Unimodality, Independence Lead to NP-Hardness of Interval Probability Problems

Daniel J. Berleant[1], Olga Kosheleva[2],
Vladik Kreinovich[2], and Hung T. Nguyen[3]

[1]Department of Electrical and Computer Engineering
Iowa State University, Ames, IA 50011, USA
berleant@iastate.edu
[2]NASA Pan-American Center for
Earth and Environmental Studies (PACES)
University of Texas, El Paso, TX 79968, USA
olgak@utep.edu, vladik@utep.edu
[3]Department of Mathematical Sciences
New Mexico State University, Las Cruces, NM 88003, USA
hunguyen@nmsu.edu

Abstract

In many real-life situations, we only have partial information about probabilities. This information is usually described by bounds on moments, on probabilities of certain events, etc. – i.e., by characteristics $c(p)$ which are linear in terms of the unknown probabilities p_j. If we know interval bounds on some such characteristics $\underline{a}_i \le c_i(p) \le \overline{a}_i$, and we are interested in a characteristic $c(p)$, then we can find the bounds on $c(p)$ by solving a linear programming problem.

In some situations, we also have additional conditions on the probability distribution – e.g., we may know that the two variables x_1 and x_2 are independent, or that the joint distribution of x_1 and x_2 is unimodal. We show that adding each of these conditions makes the corresponding interval probability problem NP-hard.

1 Introduction

Interval probability problems can be often reduced to linear programming (LP). In many real-life situations, in addition to the *intervals* $[\underline{x}_i, \overline{x}_i]$ of possible values of the unknowns $x_1, \ldots, x_n$, we also have partial information about the *probabilities* of different values within these intervals.

This information is usually given in terms of bounds on the standard characteristics $c(p)$ of the corresponding probability distribution p, such as the k-th moment $M_k \stackrel{\rm def}{=} \int x^k \cdot \rho(x)\,dx$ (where $\rho(x)$ is the probability density), the values of the cumulative distribution function (cdf) $F(t) \stackrel{\rm def}{=} {\rm Prob}(x \le t) = \int_{-\infty}^{t} \rho(x)\,dx$ of some of the variables, etc. Most of these characteristics are linear in terms of $\rho(x)$ – and many other characteristics like central moments are combinations of linear characteristics: e.g., variance V can be expressed as $V = M_2 - M_1^2$.

A typical practical problem is when we know the ranges of some of these characteristics $\underline{a}_i \le c_i(p) \le \overline{a}_i$, and we want to find the range of possible values of some other characteristic $c(p)$. For example, we know the bounds on the marginal cdfs for the variables x_1 and x_2, and we want to find the range of values of the cdf for $x_1 + x_2$.

In such problems, the range of possible values of $c(p)$ is an interval $[\underline{a}, \overline{a}]$. To find $\underline{a}$ (correspondingly, $\overline{a}$), we must minimize (correspondingly, maximize) the linear objective function $c(p)$ under linear constraints — i.e., solve a linear programming (LP) problem; see, e.g., [17, 18, 19, 26].

Other simple examples of linear conditions include bounds on the values of the density function $\rho(x)$; see, e.g., [16].

Interval probability problems also naturally arise in the analysis of not fully identified probabilistic models – e.g., when we need to estimate population-related parameters (such as variance or covariance) and some data points are missing. These models often results in problems that are, essentially, interval computation problems. Economists recently became interested in the analysis of such models. For an overview of these models and their economic applications see, e.g., [11, 20] and references therein.

Similar problems also naturally appear in robust (= multi-prior) Bayesian analysis. If a set of priors is convex, this approach essentially boils down to interval probabilities as well.

In both economic and Bayesian applications, some of the resulting problems are non-linear, but many problems are linear and can be thus solved by LP techniques.

Comment. In some practically important cases, the traditional formulation of the problem is not explicitly related to LP, but it is possible to reformulate these problems in LP terms. For example, when we select a new strategy for a company (e.g., for an electric company), one of the reasonable criteria is that the expected monetary gain should be not smaller than the expected gain for a previously known strategy. In many case, for each strategy, we can estimate the probability of different production values – e.g., the probability $F(t) = {\rm Prob}(x \le t)$ that we will produce the amount $\le t$. However, the utility $u(t)$ corresponding to producing t depends on the future prices and is not well known; therefore, we cannot predict the exact value of the expected utility $\int u(x) \cdot \rho(x)\,dx$. One way to handle this situation is require that for *every* monotonic utility function $u(t)$, the expected utility under the new strategy – with probability density function (pdf) $\rho(x)$ and cdf $F(x)$ – is larger than or equal to the expected utility under the

old strategy – with pdf $\rho_0(x)$ and cdf $F_0(x)$: $\int u(x) \cdot \rho(x)\,dx \ge \int u(x) \cdot \rho_0(x)\,dx$. This condition is called *first order stochastic dominance*. It is known that this condition is equivalent to the condition that $F(x) \le F_0(x)$ for all x.

Indeed, the condition is equivalent to

$$\int_0^t u(x) \cdot (\rho(x) - \rho_0(x))\,dx \ge 0.$$

Integrating by part, we conclude that

$$-\int_0^t u'(x) \cdot (F(x) - F_0(x))\,dx \ge 0;$$

since $u(x)$ is non-decreasing, the derivative $u'(x)$ can be an arbitrary non-negative function; so, the above condition is indeed equivalent to $F(x) \le F_0(x)$ for all x.

Each of these inequalities is linear in terms of $\rho(x)$ – so, optimizing a linear objective function under the constraints $F(x) \ge F_0(x)$ is also a LP problem.

This requirement may be too restrictive; in practice, preferences have the property of risk aversion: it is better to gain a value x with probability 1 than to have either 0 or $2x$ with probability 1/2. In mathematical terms, this condition means that the corresponding utility function $u(x)$ is concave. It is therefore reasonable to require that for all such *risk aversion* utility functions $u(x)$, the expected utility under the new strategy is larger than or equal to the expected utility under the old strategy. This condition is called *second order stochastic dominance* (see, e.g., [7, 8, 22, 23]), and it known to be equivalent to the condition that $\int_0^t F(x)\,dx \le \int_0^t F_0(x)\,dx$.

Indeed, the condition is equivalent to

$$\int_0^t u(x) \cdot (\rho(x) - \rho_0(x))\,dx \ge 0$$

for every concave function $u(x)$. Integrating by part twice, we conclude that

$$\int_0^t u''(x) \cdot \left(\int_0^x F(z)\,dz - \int_0^x F_0(z)\,dz \right) dx \ge 0.$$

Since $u(x)$ is concave, the second derivative $u''(x)$ can be an arbitrary non-positive function; so, the above condition is indeed equivalent to $\int_0^t F(x)\,dx \le \int_0^t F_0(x)\,dx$ for all t.

The cdf $F(x)$ is a linear combination of the values $\rho(x)$; thus, its integral $\int F(x)\,dx$ is also linear in $\rho(x)$, and hence the above condition is still linear in terms of the values $\rho(x)$. Thus, we again have a LP problem; for details, see [2].

Most of the corresponding LP problems can be efficiently solved. Theoretically, some of these LP problems have infinitely many variables $\rho(x)$,

but in practice, we can discretize each coordinate and thus, get a LP problem with finitely many variables.

There are known efficient algorithms and software for solving LP problems with finitely many variables. These algorithms require polynomial time ($\leq n^k$) to solve problems with $\leq n$ unknowns and $\leq n$ constraints; these algorithms are actively used in imprecise probabilities; see, e.g., [1, 4, 5, 6].

For example, for the case of two variables x_1 and x_2, we may know the probabilities $p_i = p(x_1 \in [i, i+1])$ and $q_j = p(x_2 \in [j, j+1])$ for finitely many intervals $[i, i+1]$. Then, to find the range of possible values of, e.g.,

$$\text{Prob}(x_1 + x_2 \leq k),$$

we can consider the following linear programming problem: the unknowns are

$$p_{i,j} \stackrel{\text{def}}{=} p(x_1 \in [i, i+1] \,\&\, x_2 \in [j, j+1]),$$

the constraints are $p_{i,j} \geq 0$, $p_{i,1} + p_{i,2} + \ldots = p_i$, $p_{1,j} + p_{2,j} + \ldots = q_j$, and the objective function is $\sum\limits_{i,j:i+j\leq k} p_{i,j}$.

Comment. The only LP problems for which there may not be an efficient solution are problems involving a large amount of variables v. If we discretize each variable into n intervals, then overall, we need n^v unknowns $p_{i_1,i_2,\ldots,i_v}$ ($1 \leq i_1 \leq n$, $1 \leq i_2 \leq n$, $\ldots$, $1 \leq i_v \leq n$) to describe all possible probability distributions. When v grows, the number of unknowns grows exponentially with v and thus, for large v, becomes unrealistically large.

It is known (see, e.g., [13]) that this exponential increase in complexity is inherent to the problem: e.g., for v random variables $x_1, \ldots, x_v$ with known marginal distributions, the problem of finding the exact bounds on the cdf for the sum $x_1 + \ldots + x_v$ is NP-hard.

Beyond LP. There are important practical problems which lie outside LP. One example is problems involving *independence*, when constraints are linear in $p(x, y) = p(x) \cdot p(y)$ and thus, bilinear in $p(x)$ and $p(y)$. In this paper, we show that the corresponding range estimation problem is NP-hard.

Another example of a condition which cannot be directly described in terms of LP is the condition of *unimodality*. This notion is very practically useful, because in many practical situations, we do not know the exact probability distribution, but we do know that this (unknown) distribution is unimodal. For example, in statistical analysis, a unimodal distribution means that we have the data corresponding to a single cluster, while a multi-modal distribution would mean that we have a mixture of data corresponding to different clusters – a mixture that needs to be separated before we start statistical analysis of this cluster (see, e.g., [15]); other applications of uni- and multi-modality – and corresponding interval-related problems and algorithms – are presented, e.g., in [24, 25]. Often, when we take unimodality into account, we can drastically improve the resulting estimates; see, e.g., [9, 21].

For a one-variable distribution with probabilities $p_1, \ldots, p_n$, unimodality means that there exists a value m ("mode") such that p_i increase (non-strictly = weakly) until m and then decreases after m:

$$p_1 \le p_2 \le \ldots \le p_{m-1} \le p_m \ge p_{m+1} \ge \ldots \ge p_{n-1} \ge p_n.$$

When the location of the mode is known, we get several linear inequalities, so we can still use efficient techniques such as LP; see, e.g., [9, 27].

For a 1-D case, if we do not know the location of the mode, we can try all n possible locations and solve n corresponding LP problems. Since each LP problem requires a polynomial time to run, running n such problems still requires a polynomial time.

In this paper, we show that if we assume unimodality in the 2-D case, then the range estimation problem also becomes NP-hard – and therefore, that, unless P=NP, no algorithm can solve all the instances of this problem in polynomial time.

Comments.

- This paper builds on our previous work: our motivations are very similar to the ones described in our conference paper [2], and two of the results – Theorem 3 and Theorem 6 – first appeared in our conference paper [3]. All other results are new.
- Other possible restrictions on probability may involve bounds on the *entropy* of the corresponding probability distributions; such problems are also, in general, NP-hard [12].

2 Adding Unimodality Makes Interval Probability Problems NP-Hard

Before we formulate our first result, let us recall the definition of a unimodal distribution.

For a continuous distribution (1-D or 2-D) with density $\rho(x)$, a *mode* is usually defined as a value m at which the probability density function $\rho(x)$ attains a (non-zero) local maximum, i.e., at which $\rho(m) \ge \rho(y)$ for all y from some neighborhood of the point m. A distribution with a unique mode is called *unimodal*, a distribution with exactly two modes is called *bimodal*, etc.

For most distributions, this definition captures an intuitive meaning of unimodality. For example, in the 1-D case, the graph of a function $\rho(x)$ corresponding to a unimodal distribution has exactly one local maximum m, i.e., the function $\rho(x)$ increases for $x \le m$, and then decreases. The graph of the function $\rho(x)$ for a bimodal distribution has two local maxima m_1 and m_2, so the function $\rho(x)$ grows for $x \le m_1$, then it start decreasing, then increases again until it reaches m_2, and then finally decreases to 0. Intuitively, a multi-modal distribution is a distribution for which the density function $\rho(x)$ oscillates a lot.

The above definition, however, may be somewhat counter-intuitive if we try to apply it to, e.g., a uniform distribution on an interval $[0,1]$. In this distribution, $\rho(x) = 1$ for all $x \in [0,1]$, and $\rho(x) = 0$ for all other x. From the purely mathematical viewpoint, all the values $x \in [0,1]$ are local maxima, and thus, the uniform distribution is classified as multi-modal. However, in this example, the density function $\rho(x)$ stays constant and does not oscillate at all. It is therefore reasonable to consider a modified definition of unimodality, in which a local maximum is defined as either a connected set S of points (which may be a 1-point set) such that:

- $\rho(x)$ is constant on this set S, and
- $\rho(y)$ is smaller than this constant for all values y from some neighborhood of the set S.

Under this new definition, the uniform distribution is unimodal, with the only local maximum set $S = [0,1]$. To distinguish between these two definitions, we will call unimodality according to the traditional definition *strict unimodality*, and the new definition simply *unimodality*.

Both definitions can be naturally extended to a discrete case. In the 1-D case, each value i has two natural neighbors: $i-1$ and $i+1$; we can also consider the point i to be its own neighbor. In the 2-D case, points (i,j) and (i',j') are neighbors if i is a neighbor of i' and j is a neighbor of j'.

We will show that under discrete analogues of both definitions, the problem becomes NP-hard.

Definition 1 *Let $n_1 > 0$ and $n_2 > 0$ be given integers. By a* probability distribution*, we mean a collection of real numbers $p_{i,j} \geq 0$, $1 \leq i \leq n_1$, and $1 \leq j \leq n_2$, such that $\sum\limits_{i=1}^{n_1} \sum\limits_{j=1}^{n_2} p_{i,j} = 1$.*

Definition 2

- *We say that two integers i and i' are* neighbors *if $|i - i'| \leq 1$.*
- *We say that two points (i,j) and (i',j') are* neighbors *if i and i' are neighbors, and j and j' are neighbors.*

Definition 3 *Let $p_{i,j}$ be a probability distribution.*

- *We say that a point (i,j) is a mode (or a* local maximum*) if $p_{i,j} > 0$ and for every neighbor (i',j') of this point, $p_{i,j} \geq p_{i',j'}$.*
- *We say that a distribution $p_{i,j}$ is* strictly unimodal *if it has exactly one mode.*

Definition 4 *Let $p_{i,j}$ be a probability distribution.*

- *A set of integer points S is called* connected *if we can connect every two points $x = (i,j)$ and $x' = (i',j')$ from this set can be connected by a sequence of points in which every two sequential points are neighbors.*

- *A connected set S is called a* mode set *(or a* local maximum set*) if all the points $(i,j) \in S$ have the same value of $p_{i,j}$, and every point (i',j') outside S which is a neighbor to one of the points from S has a smaller value of p: $p_{i',j'} < p_{i,j}$.*
- *We say that a distribution $p_{i,j}$ is* unimodal *if it has exactly one mode set.*

Definition 5 *By a* linear constraint *on the probability distribution, we mean the constraint of the type $\underline{b} \le \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} b_{i,j} \cdot p_{i,j} \le \overline{b}$ for some given values $\underline{b}$, $\overline{b}$, and $b_{i,j}$.*

Definition 6

- *By an* interval probability problem under strict unimodality constraint*, we mean the following problem: given a find list of linear constraints, check whether there exists a strictly unimodal distribution which satisfies all these constraints.*
- *By an* interval probability problem under unimodality constraint*, we mean the following problem: given a find list of linear constraints, check whether there exists a unimodal distribution which satisfies all these constraints.*

Theorem 1 *Interval probability problem under unimodality constraint is NP-hard.*

Theorem 2 *Interval probability problem under strict unimodality constraint is NP-hard.*

Comment. So, under the unimodality constraints, even checking whether a system of linear constraints is consistent – i.e., whether the range of a given characteristic is empty – is computationally difficult (NP-hard).

Proof of Theorem 1. We will show that if we can check, for every system of linear constraints, whether this system is consistent or not under unimodality, then we would be able to solve a *partition* problem which is known to be NP-hard [10, 14]. The partition problem consists of the following: given n positive integers $s_1, \ldots, s_n$, check whether exist n integers $\varepsilon_i \in \{-1, 1\}$ for which $\varepsilon_1 \cdot s_1 + \ldots + \varepsilon_n \cdot s_n = 0$.

Indeed, for every instance of the partition problem, we form the following system of constraints: $n_1 = 2$, $n_2 = n + 2$,

- $p_{1,1} = p_{1,n+2} = 1/(n+2)$, $p_{2,1} = p_{2,n+2} = 0$;
- $p_{1,j+1} + p_{2,j+1} = 1/(n+2)$ for every $j = 1, \ldots, n$;
- $\sum_{j=1}^{n} (-s_j \cdot p_{1,j+1} + s_j \cdot p_{2,j+1}) = 0$.

Let us prove that this system is consistent if and only if the original instance of the partition problem has a solution.

"If" part. If the original instance has a solution $\varepsilon_i \in \{-1, 1\}$, then, for every j from 1 to n, we can take:

- if $\varepsilon_j = -1$, then we take $p_{1,j+1} = 1/(n+2)$ and $p_{2,j+1} = 0$;
- if $\varepsilon_j = 1$, then we take $p_{1,j+1} = 0$ and $p_{2,j+1} = 1/(n+2)$.

We also take $p_{1,1} = p_{1,n+2} = 1/(n+2)$ and $p_{2,1} = p_{2,n+2} = 0$.

The resulting distribution is unimodal. Indeed, every probability $p_{i,j}$ in this distribution is either 0, or $1/(n+2)$. The set S of all the points at which $p_{i,j} = 1/(n+2)$ is connected, so this set is the mode set – and clearly the only mode set.

Let us check that this distribution satisfies all the desired constraints. We explicitly selected $p_{1,1} = p_{1,n+2} = 1/(n+2)$ and $p_{2,1} = p_{2,n+2} = 0$, so the first constraint is satisfied. It is easy to check that for every j, we have $p_{1,j+1} + p_{2,j+1} = 1/(n+2)$. Finally, due to our choice of $p_{i,j}$, we conclude that $-s_j \cdot p_{1,j+1} + s_j \cdot p_{2,j+1} = \dfrac{1}{n+2} \cdot \varepsilon_j \cdot s_j$ and thus,

$$\sum_{j=1}^{n} (-s_j \cdot p_{1,j+1} + s_j \cdot p_{2,j+1}) = \frac{1}{n+2} \cdot \sum_{j=1}^{n} \varepsilon_j \cdot s_j = 0.$$

"Only if" part. Vice versa, let us assume that we have a unimodal distribution $p_{i,j}$ for which all the desired constraints are satisfied.

For $j = 1$ and $j = n+2$, we have $p_{1,j} = 1/(n+2)$ and $p_{2,j} = 0$. For all other j, we have $p_{1,j} + p_{2,j} = 1/(n+2)$ and thus, $p_{1,j} \le 1/(n+2)$ and $p_{2,j} \le 1/(n+2)$; so, all the probabilities $p_{i,j}$ are smaller than or equal to $1/(n+2)$, and $\max\limits_{i,j} p_{i,j} = 1/(n+2) = p_{1,1} = p_{1,n+2}$. Since the value $p_{1,1}$ is a global maximum, the point $(1, 1)$ has to be a part of a local maximum (mode) set. Similarly, the point $(1, n+2)$ has to be a part of a mode set. Since the distribution is unimodal, there exists only one mode set, so this set S must contain both $(1, 1)$ and $(1, n+2)$. By definition of a mode set, S is connected, and we must have the same probability $p_{i,j} = 1/(n+2)$ for all the points $(i, j) \in S$.

Since the set S is connected, the elements $(1, 1) \in S$ and $(1, n+2) \in S$ must be connectable by a sequence of points in which every two sequential points are neighbors. When the points are neighbors, their coordinates cannot differ by more than 1. Thus, the second coordinates of the connecting points from S must take all the integers from 1 to $n+2$, without any gaps. Hence, for every j from 1 to n, the set S must contain one of the two points $(m(j), j+1)$, for some $m(j) \in \{1, 2\}$. Here:

- If $m(j) = 1$, i.e., if $(1, j+1) \in S$, then $p_{1,j+1} = 1/(n+2)$ and therefore, since $p_{1,j+1} + p_{2,j+1} = 1/(n+2)$, we conclude that $p_{2,j+1} = 0$.
- If $m(j) = 2$, i.e., if $(2, j+1) \in S$, then $p_{2,j+1} = 1/(n+2)$ and therefore, since $p_{1,j+1} + p_{2,j+1} = 1/(n+2)$, we conclude that $p_{1,j+1} = 0$.

If we denote $\varepsilon_j \stackrel{\text{def}}{=} 2m(j) - 3$, then we conclude that $\varepsilon_j \in \{-1, 1\}$. For each j, we have

$$-s_j \cdot p_{1,j+1} + s_j \cdot p_{2,j+1} = \varepsilon_j \cdot s_j \cdot \frac{1}{n+2},$$

hence from the constraint

$$\sum_{j=1}^{n}(-s_j \cdot p_{1,j+1} + s_j \cdot p_{2,j+1}) = \frac{1}{n+2} \cdot \sum_{j=1}^{n} \varepsilon_j \cdot s_j = 0,$$

we conclude that $\sum \varepsilon_j \cdot s_j = 0$, i.e., that the original instance of the partition problem has a solution.

The theorem is proven.

Comment. The above constraints are not just mathematical tricks, they have a natural interpretation if for x_1, we take the values -1 and 1 as corresponding to $i = 1, 2,$, and for x_2, we take the values $0, s_1, \ldots, s_n, S$, where $S > \max s_i$. Then:

- the constraint $p_{1,j+1} + p_{2,j+1} = 1/(n+2)$ means that $\text{Prob}(x_2 = s_i) = 1/(n+2)$ for all n values s_i, and
- the constraint $\sum\limits_{j=1}^{n}(-s_j \cdot p_{1,j+1} + s_j \cdot p_{2,j+1}) = 0$ means that the conditional expected value of the product is 0: $E[x_1 \cdot x_2 \,|\, 0 < x_2 < S] = 0$.

So, the difficult-to-solve problem is to check whether it is possible that

$$E[x_1 \cdot x_2 \,|\, 0 < x_2 < S] = 0$$

for some unimodal distribution for which the marginal distribution on x_2 is "uniform", and for which on the edges, i.e., for $x_2 = 0$ and $x_2 = S$, we have $x_1 = -1$.

Proof of Theorem 2. We will prove this theorem by using the same reduction to the partition problem that we used in the previous proof.

Let us select a small real number $\alpha > 0$, and for every instance $s_1, \ldots, s_n$ of the partition problem, form the following system of constraints: $n_1 = 2$, $n_2 = n + 2$,

- $p_{1,1} = \dfrac{1}{n+2} \cdot \left(1 + \alpha \cdot \left(1 - \dfrac{n+3}{2}\right)\right)$; $p_{2,1} = 0$;
- $p_{1,n+2} = \dfrac{1}{n+2} \cdot \left(1 + \alpha \cdot \left((n+2) - \dfrac{n+3}{2}\right)\right)$; $p_{2,n+2} = 0$;
- $p_{1,j+1} + p_{2,j+1} = \dfrac{1}{n+2} \cdot \left(1 + \alpha \cdot \left((j+1) - \dfrac{n+3}{2}\right)\right)$ for every $j = 1, \ldots, n$;

- $-\alpha \cdot \dfrac{n}{2} \le \sum\limits_{j=1}^{n} (-s_j \cdot p_{1,j+1} + s_j \cdot p_{2,j+1}) \le \alpha \cdot \dfrac{n}{2}$.

Let us prove that for sufficient small α this system is consistent if and only if the original instance of the partition problem has a solution.

The condition that α is sufficiently small includes the requirement that $\alpha < p_{1,1}$ which is clearly satisfied for all sufficiently small values α.

"If" part. If the original instance has a solution $\varepsilon_i \in \{-1, 1\}$, then, for every j from 1 to n, we can take:

- if $\varepsilon_j = -1$, then we take $p_{1,j+1} = \dfrac{1}{n+2} \cdot \left(1 + \alpha \cdot \left((j+1) - \dfrac{n+3}{2}\right)\right)$ and $p_{2,j+1} = 0$;
- if $\varepsilon_j = 1$, then we take $p_{2,j+1} = \dfrac{1}{n+2} \cdot \left(1 + \alpha \cdot \left((j+1) - \dfrac{n+3}{2}\right)\right)$ and $p_{1,j+1} = 0$.

We also take $p_{1,1} = \dfrac{1}{n+2} \cdot \left(1 + \alpha \cdot \left(1 - \dfrac{n+3}{2}\right)\right)$,

$$p_{1,n+2} = \frac{1}{n+2} \cdot \left(1 + \alpha \cdot \left((n+2) - \frac{n+3}{2}\right)\right),$$

and $p_{2,1} = p_{2,n+2} = 0$.

The resulting distribution is strictly unimodal. Indeed, every non-zero probability $p_{i,j}$ in this distribution is strictly smaller than the non-zero probability corresponding to $j+1$, and the only local maximum (mode) is the point $(1, n+2)$.

Let us check that this distribution satisfies all the desired constraints. We explicitly selected the desired values for $p_{1,1}$, $p_{1,n+2}$, and $p_{2,1} = 0$, and $p_{2,n+2} = 0$, so the first constraint is satisfied. It is easy to check that for every j, we have the desired value for $p_{1,j+1} + p_{2,j+1}$.

As in the proof of the previous result, we conclude that the α-free part of the sum $-s_j \cdot p_{1,j+1} + s_j \cdot p_{2,j+1} = \dfrac{1}{n+2} \cdot \varepsilon_j \cdot s_j$ is 0. The α-dependent part can be bounded by $\alpha \cdot \dfrac{1}{n+2} \cdot \sum\limits_{j=1}^{n} \left| j - \dfrac{n+3}{2} \right|$. Each term in the sum is bounded by $\dfrac{n+2}{2}$, hence the sum is bounded by $\dfrac{n \cdot (n+2)}{2}$, and the entire expression is bounded by $\alpha \cdot \dfrac{n}{2}$, as desired.

"Only if" part. Vice versa, let us assume that we have a strictly unimodal distribution $p_{i,j}$ for which all the desired constraints are satisfied.

Due to our selection of constraints, the sum $p_{1,j} + p_{2,j}$ increases with j; for $j = 1$ and $j = n+2$, this whole sum corresponds to just one value $p_{1,j}$. Thus,

for every $j < n+2$ and for every i, we have

$$p_{i,j} \le p_{1,j} + p_{2,j} < p_{1,n+2} + p_{2,n+2} = p_{1,n+2}.$$

So, the point $(1, n+2)$ is the global maximum of the probability distribution $p_{i,j}$. Since the distribution $p_{i,j}$ is strictly unimodal, the function $p_{i,j}$ only has one local maximum – the same as its global maximum $(1, n+2)$.

Let us use this conclusion to prove that there exist values $m(1), \ldots, m(n)$ for which

$$p_{1,1} < p_{m(1),2} < \ldots < p_{m(j),j+1} < \ldots < p_{m(n),n+1} < p_{1,n+2},$$

$p_{m(j),j+1} > p_{3-m(j),j+1}$ and $p_{3-m(j),j+1} < \alpha$ for all $j = 1, \ldots, n$.

We will prove the existence of the desired sequence $m(j)$ by induction: namely, we prove that for every j from 1 to n, there exists a sequence such that

$$p_{1,1} < p_{m(1),2} < \ldots < p_{m(j),j+1},$$

and for which $p_{m(j),j+1} > p_{3-m(j),j+1}$ and $p_{3-m(j),j+1} < \alpha$.

Induction base $j = 1$. Since $p_{1,1}$ is not a local maximum, and $p_{1,1} > p_{2,1} = 0$, we must therefore have a larger value of $p_{i,j}$ at one of the other two neighboring points. So, either for $m(1) = 1$, or for $m(1) = 2$, we must have $p_{1,1} < p_{m(1),2}$.

Since $p_{m(1),2} > p_{1,1} = \dfrac{1}{n+2} \cdot \left(1 + \alpha \cdot \left(1 - \dfrac{n+3}{2}\right)\right)$ and

$$p_{m(1),2} + p_{3-m(1),2} = \frac{1}{n+2} \cdot \left(1 + \alpha \cdot \left(2 - \frac{n+3}{2}\right)\right),$$

we conclude that

$$p_{3-m(1),2} = (p_{m(1),2} + p_{3-m(1),2}) - p_{m(1),2} <$$

$$(p_{m(1),2} + p_{3-m(1),2}) - p_{1,1} = \frac{1}{n+2} \cdot \alpha.$$

This value is $< \alpha$. Since we selected α for which $\alpha < p_{1,1}$, the value $p_{3-m(1),2}$ is smaller than $p_{1,1}$ and hence, smaller that $p_{m(1),2}$.

Induction step. Let us assume that we have already proven the existence of the sequence $m(1), \ldots, m(j-1)$, and let us prove the existence of the term $m(j)$. From the induction assumption, we conclude that $p_{m(j-1),j} > p_{3-m(j-1),j}$ and that $p_{m(j-1),j} > p_{m(j-2),j-1}$. Since $p_{m(j-2),j-1} > p_{3-m(j-2),j-1}$, we thus conclude that $p_{m(j-1),j} > p_{3-m(j-2),j-1}$. Thus, the value $p_{m(j-1),j}$ is larger than the values at 3 neighboring points. Since this value cannot be a local maximum, this means that there must exist a neighboring point with a larger value, i.e., there must exist $m(j) \in \{1, 2\}$ for which $p_{m(j-1),j} < p_{m(j),j+1}$.

To complete the induction proof, we must show that $p_{3-m(j),j+1} < \alpha$ and $p_{m(j),j+1} > p_{3-m(j),j+1}$. Indeed, since

$$p_{m(j),j+1} > p_{1,1} = \frac{1}{n+2} \cdot \left(1 + \alpha \cdot \left(1 - \frac{n+3}{2}\right)\right)$$

and

$$p_{m(j),j+1} + p_{3-m(j),j+1} = \frac{1}{n+2} \cdot \left(1 + \alpha \cdot \left((j+1) - \frac{n+3}{2}\right)\right),$$

we conclude that

$$p_{3-m(j),j+1} = (p_{m(j),j+1} + p_{3-m(j),j+1}) - p_{m(j),j+1} <$$

$$(p_{m(j),j+1} + p_{3-m(j),j+1}) - p_{1,1} = \frac{j}{n+2} \cdot \alpha.$$

For $j \le n$, the value $\frac{j}{n+2} \cdot \alpha$ is smaller than α. Since $p_{3-m(j),j+1} < \alpha$, $\alpha < p_{1,1}$, and $p_{1,1} < p_{m(j),j+1}$, we thus conclude that $p_{3-m(j),j+1} < p_{m(j),j+1}$.

For each j from 1 to n, we have $0 \le p_{3-m(j),j+1} < \alpha$ and $p_{1,1} < p_{m(j),j+1} < p_{1,n+2}$. By our selection of $p_{1,1}$ and $p_{1,n+2}$, we conclude that $p_{1,n+2} - p_{1,1} = \frac{n+1}{n+2} \cdot \alpha < \alpha$, hence $p_{1,n+2} < p_{1,1} + \alpha$, and so $p_{1,1} < p_{m(j),j+1} < p_{1,1} + \alpha$.

Let us $\varepsilon_j \stackrel{\text{def}}{=} 2m(j) - 3$, then $\varepsilon_j \in \{-1, 1\}$. If $m(j) = 2$, then $\varepsilon_j = 1$, and from $p_{1,1} < p_{2,j+1} < p_{1,1} + \alpha$ and $0 \le p_{3-m(j),j+1} < \alpha$, we conclude that $p_{1,1} - \alpha < -p_{1,j+1} + p_{2,j+1} < p_{1,1} + \alpha$, hence

$$s_j \cdot p_{1,1} - \alpha \cdot s_j < -s_j \cdot p_{1,j+1} + s_j \cdot p_{2,j+1} < s_j \cdot p_{1,1} + \alpha \cdot s_j.$$

Similarly, if $m(j) = 1$ and $\varepsilon_j = -1$, we conclude that

$$-s_j \cdot p_{1,1} - \alpha \cdot s_j < -s_j \cdot p_{1,j+1} + s_j \cdot p_{2,j+1} < -s_j \cdot p_{1,1} + \alpha \cdot s_j.$$

In general,

$$\varepsilon_j \cdot s_j \cdot p_{1,1} - \alpha \cdot s_j < -s_j \cdot p_{1,j+1} + s_j \cdot p_{2,j+1} < \varepsilon_j \cdot s_j \cdot p_{1,1} + \alpha \cdot s_j.$$

Thus,

$$\sum_{j=1}^{n} \varepsilon_j \cdot s_j \cdot p_{1,1} - \alpha \cdot \sum_{j=1}^{n} s_j < \sum_{j=1}^{n} (-s_j \cdot p_{1,j+1} + s_j \cdot p_{2,j+1}) < \sum_{j=1}^{n} \varepsilon_j \cdot s_j \cdot p_{1,1} + \alpha \cdot \sum_{j=1}^{n} s_j.$$

Since the constraints are satisfied, we have

$$-\alpha \cdot \frac{n}{2} \le \sum_{j=1}^{n} (-s_j \cdot p_{1,j+1} + s_j \cdot p_{2,j+1}) \le \alpha \cdot \frac{n}{2}$$

and therefore,

$$\sum_{j=1}^{n} \varepsilon_j \cdot s_j \cdot p_{1,1} - \alpha \cdot \sum_{j=1}^{n} s_j < \alpha \cdot \frac{n}{2}$$

and

$$\sum_{j=1}^{n} \varepsilon_j \cdot s_j \cdot p_{1,1} + \alpha \cdot \sum_{j=1}^{n} s_j > -\alpha \cdot \frac{n}{2}.$$

Dividing both sides of these inequalities by $p_{1,1} = \frac{1}{n+2} \cdot \left(1 + \alpha \cdot \left(1 - \frac{n+3}{2}\right)\right)$ and combining these two inequalities, we conclude that $-\beta < \sum\limits_{j=1}^{n} \varepsilon_j \cdot s_j < \beta$, where

$$\beta = \frac{\alpha \cdot \left(\sum\limits_{j=1}^{n} s_j + \frac{n}{2}\right)}{\frac{1}{n+2} \cdot \left(1 + \alpha \cdot \left(1 - \frac{n+3}{2}\right)\right)}.$$

When $\alpha \to 0$, we have $\alpha / p_{1,1} \to 0$ and $\beta \to 0$. Thus, for sufficiently small α, we have $\beta < 1$. For such α, since the sum $\sum\limits_{j=1}^{n} \varepsilon_j \cdot s_j$ is an integer, its only possible value is 0. So, the original instance of the partition problem has a solution.

The theorem is proven.

3 Adding Conditional Unimodality Makes Interval Probability Problems NP-Hard

In some practical situations, we are not sure that the 2-D distribution is unimodal, but we know that, for every value of x_2, the corresponding 1-D *conditional* distribution for x_1 is unimodal. In this case, to describe this as a LP problem, we must select a mode for every x_2. If there are n values of x_2, and at least 2 possible choices of mode location, then we get an exponential amount of 2^n possible choices. In this section, we show that this problem is also NP-hard – and therefore, that, unless P=NP, no algorithm can solve it in polynomial time.

Definition 7 *Let $n_1 > 0$ and $n_2 > 0$ be given integers, and let $p_{i,j}$ be a probability distribution.*

- *We say that the distribution $p_{i,j}$ is* conditionally unimodal in the 1st variable *(or* 1-unimodal*, for short) if for every j from 1 to n_2, there exists a value $m(j)$ such that $p_{i,j}$ grows with i for $i \le m(j)$ and decreases with i for $i \ge m(j)$:*

$$p_{1,j} \le p_{2,j} \le \ldots \le p_{m(j),j} \ge p_{m(j)+1,j} \ge \ldots \ge p_{n_1,j}.$$

- *We say that the distribution $p_{i,j}$ is* conditionally unimodal in the 2nd variable *(or* 2-unimodal*, for short) if for every i from 1 to n_1, there exists a value $m(i)$ such that $p_{i,j}$ grows with j for $j \le m(i)$ and decreases with j for $j \ge m(i)$:*

$$p_{i,1} \le p_{i,2} \le \ldots \le p_{i,m(i)} \ge p_{i,m(i)+1} \ge \ldots \ge p_{i,n_2}.$$

- *We say that the distribution $p_{i,j}$ is* conditionally unimodal *if it is both 1-unimodal and 2-unimodal.*

- *By an* interval probability problem under 1-unimodality constraint*, we mean the following problem: given a find list of linear constraints, check whether there exists a 1-unimodal distribution which satisfies all these constraints.*

- *By an* interval probability problem under conditional unimodality constraint*, we mean the following problem: given a find list of linear constraints, check whether there exists a conditionally unimodal distribution which satisfies all these constraints.*

Theorem 3 *Interval probability problem under 1-unimodality constraint is NP-hard.*

Comments.

- If the coefficients of the linear equalities are rational, then this problem is in the class NP and is, thus, not only NP-hard but also NP-complete. (The authors are thankful to the referees for this observation.)

- This result clearly means that the interval probability problem under 2-unimodality constraint is also NP-complete.

Theorem 4 *Interval probability problem under conditional unimodality constraint is NP-hard.*

Comments. Similarly to Theorem 3, one can easily see that if the coefficients of the linear equalities are rational, then this problem is in the class NP and is, thus, not only NP-hard but also NP-complete.

Proof of Theorem 3. We will prove this theorem by using the same reduction to the partition problem that we used in the previous proofs.

Specifically, for every instance of the partition problem, we form the following system of constraints: $n_1 = 3$, $n_2 = n$,

- $p_{2,j} = 0$ for every $j = 1, \ldots, n_2$,

- $p_{1,j} + p_{2,j} + p_{3,j} = 1/n$ for every $j = 1, \ldots, n_2$;

- $\sum_{j=1}^{n_2}(-s_j \cdot p_{1,j} + s_j \cdot p_{3,j}) = 0$.

Let us prove that this system is consistent if and only if the original instance of the partition problem has a solution.

"If" part. If the original instance has a solution $\varepsilon_i \in \{-1, 1\}$, then, for every j from 1 to n_2, we can take $p_{2+\varepsilon_j,j} = 1/n$ and $p_{i,j} = 0$ for $i \neq 2 + \varepsilon_j$. In other words:

- if $\varepsilon_j = -1$, then we take $p_{1,j} = 1/n$ and $p_{2,j} = p_{3,j} = 0$;
- if $\varepsilon_j = 1$, then we take $p_{1,j} = p_{2,j} = 0$ and $p_{3,j} = 1/n$.

The resulting distribution is 1-unimodal: indeed, for each j, its mode is the value $1 + \varepsilon_j$. Let us check that it satisfies all the desired constraints. It is easy to check that for every j, we have $p_{2,j} = 0$ and $p_{1,j} + p_{2,j} + p_{3,j} = 1/n$. Finally, due to our choice of $p_{i,j}$, we conclude that $-s_j \cdot p_{1,j} + s_j \cdot p_{3,j} = \dfrac{1}{n} \cdot \varepsilon_j \cdot s_j$ and thus,

$$\sum_{j=1}^{n_2}(-s_j \cdot p_{1,j} + s_j \cdot p_{3,j}) = \frac{1}{n} \cdot \sum_{j=1}^{n_2} \varepsilon_j \cdot s_j = 0.$$

"Only if" part. Vice versa, let us assume that we have a 1-unimodal distribution $p_{i,j}$ for which all the desired constraints are satisfied. Since the distribution is 1-unimodal, for every j, there exists a mode $m(j) \in \{1, 2, 3\}$ for which the values $p_{i,j}$ increase for $i \leq m(j)$ and decrease for $i \geq m(j)$. This mode cannot be equal to 2, because otherwise, the value $p_{2,j} = 0$ will be the largest of the three values $p_{1,j}$, $p_{2,j}$, and $p_{3,j}$ hence all three values will be 0 – which contradicts to the constraint $p_{1,j} + p_{2,j} + p_{3,j} = 1/n$. Thus, this mode is either 1 or 3:

- if the mode is 1, then due to monotonicity, we have $0 = p_{2,j} \geq p_{3,j}$ hence $p_{3,j} = p_{2,j} = 0$;
- if the mode is 3, then due to monotonicity, we have $p_{1,j} \leq p_{2,j} = 0$ hence $p_{1,j} = p_{2,j} = 0$.

In both case, for each j, only one value of $p_{i,j}$ is different from 0 – the value $p_{m(j),j}$. Since the sum of these three values is $1/n$, this non-zero value must be equal to $1/n$. If we denote $\varepsilon_j \stackrel{\text{def}}{=} m(j) - 2$, then we conclude that $\varepsilon_j \in \{-1, 1\}$. For each j, we have

$$-s_j \cdot p_{1,j} + s_j \cdot p_{3,j} = \varepsilon_j \cdot s_j \cdot (1/n),$$

hence from the constraint

$$\sum_{j=1}^{n_2}(-s_j \cdot p_{1,j} + s_j \cdot p_{3,j}) = \frac{1}{n} \cdot \sum_{j=1}^{n_2} \varepsilon_j \cdot s_j = 0,$$

we conclude that $\sum \varepsilon_j \cdot s_j = 0$, i.e., that the original instance of the partition problem has a solution.

The theorem is proven.

Comment. The above constraints are not just mathematical tricks, they have a natural interpretation if for x_1, we take the values -1, 0, and 1 as corresponding to $i = 1, 2, 3$, and for x_2, we take the values $s_1, \ldots, s_n$. Then:

- the constraint $p_{2,j} = 0$ means that $\text{Prob}(x_1 = 0) = 0$;
- the constraint $p_{1,j} + p_{2,j} + p_{3,j} = 1/n$ means that $\text{Prob}(x_2 = s_i) = 1/n$ for all n values s_i, and
- the constraint $\sum\limits_{j=1}^{n_2} (-s_j \cdot p_{1,j} + s_j \cdot p_{3,j}) = 0$ means that the expected value of the product is 0: $E[x_1 \cdot x_2] = 0$.

So, the difficult-to-solve problem is to check whether it is possible that $E[x_1 \cdot x_2] = 0$ and $\text{Prob}(x_1 = 0) = 0$ for some 1-unimodal distribution for which the marginal distribution on x_2 is "uniform".

Proof of Theorem 4. To prove this result, we will reduce, to this problem, the same partition problem as in the proof of the previous theorems.

For every instance of the partition problem, we form the following system of constraints: $n_1 = 3 \cdot n$, $n_2 = n$,

- $p_{i,j} = 0$ for every $i = 1, \ldots, n_2$ and for every $i \neq 3 \cdot j$ and $i \neq 3 \cdot j - 2$;
- $\sum\limits_{i=1}^{n_1} p_{i,j} = 1/n$ for every $j = 1, \ldots, n_2$;
- $\sum\limits_{j=1}^{n_2} (-s_j \cdot p_{3 \cdot j-2,j} + s_j \cdot p_{3 \cdot j,j}) = 0$.

Let us prove that this system is consistent if and only if the original instance of the partition problem has a solution.

"If" part. If the original instance has a solution $\varepsilon_i \in \{-1, 1\}$, then, for every j from 1 to n_2, we can take $p_{3 \cdot j-1+\varepsilon_j,j} = 1/n$ and $p_{i,j} = 0$ for $i \neq 3 \cdot j - 1 + \varepsilon_j$. In other words:

- if $\varepsilon_j = -1$, then we take $p_{3 \cdot j-2,j} = 1/n$ and $p_{3 \cdot j,j} = 0$;
- if $\varepsilon_j = 1$, then we take $p_{3 \cdot j-2,j} = 0$ and $p_{3 \cdot j,j} = 1/n$.

The resulting distribution is 1-unimodal: indeed, for each j, its mode is the value $3 \cdot j - 1 + \varepsilon_j$. Similarly, it is 2-unimodal, because for each i, only one probability $p_{i,j}$ may be different from 0. Similarly to the proof of Theorem 3, we can check that this distribution satisfies all the desired constraints.

"Only if" part. Vice versa, let us assume that we have a conditionally unimodal distribution $p_{i,j}$ for which all the desired constraints are satisfied. Since the distribution is conditionally unimodal, it is 1-unimodal, so for every j, there exists a mode $m(j) \in \{3\cdot j-2, 3\cdot j\}$ for which the values $p_{i,j}$ increase for $i \le m(j)$ and decrease for $i \ge m(j)$. Similarly to the proof of Theorem 3, we conclude that for each j, only one value of $p_{i,j}$ is different from 0 – the value $p_{m(j),j}$, and this non-zero value is equal to $1/n$. If we denote $\varepsilon_i \stackrel{\text{def}}{=} m(i) - (3 \cdot i - 1)$, then we conclude that $\varepsilon_i \in \{-1, 1\}$. For each j, we have

$$-s_j \cdot p_{3\cdot j-2,j} + s_j \cdot p_{3\cdot j,j} = \varepsilon_j \cdot s_j \cdot (1/n),$$

hence from the constraint

$$\sum_{j=1}^{n_2}(-s_j \cdot p_{3\cdot j-2,j} + s_j \cdot p_{3\cdot j,j}) = \frac{1}{n} \cdot \sum_{j=1}^{n_2} \varepsilon_j \cdot s_j = 0,$$

we conclude that $\sum \varepsilon_i \cdot s_i = 0$, i.e., that the original instance of the partition problem has a solution.

The theorem is proven.

Comment. We can get a natural interpretation of the above constraints if for $s \stackrel{\text{def}}{=} 3 \cdot \max_i s_i$, we take the values $x_2 = s \cdot j$ corresponding to $j = 1, \ldots, n_2$, and for i:

- we take the value $x_1 = s \cdot j$ corresponding to $i = 3 \cdot j - 1$;
- we take the value $x_1 = s \cdot j - s_j$ corresponding to $i = 3 \cdot j - 2$; and
- we take the value $x_1 = s \cdot j + s_j$ corresponding to $i = 3 \cdot j$.

In this interpretation, the above constraint $\sum\limits_{j=1}^{n_2}(-s_j \cdot p_{3\cdot j-2,j} + s_j \cdot p_{3\cdot j,j}) = 0$ simply means that $E[x_1] = E[x_2]$.

4 Adding Quasiconcavity Makes Interval Probability Problems NP-Hard

Another formalization of the intuitive notion of a unimodal ("single-cluster") distribution is that for every α, the level sets $\{x : \rho(x) \ge \alpha\}$ should be convex. Such distributions are called *quasiconcave*. Let us show that under this formalization, the problem is also NP-hard.

Definition 8 *We say that a set S of integer points (i, j) is* convex *if it contains all integer points from its convex hull.*

Comment. For example, the set S_1 consisting of the points (0,0), (0,1), (1,0), and (1,1) is convex – because its convex hull is the unit square $[0,1] \times [0,1]$, and S contains all 4 integer points from the unit square.

On the other hand, the set S_2 consisting of the points (0,0) and (0,2) is not convex, because its convex hull contains an integer point (0,1) which is not contained in S_2.

Definition 9

- *We say that a probability distribution $p_{i,j}$ is* quasiconcave *if for every value p_0, the set $\{(i,j) : p_{i,j} \geq p_0\}$ is convex.*
- *By an* interval probability problem under quasiconcavity constraint*, we mean the following problem: given a find list of linear constraints, check whether there exists a quasiconcave distribution which satisfies all these constraints.*

Theorem 5 *Interval probability problem under quasiconcavity constraint is NP-hard.*

Proof of Theorem 5. We will prove this theorem by using the same reduction to the partition problem that we used in the previous proofs.

Specifically, for every instance $s_1, \ldots, s_n$ of the partition problem, we select the value $\alpha = \dfrac{2}{n \cdot (3n+4)}$ and then form the following system of constraints: $n_1 = 3$, $n_2 = n$,

- for every j from 1 to n: $p_{1,j} \geq j \cdot \alpha$, $p_{2,j} = j \cdot \alpha$, $p_{3,j} \geq j \cdot \alpha$, and $p_{1,j} + p_{3,j} = (2j + 0.5) \cdot \alpha$, and
- $\sum\limits_{j=1}^{n_2} (-s_j \cdot p_{1,j} + s_j \cdot p_{3,j}) = 0$.

Comment. The value α is selected so as to make $\sum\limits_{i=1}^{3} \sum\limits_{j=1}^{n} p_{i,j} = 1$: for each j, we have $p_{1,j} + p_{2,j} + p_{3,j} = (3j + 0.5) \cdot \alpha$, hence

$$\sum_{i,j} p_{i,j} = \alpha \cdot \sum_{j=1}^{n} (3j + 0.5) = \alpha \cdot \left(3 \frac{n \cdot (n+1)}{2} + \frac{n}{2} \right) = \alpha \cdot \frac{n \cdot (3n+4)}{2}.$$

"If" part. If the original instance has a solution $\varepsilon_i \in \{-1, 1\}$, then, for every j from 1 to n, we can take:

- if $\varepsilon_j = -1$, then we take $p_{1,j} = (j + 0.5) \cdot \alpha$ and $p_{2,j} = p_{3,j} = j \cdot \alpha$;
- if $\varepsilon_j = 1$, then we take $p_{1,j} = p_{2,j} = j \cdot \alpha$ and $p_{3,j} = (j + 0.5) \cdot \alpha$.

Let us prove that the resulting distribution is quasiconcave. Indeed, possible values of $p_{i,j}$ are $j \cdot \alpha$ and $(j + 0.5) \cdot \alpha$.

- for $p_0 = j \cdot \alpha$, the level set $S = \{(i, j) \,|\, p_{i,j} \geq p_0\}$ consists of all the rows $\{(1, j), (2, j), (3, j\}$ starting from j-th – which is clearly a convex set;
- for $p_0 = (j + 0.5) \cdot \alpha$, the level set $S = \{(i, j) \,|\, p_{i,j} \geq p_0\}$ consists of all the rows starting from $(j + 1)$-th plus two elements (midpoint $i = 2$ plus one of the endpoints) from row j – which is also clearly a convex set.

Let us check that it satisfies all the desired constraints. It is easy to check that for every j, we have $p_{1,j} \geq j \cdot \alpha$, $p_{2,j} = j \cdot \alpha$, $p_{3,j} \geq j \cdot \alpha$, and $p_{1,j} + p_{3,j} = (2j + 0.5) \cdot \alpha$.

Finally, due to our choice of $p_{i,j}$, we conclude that $-s_j \cdot p_{1,j} + s_j \cdot p_{3,j} = 0.5 \cdot \alpha \cdot \varepsilon_j \cdot s_j$ and thus,

$$\sum_{j=1}^{n} (-s_j \cdot p_{1,j} + s_j \cdot p_{3,j}) = 0.5 \cdot \alpha \cdot \sum_{j=1}^{n} \varepsilon_j \cdot s_j = 0.$$

"Only if" part. Vice versa, let us assume that we have a quasiconcave distribution $p_{i,j}$ for which all the desired constraints are satisfied. Let us fix j from 1 to n. For $p_0 = \min(p_{1,j}, p_{3,j})$, we have $p_{1,j} \geq p_0$ and $p_{3,j} \geq p_0$ and therefore, both points $(1, j)$ and $(3, j)$ belong to the level set $S = \{(i, j) \,|\, p_{i,j} \geq p_0\}$.

Since the distribution is quasiconcave, the set S is convex, and thus, the set S must also contain the midpoint $(2, j) = \dfrac{(1, j) + (3, j)}{2}$. The fact that $(2, j) \in S$ means that $p_{2,j} \geq p_0 = \min(p_{1,j}, p_{3,j})$. One of our constraints is that $p_{2,j} = j \cdot \alpha$, so we can conclude that the smallest of the two values $p_{1,j}$ and $p_{3,j}$ must be $\leq j \cdot \alpha$. Since, according to our constraints, both values $p_{1,j}$ and $p_{3,j}$ must be $\geq j \cdot \alpha$, we thus conclude that one of these two values must be exactly equal to $j \cdot \alpha$. Since their sum must be equal to $(2j + 0.5) \cdot \alpha$, we deduce that the other value is equal to $(j + 0.5) \cdot \alpha$.

Let us define ε_j as follows:

- $\varepsilon_j = -1$ if $p_{1,j} = (j + 0.5) \cdot \alpha$ and $p_{3,j} = j \cdot \alpha$, and
- $\varepsilon_j = 1$ if $p_{1,j} = j \cdot \alpha$ and $p_{3,j} = (j + 0.5) \cdot \alpha$.

For each j, we have

$$-s_j \cdot p_{1,j} + s_j \cdot p_{3,j} = 0.5 \cdot \alpha \cdot \varepsilon_j \cdot s_j,$$

hence from the constraint

$$\sum_{j=1}^{n} (-s_j \cdot p_{1,j} + s_j \cdot p_{3,j}) = 0.5 \cdot \alpha \cdot \sum_{j=1}^{n} \varepsilon_j \cdot s_j = 0,$$

we conclude that $\sum \varepsilon_j \cdot s_j = 0$, i.e., that the original instance of the partition problem has a solution.

The theorem is proven.

5 Adding Independence Makes Interval Probability Problems NP-Hard

In general, in statistics, independence makes problems easier. We will show, however, that for interval probability problems, the situation is sometimes opposite: the addition of independence assumption turns easy-to-solve problems into NP-hard ones.

Definition 10 *Let $n_1 > 0$ and $n_2 > 0$ be given integers.*

- *By an* independent probability distribution*, we mean a collection of real numbers $p_i \geq 0$, $1 \leq i \leq n_1$, and q_j, $1 \leq j \leq n_2$, such that $\sum\limits_{i=1}^{n_1} p_i = \sum\limits_{j=1}^{n_2} q_j = 1$.*
- *By a* linear constraint *on the independent probability distribution, we mean a constraint of the form*

$$\underline{b} \leq \sum_{i=1}^{n_1} a_i \cdot p_i + \sum_{j=1}^{n_2} b_j \cdot q_j + \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} c_{i,j} \cdot p_i \cdot q_j \leq \overline{b}$$

  for some given values $\underline{b}$, $\overline{b}$, a_i, b_j, and $c_{i,j}$.
- *By an* interval probability problem under independence constraint*, we mean the following problem: given a find list of linear constraints, check whether there exists an independent distribution which satisfies all these constraints.*

Comment. Independence means that $p_{i,j} = p_i \cdot q_j$ for every i and j. The above constraints are linear in terms of these probabilities $p_{i,j} = p_i \cdot q_j$.

Theorem 6 *Interval probability problem under independence constraint is NP-hard.*

Proof. To prove this theorem, we will reduce the problem in question to the same known NP-hard problem as in the proof of Theorem 1: to the partition problem.

For every instance of the partition problem, we form the following system of constraints: $n_1 = n_2 = n$,

- $p_i - q_i = 0$ for every i from 1 to n;
- $S_i \cdot p_i - p_i \cdot q_i = 0$ for all i from 1 to n,

where

$$S_i \stackrel{\text{def}}{=} \frac{2 \cdot s_i}{\sum\limits_{k=1}^{n} s_k}.$$

Let us prove that this system is consistent if and only if the original instance of the partition problem has a solution.

Indeed, if the original instance has a solution $\varepsilon_i \in \{-1, 1\}$, then, for every i from 1 to n, we can take $p_i = q_i = \frac{1+\varepsilon_i}{2} \cdot S_i$, i.e.:

- if $\varepsilon_i = -1$, we take $p_i = q_i = 0$;
- if $\varepsilon_i = 1$, we take $p_i = q_i = S_i$.

Let us show that for this choice, $\sum\limits_{i=1}^{n} p_i = \sum\limits_{j=1}^{n} q_j = 1$. Indeed,

$$\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} \frac{1+\varepsilon_i}{2} \cdot S_i = \frac{1}{2} \cdot \sum_{i=1}^{n} S_i + \frac{1}{2} \cdot \sum_{i=1}^{n} \varepsilon_i \cdot S_i.$$

By definition of $S_i = \frac{2 \cdot s_i}{\sum\limits_{k=1}^{n} s_k}$, we have

$$\sum_{i=1}^{n} S_i = 2 \cdot \frac{\sum\limits_{i=1}^{n} s_i}{\sum\limits_{k=1}^{n} s_k} = 2,$$

and

$$\sum_{i=1}^{n} \varepsilon_i \cdot S_i = 2 \cdot \frac{\sum\limits_{i=1}^{n} \varepsilon_i \cdot s_i}{\sum\limits_{k=1}^{n} s_k}.$$

Since $\sum\limits_{i=1}^{n} \varepsilon_i \cdot s_i = 0$, the second sum is 0, hence $\sum\limits_{i=1}^{n} p_i = 1$.

In both cases $\varepsilon_i = \pm 1$, we have $S_i \cdot p_i - p_i \cdot q_i = 0$, so all the constraints are indeed satisfied.

Vice versa, if the constraints are satisfied, this means that for every i, we have $p_i = q_i$ and $S_i \cdot p_i - p_i \cdot q_i = p_i \cdot (S_i - q_i) = p_i \cdot (S_i - p_i) = 0$, so $p_i = 0$ or $p_i = S_i$. Thus, the value p_i/S_i is equal to 0 or 1, hence the value $\varepsilon_i \stackrel{\text{def}}{=} 2 \cdot (p_i/S_i) - 1$ takes values -1 or 1. In terms of ε_i, we have $p_i/S_i = \frac{1+\varepsilon_i}{2}$, hence $p_i = \frac{1+\varepsilon_i}{2} \cdot S_i$. Since $\sum\limits_{i=1}^{n} p_i = 1$, we conclude that

$$\sum_{i=1}^{n} p_i = \frac{1}{2} \cdot \sum_{i=1}^{n} S_i + \frac{1}{2} \cdot \sum_{i=1}^{n} \varepsilon_i \cdot S_i = 1.$$

We know that $\frac{1}{2} \cdot \sum_{i=1}^{n} S_i = 1$, hence $\sum_{i=1}^{n} \varepsilon_i \cdot S_i = 0$. We know that this sum is proportional to $\sum_{i=1}^{n} \varepsilon_i \cdot s_i$, hence $\sum_{i=1}^{n} \varepsilon_i \cdot s_i = 0$ – i.e., the original instance of the partition problem has a solution.

The theorem is proven.

Comment. In our proof, we only used the case in which $n_1 = n_2$, i.e., in which there are exactly as many different values p_i as there are different values of q_j. Thus, we have proved, in effect, a stronger result: that even if we restrict the interval probability problem under independence constraint to such cases, we still get an NP-hard problem.

Let us clarify this observation. Informally, a problem $\mathcal{P}$ is NP-hard if every instance of any problem $\mathcal{Q}$ from the class NP can be effectively reduced to an instance of this problem $\mathcal{P}$. If a subproblem $\mathcal{P}'$ of the problem $\mathcal{P}$ is NP-hard, this means that we can reduce every instance of an NP-hard problem to an instance of this subproblem $\mathcal{P}'$ – and thus, to an instance of the problem $\mathcal{P}$. So, if a subproblem $\mathcal{P}'$ of the original problem $\mathcal{P}$ is NP-hard, then the problem $\mathcal{P}$ is NP-hard as well. In this particular case, since the subproblem $\mathcal{P}'$ formed by all instances with $n_1 = n_2$ is NP-hard, the original problem is NP-hard as well.

Informal Open Question

In the proofs of some of our theorems, we produced a natural interpretation for the constraints described in the proofs. For other proofs, the only constraints we could find are purely mathematical. It would be nice to come up with alternative proofs of these results – proofs based on more natural constraints.

Acknowledgments

This work was supported in part by NASA under cooperative agreement NCC5-209, NSF grant EAR-0225670, NIH grant 3T34GM008048-20S1, and Army Research Lab grant DATM-05-02-C-0046.

The authors are very thankful to the participants of the 4th International Symposium on Imprecise Probabilities and Their Applications ISIPTA'05 (Carnegie Mellon University, July 20–24, 2005), especially to Mikelis Bickis (University of Saskatchewan, Canada), Arthur P. Dempster (Harvard University), and Damjan Ŝkulj (University of Ljubljana, Slovenia), and to the anonymous referees for valuable discussions.

References

[1] D. Berleant, M.-P. Cheong, C. Chu, Y. Guan, A. Kamal, G. Sheblé, S. Ferson, and J. F. Peters, Dependable handling of uncertainty, *Reliable Computing*, 2003, Vol. 9, No. 6, pp. 407–418.

[2] D. Berleant, M. Dancre, J. Argaud, and G. Sheblé, Electric company portfolio optimization under interval stochastic dominance constrraints, In: F. G. Cozman, R. Nau, and T. Seidenfeld, *Proceedings of the 4th International Symposium on Imprecise Probabilities and Their Applications ISIPTA'05*, Pittsburgh, Pennsylvania, July 20–24, 2005, pp. 51–57.

[3] D. J. Berleant, O. Kosheleva, and H. T. Nguyen, "Adding Unimodality or Independence Makes Interval Probability Problems NP-Hard", *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems IPMU'06*, Paris, France, July 2–7, 2006 (to appear).

[4] D. Berleant, L. Xie, and J. Zhang, Statool: a tool for Distribution Envelope Determination (DEnv), an interval-based algorithm for arithmetic on random variables, *Reliable Computing*, 2003, Vol. 9, No. 2, pp. 91–108.

[5] D. Berleant and J. Zhang, Using Pearson correlation to improve envelopes around the distributions of functions, *Reliable Computing*, 2004, Vol. 10, No. 2, pp. 139–161.

[6] D. Berleant and J. Zhang, Representation and Problem Solving with the Distribution Envelope Determination (DEnv) Method, *Reliability Engineering and System Safety*, 2004, Vol., 85, No. 1–3.

[7] A. Borglin and H. Keiding, Stochastic dominance and conditional expectation an insurance theoretical approach, *The Geneva Papers on Risk and Insurance Theory*, 2002, Vol. 27, pp. 31-48.

[8] A. Chateauneuf, On the use of capacities in modeling uncertainty aversion and risk aversion, *Journal of Mathematical Economics*, 1991, Vol. 20, pp. 343-369.

[9] S. Ferson, *RAMAS Risk Calc 4.0*, CRC Press, Boca Raton, Florida.

[10] M. R. Garey and D. S. Johnson, *Computers and Intractability, a Guide to the Theory of NP-Completeness*, W.H. Freemand and Company, San Francisco, CA, 1979.

[11] J. Horowitz, C. F. Manski, M. Ponomareva, and J. Stoye, Computation of bounds on population parameters when the data are incomplete, *Reliable Computing*, 2003, Vol. 9, pp. 419–440.

[12] G. J. Klir, G. Xiang, and O. Kosheleva, "Estimating information amount under interval uncertainty: algorithmic solvability and computational complexity", *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems IPMU'06*, Paris, France, July 2–7, 2006 (to appear).

[13] V. Kreinovich and S. Ferson, Computing Best-Possible Bounds for the Distribution of a Sum of Several Variables is NP-Hard, *International Journal of Approximate Reasoning*, 2006, Vol. 41, pp. 331–342.

[14] V. Kreinovich, A. Lakeyev, J. Rohn, P. Kahl, *Computational complexity and feasibility of data processing and interval computations*, Kluwer, Dordrecht, 1997.

[15] V. Kreinovich, E. J. Pauwels, S. Ferson, and L. Ginzburg, A Feasible Algorithm for Locating Concave and Convex Zones of Interval Data and Its Use in Statistics-Based Clustering, *Numerical Algorithms*, 2004, Vol. 37, pp. 225–232.

[16] V. G. Krymsky, Computing Interval Estimates for Components of Statistical Information with Respect to Judgements on Probability Density Functions, In: J. Dongarra, K. Madsen, and J. Wasniewski (eds.), *PARA'04 Workshop on State-of-the-Art in Scientific Computing*, Springer Lecture Notes in Computer Science, 2005, Vol. 3732, pp. 151–160.

[17] V. Kuznetsov, *Interval Statistical Models*, Radio i Svyaz, Moscow, 1991 (in Russian).

[18] V. P. Kuznetsov, Interval methods for processing statistical characteristics, *Proceedings of the International Workshop on Applications of Interval Computations APIC'95*, El Paso, Texas, February 23–25, 1995 (a special supplement to the journal *Reliable Computing*), pp. 116–122.

[19] V. P. Kuznetsov, Auxiliary problems of statistical data processing: interval approach, *Proceedings of the International Workshop on Applications of Interval Computations APIC'95*, El Paso, Texas, February 23–25, 1995 (a special supplement to the journal *Reliable Computing*), pp. 123–129.

[20] C. F. Manski, *Partial Identification of Probability Distributions*, Springer Verlag, New York, 2003.

[21] S. Sivaganesan and J. O. Berger, Ranges of Posterior Measures for Priors with Unimodal Contaminations, *Annals of Statistics*, 1989, Vol. 17, No. 2, pp. 868–889

[22] D. Skulj, "Generalized conditioning in neighbourhood models", In: F. G. Cozman, R. Nau, and T. Seidenfeld, *Proceedings of the 4th International Symposium on Imprecise Probabilities and Their Applications ISIPTA'05*, Pittsburgh, Pennsylvania, July 20–24, 2005.

[23] D. Skulj, *A role of Jeffrey's rule of conditioning in neighborhood models*, to appear.

[24] K. Villaverde and V. Kreinovich, A linear-time algorithm that locates local extrema of a function of one variable from interval measurement results, *Interval Computations*, 1993, No. 4, pp. 176–194.

[25] K. Villaverde and V. Kreinovich, Parallel algorithm that locates local extrema of a function of one variable from interval measurement results, *Reliable Computing*, 1995, Supplement (Extended Abstracts of APIC'95: International Workshop on Applications of Interval Computations, El Paso, TX, Febr. 23–25, 1995), pp. 212–219.

[26] P. Walley, *Statistical Reasoning with Imprecise Probabilities*, Chapman & Hall, N.Y., 1991.

[27] J. Zhang and D. Berleant, Arithmetic on random variables: squeezing the envelopes with new joint distribution constraints, *Proceedings of the 4th International Symposium on Imprecise Probabilities and Their Applications ISIPTA'05*, Pittsburgh, Pennsylvania, July 20–24, 2005, pp. 416–422.