((x-1)^{m-1}\left((m+1)x-1\right)\right) = O(d^2)$$

whenever $0\leq x\leq 2$. It follows that

$$\sum_{0\leq m\leq d-2}\Phi^m[h_d] = O\left(2^{-d}d^{-2}\Gamma\left(\frac{1}{d-1}\right)^d n^{-\frac{1}{d-1}}\rho_0 e^{\rho_0}\right),$$

which holds uniformly as long as $e^{\rho_0}\leq 2$. This is how the upper limit of d in (18) arises.

In such a case,

$$\sum_{0\leq m\leq d-2}\Phi^m[h_d] = O\left(2^{-d}d^{-1}\Gamma\left(\frac{1}{d-1}\right)^d n^{-\frac{1}{d-1}-\frac{2}{d^2}}\right).$$

We consider now $\Phi^j[\eta_d]$. Note that an exponentially small term remains exponentially small under the Φ-operator because

$$\int_{(1,\infty)^j}(v_1\cdots v_j)^{-1-\alpha}e^{-nv_1\cdots v_j}\mathbf{dv}\sim n^{-j}e^{-n}.$$

So all terms of the forms $\Phi^m[\eta_d]$ are asymptotically negligible. And we then deduce (38). ■

More calculations give

$$\begin{aligned}\frac{f_d(n)}{\frac{1}{d-1}\Gamma\left(\frac{1}{d-1}\right)^d} &= \frac{1-e^{-\rho}}{2-e^{-\rho}}+\frac{\rho e^{-\rho}}{(2-e^{-\rho})^3}\left(\frac{2\rho-1+(\rho+\frac{1}{2})e^{-\rho}}{d}\right.\\ &\left.+\frac{2(\rho-3)+(\rho+3)\,e^{-\rho}}{d^3}\log n\right)+O\left(\frac{\rho e^{-\rho}}{d^2}(\rho^3+1)\left(1+\frac{\log^2 n}{d^4}\right)\right).\end{aligned}$$

Note that the range (14) arises because we had to drop factors of the form $(-1)^j$ in estimating the sum of $h_d(n)$. With a more careful analysis along the same inductive line, we can extend the range of uniformity of (38).