Finding Best Tuple via Interaction with Error-prone User Input

Qixu Chen, Raymond Chi-Wing Wong
The Hong Kong University of Science and Technology
qchenax@connect.ust.hk,raywong@cse.ust.hk

ABSTRACT

Given a large dataset containing a lot of points, in order to find the point which is the most interesting to a user efficiently, we could ask the user *interactively* a number of questions, each requiring the user to compare 2 points for choosing a more preferred point, and obtain his/her answers to learn his/her preference (represented by a utility function). With the utility function learnt, the most interesting point could be determined easily. In the database community, this interaction becomes more and more popular because it does not require the user to know his/her utility function and could return the most interesting point to the user, which overcomes the disadvantages of both traditional top-k queries (which requires to know the utility function in advance) and skyline queries (which could return a lot of possible points as an output). In the real world, the user may make mistakes (carelessly), which means that s/he may answer some of the questions *wrongly*. Unfortunately, existing interaction algorithms may find the *undesirable* point based on the *wrongly learnt* utility function because they assume that all answers from the user are 100% correct. In particular, even if the user answers only 1 wrong answer, the output of the existing algorithms may be far away from the users' real need. Motivated by this, in this paper, we propose a new problem of finding the most interesting point via interaction which is robust to possible mistakes made by a user. Besides, we propose (1) an algorithm that asks an asymptotically optimal number of questions when the dataset contains 2 dimensions and (2) two algorithms with provable performance guarantee when the dataset contains d dimensions where $d \geq 2$. Experiments on real and synthetic datasets show that our algorithms outperform the existing ones with a higher accuracy with only a small number of questions asked.

1 INTRODUCTION

A database system may contain millions of points (or tuples). In order to help the user to find his/her interesting point, we need queries to obtain a representative set of points consisting of his/her interesting point. Such queries can be considered as *multi-criteria decision making* problems and they can be applied in various domains, including house buying, car purchase and job search. For example, in a car purchasing database where each car is described by some attributes, Alice wants to find a car with a low gasoline consumption and a high speed, and is as cheap as possible. Here, gasoline consumption, speed and price are some attributes that Alice would consider when buying a car.

There are two popular types of traditional multi-criteria decision making queries, namely *top-k* and *skyline* [5]. The top-k query measures the *utility* of tuples based on a *utility function* provided by the user. A high utility indicates that the corresponding tuple is more preferred. The query returns k tuples with the highest utility in the dataset. Top-k requires the knowledge of the user's exact utility function. On the other hand, the skyline query does not need this information and uses a *"dominant"* concept. Specifically, a tuple p is said to *dominate* a tuple q if p is not worse than q in any attribute and is better than q in at least one attribute. Intuitively, p is better than q w.r.t. all *monotonic* utility functions. The skyline query returns the set of tuples that are not dominated by any tuple in the dataset. Unfortunately, the size of the returned set is uncontrollable and could be as large as the database size in the worst case.

Motivated by this, a novel interactive framework [19, 27, 29] was proposed to overcome the disadvantages of both the top-k query (requiring a given utility function) and the skyline query (returning an output with an uncontrollable size). Intuitively, it asks the user a number of *rounds* of simple questions and returns the most interesting tuple to the user. The interactive system not only does not require the user to provide an exact utility function, but also can control the size of the returned set (i.e., the only one tuple returned). Therefore, it does not have the limitation of both top-k and skyline queries. A widely applied form of question [19, 27, 29] is to display 2 points in each round, and the user is asked to select the preferred point. Consider the car purchasing scenario. The interactive framework simulates a sales assistant that asks Alice to indicate her preference among several pairs of cars, and make recommendations based on the answers of Alice.

However, in real world, when answering questions, people make occasional mistakes/errors due to various reasons. For example, in a simple task of selecting the correct switch among two given switches that are dissimilar in shape, a human knowing which switch is correct can still select the wrong one with probability 0.1% [14]. For operations that require some care, the error probability can range from 1% to 3%, which may even increase under high mental stress. Simple calculation shows that if the system interacts with the user for 10 rounds (i.e., asks the user a sequence of 10 questions), a user has at least 10%-30% chance to make at least one mistake. For example, Alice may indicate her preference to a cheap car with high gas consumption among some similar cars due to a careless mistake, even though she intends to buy a low-consumption car. Note that in the real world, a meticulous sales assistant will notice the inconsistency and check with Alice.

It is worth mentioning that making a "small" mistake can lead to unforgettable and unchangeable consequences. Let us give two real cases. The first case is about the selection of the tertiary school, one of the critical milestones of one's life. In 2020, an 18-year-old student in Guangdong province in Mainland China who obtained a top-tier score from the National College Entrance Examination in China (also called gaokao) was admitted to a mediocre college with a similar name to his target university due to his mistake of choosing a wrong school in the tertiary school selection system [28]. The second case is about a huge financial loss due to a well-known "fat finger error" (which refers to an error made by the operator in the trading system when making a wrong deal in the trading system by mis-clicking or pressing a wrong key). In 2018, Samsung

Securities made a wrong transaction worth 100 billion dollars due to a fat finger error, which could incur a loss of 428 million dollars, 12.17% of the company's market capitalization [1].

Motivated by this, in this paper, we propose a new problem called the *interactive best point retrieval problem* considering error-prone user input, which is more realistic. Roughly speaking, our problem is to find the best point in a dataset D for a user with an *unknown* utility vector u, in the scenario that the user is "imperfect" and makes a random error with probability at most θ for each question requiring a user to select one preferred point among 2 points displayed, where θ is called an *error rate* and is a user parameter. In other words, the user chooses the preferred point with probability at least $1-\theta$ and chooses the other point with probability at most θ. Note that θ can be obtained from some channels like user behavior studies [14]. In our user study, θ is found to be 4.5%.

Unfortunately, most (if not all) existing interaction algorithms [19, 27, 29] without considering user errors may find the *undesirable* point based on the *wrongly learnt* utility function since they assume that the user never makes mistake. In our experiment, these algorithms return incorrect results. For example, on a dataset with 1M points, the accuracies of all closely related algorithms *UtilityApprox* [19], *UH-Simplex* [29] and *HDPI* [27] are at most 74% only.

Furthermore, all adapted versions of existing interactive algorithms considering user errors [12, 22] for this problem do not perform well. In our experiment, the accuracies of the adapted versions of *Active-Ranking* [12] and *Preference-Learning* [22] on a dataset with size 1M are at most 67% only, which is not acceptable.

Contributions. We summarize our contributions as follows. Firstly, we are the first to propose the best point retrieval problem considering random interaction errors during user interaction. Secondly, we prove a lower bound on the expected number of questions needed to determine the best point with a desired confidence threshold. Thirdly, we propose (1) an algorithm with an asymptotically optimal number of questions asked when the dataset contains two dimensions and (2) two solutions with provable guarantee in terms of both the number of questions asked and the confidence on finding the best point when the dataset contains d dimensions where $d \geq 2$. Fourthly, we conducted comprehensive experiments to demonstrate the superiority of the proposed methods. The results show that our algorithms maintain a high accuracy (e.g., nearly to 100% in most experiments) in finding the best point using only a small number of questions, but existing approaches either ask too many questions (e.g., twice as many as ours), or are much inaccurate (e.g., more than 10% less accuracy than ours).

Organizations. The rest of this paper is organized as follows: Section 2 gives our problem definition. Section 3 shows the related work. We introduce the algorithm for the dataset containing two dimensions in Section 4 and two algorithms for the dataset containing at least two dimensions in Section 5. In Section 6, we present the experimental results. Section 7 concludes the paper.

2 PROBLEM DEFINITION

In this section, we provide a formal definition to our problem. Firstly, in Section 2.1, we introduce some basic terminologies. Then, we formally define the random user error setting in Section 2.2. In Section 2.3, we study the lower bound of the number of questions to return the best point with a desired confidence. The input of our problem is a dataset D in a d-dimensional space. Note that each tuple in D could be described by more than d attributes, but the user is interested in exactly d of them.

2.1 Terminologies

In this paper, we use the word "tuple" and "point" interchangeably. We denote the i-th dimensional value of a point $p \in D$ by $p[i]$ where $i \in [1, d]$. Without loss of generality, the value of each dimension is normalized in range $[0, 1]$ and for each $i \in [1, d]$, there exist at least one point $p \in D$ such that $p[i] = 1$. We assume that a larger value in each dimension is more preferable to the user. If a smaller value is preferred for an attribute (e.g., price), we can modify the dimension by subtracting each value from 1 so that it satisfies the above assumption. Consider the 2-dimensional example in Table 1. We have a database $D = \{p_1, p_2, p_3, p_4, p_5, p_6, p_7\}$ and we are interested in attribute X_1 and X_2.

As widely applied in [19, 22, 27, 29], the user's preference is modeled as an unknown *linear utility function*. Specifically, we model the utility function $f : \mathbb{R}_+^d \to \mathbb{R}_+$ as a linear function $f(p) = u \cdot p$, where u is a non-negative d-dimensional real vector and $u[i]$ measures the importance of the i-th attribute. We call $f(p)$ the *utility* of p w.r.t. f and u is called the *utility vector*. In the rest of this paper, we also refer f by its utility vector u. We are interested in finding the point with the highest utility, which is the point $p_h = argmax_{p \in D}\, u \cdot p$. Note that scaling u does not change the rank of points in D and thus, does not change the best point. Therefore, without loss of generality, we assume that $\sum_{i=1}^{d} u[i] = 1$.

2.2 Handling Random Errors

The system interacts with a user for several *rounds*, until a stopping condition is satisfied. The user is asked 1 question in each round and we use the term "question" and "round" interchangeably in the rest of this paper. We adopt a popular strategy of asking questions in the literature [12, 15, 17, 22] that in each round, the system displays a pair of points, namely p_i and p_j, to the user, and the user returns the preferred point between these 2 points. Instead of assuming the user always makes correct choices, the user makes a random error with probability at most θ in each round. Specifically, let p^* denote the point with a higher utility in two points p_i and p_j, the user chooses p^* with probability at least $1-\theta$, and, with probability at most θ, s/he selects the other point. Typically, θ should not be too large since it is a careless mistake. Thus, it is natural to assume that θ is smaller than 0.5 as supported by existing statistics about the error rate described in Section 1. If θ is greater than 0.5, there is no hope that we could find the best point for this user (because the user already gives more than half of his/her answers wrongly.)

2.3 Lower Bound

We are interested in the following problem: Given an input size n, and given that the user makes an error with probability at most θ in each round, how many questions do we expect to ask to obtain the best point? [29] proves that for any dimensionality d, there exists a dataset of size n such that any comparison-based interactive algorithm must ask $\Omega(\log_2 n)$ questions to determine the best point. However, this study does not consider user errors and cannot

p	X_1	X_2	$f(p)(u = (0.3, 0.7))$
p_1	0.2	1	0.76
p_2	0.5	0.9	0.78
p_3	0.8	0.7	0.73
p_4	0.9	0.6	0.69
p_5	1	0.2	0.44
p_6	0.6	0.8	0.74
p_7	0.8	0.5	0.59

Table 1: Dataset and utility

$h_{i,j}$	$h^+_{i,j}$	$h^-_{i,j}$	intersects
$h_{1,2}$	$\{P_1\}$	$\{P_2, P_3\}$	$\{P_4, P_5\}$
$h_{1,4}$	$\{P_1, P_3, P_5\}$	$\{P_4\}$	$\{P_2\}$
$h_{2,4}$	$\{P_2, P_3, P_5\}$	$\{P_4\}$	$\{P_1\}$
$h_{2,5}$	$\{P_1, P_2, P_4\}$	$\{P_3, P_5\}$	$\{\}$
$h_{3,5}$	$\{P_3\}$	$\{P_1, P_2, P_4, P_5\}$	$\{\}$

Table 2: Table L

$h_{i,j}$	$h^+_{i,j}$	$h^-_{i,j}$	intersects
$h_{1,2}$	$\{P_1\}$	$\{P_2\}$	$\{P_4\}$
$h_{1,4}$	$\{P_1\}$	$\{P_4\}$	$\{P_2\}$
$h_{2,4}$	$\{P_2\}$	$\{P_4\}$	$\{P_1\}$

Table 3: Table L after selecting $h_{2,5}$

be applied to our problem. Therefore, we present the following theorem about the lower bound.

Theorem 1. *For any dimensionality d, given an error rate θ and a confidence parameter δ, there exists a dataset of n points such that any pairwise comparison-based interactive algorithm needs to ask $\Omega(\frac{4\theta^2(1-\theta)}{(1-2\theta)^2}(\log n)\log(\frac{\log n}{\delta}))$ rounds on expectation to determine the best point with confidence at least $1-\delta$.*

Proof sketch. We first show that in order to find the best point, $\Omega(\log n)$ questions must be asked and the expected number of errors we must handle is $\mu = \Omega(\theta \log n)$. We then prove that to let the total failure probability be less than δ, at least $k = \Omega(\frac{4\theta(1-\theta)}{(1-2\theta)^2}\log(\frac{\log n}{\delta}))$ additional questions must be asked to handle each error. The total number of rounds is therefore $\Omega(\log n + \mu k) = \Omega(\frac{4\theta^2(1-\theta)}{(1-2\theta)^2}(\log n)\log(\frac{\log n}{\delta}))$. □

3 RELATED WORK

Besides the traditional top-k and skyline queries mentioned in Section 1, various types of multi-criteria decision making queries were proposed. Two types of queries that does not require user interaction are the similarity query [2, 25] and the regret minimizing query [8, 20, 21]. The similarity query looks for tuples that are similar to a given query tuple, where the similarity is measured by a given distance function. However, the query tuple and the distance function must be provided in advance [3], which may be an unrealistic assumption in practice. On the other hand, the regret minimizing query [8, 20, 21] defines a measurement called *regret ratio*, and returns a set of tuples that minimizes the regret ratio of the user. Although regret minimizing query is not as demanding as top-k and similarity query, it is hard to achieve both a small regret ratio and a small size of return set. When a small regret ratio is fixed, the output size is usually large [8, 29].

To overcome the limitations of the above queries, some existing studies [3, 12, 19, 22, 24, 27, 29] proposed to involve user interactions. The form of interaction varies. A form of interaction that is widely adopted in [12, 19, 22, 27, 29] is to ask the user to select the favorite point among a set of displayed points. [19] proposed the interactive regret minimizing query, which aims to lower the regret ratio while keeping the output size small. [29] follows the study on regret minimization and proposed two algorithms, namely *UH-Simplex* and *UH-Random*, that only displays real tuples inside the dataset. [27] proposed interactive algorithms, *HD-PI* and *RH*, that target at searching for one of the top-k point in the database. However, all of these algorithms assume that the user never makes mistakes and completely prune some points from further consideration, making it not applicable to adapt them to handle user errors.

Besides asking a user to select one point from a set of points, there are studies focusing on other types of interactions. [30] developed algorithm *Sort-Simplex* which asks the user to give a ranking on the displayed points. [18] asks the user to partition points into *superior* and *inferior* groups to learn his/her preference. [4] proposed the interactive similarity query that learns the distance function and the query tuple via user interaction. However, it requires a user to assign *relevance scores* to hundreds of tuples to locate the query tuple. In a user's perspective, these interactions are too demanding and may affect their willingness to interact with the system. Besides, user errors are not considered in these studies.

The robustness issue is also studied in the Machine Learning (ML) and Information Retrieval (IR) literature [10, 12, 13, 22]. But, one major difference between their work and ours is that most (if not all) of them focus on a given *static* dataset without user interaction involved, but in our case, the data is created *dynamically* during user interaction. It is worth mentioning that for many algorithms in the ML/IR field, the data required to return an accurate result can be more than 10^3 [10, 13], which means that if we directly adapt them to our problem, the user needs to answer thousands of questions. [22] proposed *Preference-Learning* to learn user's preference that copes user errors by introducing a slack variant in a linear SVM. [12] proposed *Active-Ranking* and resolves possible conflict using the majority vote. Although these algorithms can be *adapted* to find the best point considering user errors, they are not efficient enough and tend to ask many more questions.

Compared with the existing studies, our work has the following advantages. Firstly, we do not require the user to provide an exact utility function or query tuple, as in the top-k query or the similarity query. Secondly, we reduce the user's effort by asking fewer questions and only asking the user to select one point (i.e., the most desirable point) in each question, while some existing algorithms ask too many questions (e.g., [12, 22]) and some other algorithms (e.g., [3, 18, 30]) ask too difficult questions. Finally, we allow the user to make unavoidable errors in interaction, which discards the unrealistic error-free assumption made in some existing studies. Even under the user error setting, we guarantee the retrieval of the best point with a desired confidence level.

4 TWO-DIMENSIONAL ALGORITHM

In this section we focus on the case where $d = 2$ and present the *2-dimensional Robust Interactive* (*2RI*) algorithm. We first introduce some important concepts in Section 4.1. Then, in Section 4.2, we

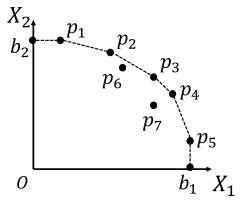


Figure 1: 2D convex hull

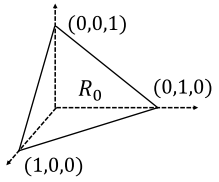


Figure 2: Utility space in 3D

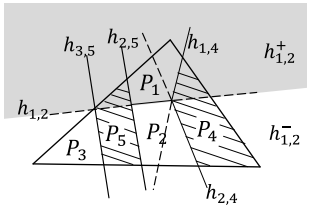


Figure 3: Partitions

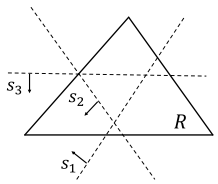


Figure 4: Halfspaces

Algorithm 1 Check(p_i, p_j, k)

```
select_p_i ← 0, select_p_j ← 0
while select_p_i < ⌈k/2⌉ and select_p_j < ⌈k/2⌉ do
    display pair (p_i, p_j) to the user
    if p_i is chosen by the user then
        select_p_i ← select_p_i + 1
    else
        select_p_j ← select_p_j + 1
return p_i if (select_p_i ≥ ⌈k/2⌉) and p_j otherwise.
```

introduce a useful checking scheme that will be used in later algorithms. Finally, we show the details of *2RI* in Section 4.3. This algorithm uses a number of rounds that is asymptotically equal to the lower bound proved in Secion 2.3 .

4.1 Preliminaries

In geometry, the *convex hull* of a dataset D, denoted by $CH(D)$, is the smallest convex set containing all points in D [9]. A point $p \in D$ is a *vertex* of $CH(D)$ if $p \notin CH(D/\{p\})$. We use $conv(D)$ to denote the set of *vertices* of $CH(D)$. Let b_i denote the point with the i-th coordinate being 1 and all other coordinates being 0. Also, denote $B = \{b_i | 1 \leq i \leq d\}$ and denote the origin as O. Consider the set $D \cup B \cup \{O\}$. In the remaining sections, when we say $conv(D)$, we mean the set of points that are both in D and in $conv(D \cup B \cup \{O\})$. We assume that there are n points in $conv(D)$, namely $p_1, p_2, \ldots, p_n$, in a clockwise order. Consider the dataset D in Table 1 and its corresponding $conv(D)$ visualized in Figure 1. In this example, $conv(D) = \{p_1, p_2, p_3, p_4, p_5\}$.

One important conclusion is that the best point must be in $conv(D)$, and thus, we need only look at points in $conv(D)$. This is because for any $p \notin conv(D)$ and any utility vector u, there must exist a point $p^* \in conv(D)$ s.t. $u \cdot p \leq u \cdot p^*$. This conclusion can also be applied to any dimensions $d \geq 2$. In the rest of this paper, unless explicitly stated, we will assume that the input to our algorithm is $conv(D)$ and we use n to denote the size of $conv(D)$.

4.2 Checking Subroutine

Before we start to introduce our algorithm, we first introduce a checking subroutine, which will be applied later in our algorithms. Intuitively, since the preference indicated by the user between two points, p_i and p_j, is incorrect with probability at most θ, we need to devise a way of "checking" to increase the confidence level of the results obtained. The details of the subroutine are shown in Algorithm 1. In short, what it does is to check the relation between two points, namely p_i and p_j, for at most k times, where k is a user parameter, and return whether p_i is more preferred to p_j based on the majority vote. k is set to an odd number to break ties. In case that we want to avoid asking repetitive questions involving same points, we could use common techniques in the field of questionnaire reliability (formally called *intra-rater reliability* [11, 23]). One example is that we could re-scale the two points in each question by multiplying with a random number between 0.95 and 1 so that these 2 new points are shown to the user and look different from the 2 original points, resulting in a question which looks different from the original question. Another example is asking correlated questions which look differently using statistical methods.

In our algorithm, before invoking the checking subroutine on points p_i and p_j, (i.e., $Check(p_i, p_j, k)$), the 2 points may have already been presented to the user once. This can also be considered as a round of checking (since the user also provided his/her preference between p_i and p_j), so instead of initializing both $select\_p_i$ and $select\_p_j$ to 0, we increment the corresponding counter by 1. For example, if the user previously indicated that p_i is preferred to p_j, then $select\_p_i$ is initialized to 1. By doing so, we can ask one fewer question for the checking subroutine, meanwhile not hurting the correctness of the algorithm.

Since the user makes random errors with probability at most θ, even this checking subroutine can fail to reveal the real relationship between p_i and p_j. Without loss of generality, assume that p_i is preferred to p_j. Let P_k denote the probability that a checking with k questions asked fails, that is, $P_k = P(Check(p_i, p_j, k) = p_j)$. Then, P_k can be computed as follows: $P_k \leq 1 - \sum_{t=0}^{\lfloor \frac{k}{2} \rfloor} \binom{k}{t} \theta^t (1-\theta)^{k-t}$.

One can observe that P_k monotonically decreases when k increases. A natural question would be how to efficiently determine the value of k if we want $P_k \leq \alpha$ where α is a desired failure probability. When α is relatively small (e.g., 0.00001), P_k can be approximated using a Gaussian function [26] and Lemma 1 can be used to compute the value of k. On the other hand, when α is relatively large (e.g., 0.01), the Gaussian approximation is not very tight, and thus we can apply a binary search on k to find the smallest value such that $P_k \leq \alpha$.

Lemma 1. *Given a user error rate θ and a desired failure probability α, the checking subroutine fails with probability at most α by setting $k = \Theta(\frac{4\theta(1-\theta)}{(1-2\theta)^2} \log(\frac{1}{\alpha}))$.*

A naive solution of the problem is then to apply this checking subroutine for every question asked to ensure their correctness. However, this solution incurs too many unnecessary questions. We seek ways of using this subroutine as few as possible and only invoke this subroutine when necessary.

4.3 2RI

We are now ready to present our 2-d algorithm *2RI*. The input is a list of points $\{p_1, \ldots, p_n\}$, sorted in a clockwise order. For the ease of illustration, we label the points from 1 to n. Intuitively, algorithm *2RI* performs a search on this sorted list by initializing a search range on this sorted list, denoted by $[L, U]$, where L and U are set to 1 and n initially, updating variables L and U to shrink/expand the search range based on the user's interaction and returning the point in the search range as an output when the search range contains only one element.

Before we describe algorithm *2RI*, we need to describe 3 concepts. The first concept is the increasing-and-decreasing trend of the utility values over the sorted list as shown in the following lemma.

Lemma 2. *Given the utility vector u and n points $p_1, p_2, \ldots, p_n$ in $conv(D)$ ordered in a clockwise direction. Let p_h be the point with the highest utility score, that is, $p_h = \arg\max_{p \in conv(D)} u \cdot p$. Then $\forall 1 \leq i < j \leq h, u \cdot p_i < u \cdot p_j$. Besides, $\forall h \leq i < j \leq n, u \cdot p_i > u \cdot p_j$.*

Proof. We first prove that $\forall 1 \leq i < j \leq h, u \cdot p_i < u \cdot p_j$. The other half can be proved symmetrically. Since p_h is the max-utility point, we have $u \cdot p_h - u \cdot p_{h-1} = u \cdot (p_h - p_{h-1}) \geq 0$. By the nature of convex hull in $2d$, for all i, j where $0 \leq i < j \leq h$, we must also have $u \cdot (p_j - p_i) = u \cdot p_j - u \cdot p_i > 0$, which concludes the proof. □

The second concept is the size of a search range. Given a range $R = [L, U]$, we define the *size* of the range R, denoted as m, to be $m = U - L + 1$.

The third concept is related to the details of an operation to be used in algorithm *2RI*. This operation is the procedure of how we could determine whether the desired point p_h (which is unknown and is to be found) is in the search range $[L', U']$ where $h \in [1, n]$, $L' \in [1, n]$ and $U' \in [1, n]$. Without loss of generality, we show how to decide if the lower bound is correct, that is, if $L' \leq h$. The case of the upper bound is then symmetric. To test the left boundary, we can simply check if $p_{L'}$ is preferred to $p_{L'-1}$ (If $p_{L'-1}$ does not exist, then the lower bound is trivially correct). According to Lemma 2, if $p_{L'}$ is preferred, then we can conclude that $L' \leq h$. Otherwise, the lower bound is wrong and we must adjust L'. In conclusion, checking the left (right) boundary requires to call the checking subroutine *Check*$(p_{L'}, p_{L'-1}, k)$ (*Check*$(p_{U'}, p_{U'+1}, k)$.).

We are ready to describe algorithm *2RI*. Initially, the search range, denoted by $[L, U]$, is initialized such that variable L is set to 1 and variable U is set to n. In the following, variables L and U are updated during the execution of the algorithm which needs interaction from the user. Specifically, we perform the following iterative process until $U = L$ (i.e., there is only one element in the search range). When $U = L$, we return p_L as the output point of this algorithm.

Specifically, the interactive process involves a number of iterations. Each iteration has the following two steps.

- **Step 1 (Search Range Shrinking):** We initialize variables L' and U' to be L and U, respectively. We initialize variable m to be the size of the initial search range (i.e., $U - L + 1$). Due to the increasing-decreasing trend of the utility value as shown in Lemma 2, we could perform a binary search on range $R' = [L', U']$ by updating variables L' and U' until the size of the updated search range $R' = [L', U']$ is at most $\lceil 2m\theta \rceil$. The reason why we choose this maximum size as $\lceil 2m\theta \rceil$ can be found in Lemma 4 in the Appendix. Here, each step involved in a binary search corresponds to the operation of asking the user to select a more preferred point among the two displayed points where these two points are p_r and p_{r+1} where $r = \lceil \frac{U'+L'}{2} \rceil$. Depending on the user's answer, by Lemma 2, the search range could be shrunk accordingly. This technique is similar to [19, 29] and details could be found therein. It is worth mentioning that we did not call any checking routine described before in this step.
- **Step 2 (Search Range Verification and Correction):** It performs the checking subroutine to verify if the desired point p_h is still inside range $[L', U']$ (i.e., R'). If yes, the algorithm proceeds to the next iteration with the confirmed range R' by updating variable L to be L' and variable U to be U'. However, if one of L' and U' is not correct, we need to perform some additional checking subroutine and update variables L' and U' accordingly. Note that either the lower bound L' or the upper bound U' is wrong, but not both. Without loss of generality, we assume that L' is wrong (i.e., $L' > h$). In the following, we want to update the search range such that both the upper bound and the lower bound of the updated search range is smaller than L'. At the same time, we also try to keep the size of the updated search range at most $\lceil 2m\theta \rceil$ w.h.p.. There are two cases where the first case is a general case and the second case is a boundary case.
  - Case (1) (i.e., $L' - \lceil 2m\theta \rceil > L$) In this case, the algorithm performs the checking subroutine again to decide whether p_h is inside range $[L' - \lceil 2m\theta \rceil, L' - 1]$. If this is the case, it updates variable L' to be $L' - \lceil 2m\theta \rceil$ and variable U' to be $L' - 1$, which essentially "shifts" the search range to this new range (i.e., $[L' - \lceil 2m\theta \rceil, L' - 1]$). Note that this new range also has size at most $\lceil 2m\theta \rceil$ and is exactly just on the left-hand-side of the original search range over the sorted list. But, if p_h is still not in this range (i.e., h is smaller than $L' - \lceil 2m\theta \rceil$), the algorithm updates variable L' to be L (i.e., the initial content of L just at the beginning of the iteration) and variable U' to be $L' - \lceil 2m\theta \rceil - 1$. Note that this new search range is on the left-hand-side of the original search range over the sorted list. In this case, it is possible that the size of this new search range could be greater than $\lceil 2m\theta \rceil$ but the chance is very low (which could be explained by Lemma 4 (in the Appendix)).
  - Case (2) (i.e., $L' - \lceil 2m\theta \rceil \leq L$) In this case, the algorithm directly sets variable L' to be L and variable U' to $L' - 1$, without performing any checking subroutine. Note that this new range also has size at most $\lceil 2m\theta \rceil$ and is exactly just on the left-hand-side of the original search range over the sorted list.

  The case where U' is wrong (i.e., $U' < h$) can also be addressed symmetrically. Finally, it sets variable L to be L', variable U to be U' and enters the next iteration.

Consider the running example as shown in Figure 1, where the input is $\{p_1, p_2, p_3, p_4, p_5\}$. Assume that $\lceil 2m\theta \rceil = 2$, $L = 1$ and $U = 5$, and by performing a binary search, we obtain $L' = 4$ and $U' = 5$. Clearly, the user made some error because p_h (= p_2 in this case) is not in the current search interval. By checking pair (p_3, p_4), the algorithm finds out that p_h is on the left of L' (i.e., $h < 4$), it then

algorithm	Number of Rounds
Verify-Point	$O(\frac{c}{1-\theta}\log n(conj+\log\log n+\log\frac{1}{\delta}))$
Verify-Space	$O(\frac{d}{1-2\theta}\log n(conj+\log\log n+\log\frac{1}{\delta}))$

Table 4: Performance of Multi-dimensional Algorithms

checks p_1, p_2 and since p_2 is more preferred to p_1, the algorithm shifts the search interval to $[2, 3]$, and enters a new iteration.

Theorem 2 presents the main results on *2RI*. Corollary 1 shows that *2RI* is asymptotically optimal.

Theorem 2. *Given an input size n, an error rate θ and a failure probability δ, 2RI finds the best point with probability at least $1-\delta$ using $O(\frac{1}{-\log 2\theta}(\log n)\log\frac{\log n}{\delta})$ rounds on expectation.*

Proof sketch. We first prove that, the expected number of iterations is $O(\log_{\frac{1}{2\theta}} n)$. Since each iteration contains $O(\log\frac{1}{2\theta}+k)$ rounds, the total number of rounds required is $O(\log n + k\log_{\frac{1}{2\theta}} n)$ with total success probability at least $1-O(P_k\log_{\frac{1}{2\theta}} n)$. Setting $\delta = O(P_k\log_{\frac{1}{2\theta}} n)$ and applying Lemma 1 yields the theorem. □

Corollary 1. *2RI is asymptotically optimal.*

Proof. Since θ is fixed, the time complexity of 2RI is asymptotically equal to the lower bound in Theorem 1. □

5 MULTI-DIMENSIONAL ALGORITHM

Although the previous algorithm works well when $d = 2$, it cannot be adapted to a higher dimensional situation due to the change of nature of the convex hull in high dimensions. A possible way of adapting it to a high-dimensional space is to follow the idea presented in [19] and estimate the utility value of each dimension one by one through constructing artificial points. This adaption, however, will be ineffective in a high dimension space since it cannot determine if the estimation on the utility value of each dimension is accurate enough, resulting in asking unnecessary questions. In this section, we present two algorithms, namely *Verify-Point* and *Verify-Space*, that can be applied to databases with the dimensionality $d \geq 2$. The complexities of the number of questions asked by these two algorithms (also called *round complexities*) are summarized in Table 4. Here, n is the input size, δ is the failure probability of retrieving the best point, $conj$ is the round complexity of *Conjecture Phase* of each of the two algorithms (to be introduced later), d is the dimensionality, and c is a data-dependent variable, which can be regarded as a small constant according to our experimental results. In our experimental results, c is at most 3. Both algorithms enjoy provable theoretical guarantee in terms of the success probability of best point retrieval and round complexity, and differ in a sense that *Verify-Space* is less dependent on the distribution of data. Both algorithms follow a 2-phase framework, and their first phase are similar. Therefore, we arrange the following sections as follows. We first introduce some preliminaries and the general algorithm framework in Section 5.1, and then introduce the first phase, called *Conjecture Phase*, of the two algorithms in Section 5.2. We then describe the second phase, called *Verification Phase*, of *Verify-Point* and *Verify-Space* in Section 5.3.1 and 5.3.2 respectively.

5.1 Preliminaries

Recall that in a d-dimensional database, the utility vector u is a d-dimensional non-negative real vector and $\sum_{i=1}^{d} u[i] = 1$. Therefore, the collection of all possible utility vectors, called the *utility space* [9] and denoted as R_0, is a $(d-1)$-dimensional polytope. For example, as shown in Figure 2, in a 3-dimensional dataset, the utility space is a triangle with vertices $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$.

For any two points, namely p_i and p_j, in $conv(D)$, we can construct a hyperplane $h_{i,j}$ that passes through the origin with normal $p_i - p_j$. $h_{i,j}$ intersects the utility space and divides it into two *halfspaces* [9]. The halfspace above (resp. below) $h_{i,j}$ is denoted as $h^+_{i,j}$ (resp. $h^-_{i,j}$), and contains all the utility vectors that ranks p_i higher (resp. lower) than p_j, or equivalently, $u \cdot p_i > u \cdot p_j$ (resp. $u \cdot p_i < u \cdot p_j$) where u is the utility vector. When the user chooses from the two displayed points, namely p_i and p_j, s/he will indicate in which halfspace her/his utility vector lies. For the ease of illustration, we use $s_{i,j}$ to denote the halfspace chosen by the user that is bounded by $h_{i,j}$. Note that based on the user's preference between p_i and p_j, $s_{i,j}$ may be either $h^+_{i,j}$ or $h^-_{i,j}$. We also denote the counterpart of $s_{i,j}$ as $s^-_{i,j}$.

For a convex polytope P, we denote the set of its vertices by V_P. The *utility range* R [29], which is defined to be the convex region that contains the true utility vector u, can be determined as follows. Initially, R is the entire utility space (i.e., R_0). When a new halfspace (e.g., $s_{i,j}$) is provided by the user's answer to the comparison between p_i and p_j, it means that $u \in s_{i,j}$ so we update R to $R \cap s_{i,j}$. Since R can be represented by the intersection of halfspaces, it is a convex polytope. Many existing studies [12, 27, 29] applied this framework and strategically choose the pair of points to reduce the number of questions asked. However, since in our problem setting, each halfspace has probability at most θ to be wrong, the utility range found by these algorithms may deviate from the real utility vector and their performance will degenerate when a user makes mistakes. To alleviate the effect of user errors, we present two algorithms that have higher error-tolerance.

The general framework of the algorithms works as follows. It runs for several *iterations* and each iteration consists of two phases, namely Phase 1, called *Conjecture Phase*, and Phase 2, called *Verification Phase*. The input to Conjecture Phase is a utility range R' indicating a convex region where the utility vector u lies, and a set C' storing some possible best points. The goal of Conjecture Phase is to "pretend" that there is no user error and interact with the user for several rounds until some *best point candidate* $p_c \in C'$ is found. During the execution of Conjecture Phase, we also maintain a set S storing all halfspaces indicated by the user. Later, in Verification Phase, we will selectively check the correctness of some halfspaces in S so that if p_c is not the *real* best point, we will still have a chance to remedy the mistake. If all the decisions made in Conjecture Phase and Verification Phase are correct, then we can conclude that p_c is the best point with high confidence. However, if some questions in Conjecture Phase are answered incorrectly, we may "waste" some rounds and focus on the wrong region of the utility space that does not contain the utility vector u. As a consequence, more than one point *can* still be the best point and at least one additional iteration is needed. In this case, with the shrunken size of the set of points under consideration, the algorithm returns back to Conjecture Phase.

It alternates between Conjecture Phase and Verification Phase until all but one point is pruned out for further consideration. Then, we return this point as the final answer.

5.2 Conjecture Phase

Given a utility range R' indicating a convex region where the utility vector u lies, and a set C' containing some possible best points, the purpose of Conjecture Phase is to locate the user's best point candidate with the least number of rounds, without considering possible interaction errors. Note that since the only information required by Verification Phase from Conjecture Phase is the user's answer on each question, or in other words, the set S containing all halfspaces indicated by the user, the design of Conjecture Phase is quite flexible. In general, it could be any comparison-based interactive algorithms considering no user errors. Here, we adapt from the algorithm *HD-PI* [27] since it has the highest round-efficiency in finding the best point candidate.

At the beginning of Conjecture Phase, S is initialized to an empty set. In each round of Conjecture Phase, the algorithm selects a pair of points, namely p_i and p_j, and asks the user to choose the preferred point. The user's answer indicates the halfspace $s_{i,j}$ where the utility vector lies, so the algorithm updates R' to $R' \cap s_{i,j}$ and inserts $s_{i,j}$ into S. It is worth mentioning that when the user makes an error, u will not lie in $s_{i,j}$. However, in Conjecture Phase, the algorithm "pretend" that the user has no error, and the potential error will be handled later by Verification Phase. After Conjecture Phase ends, the set S is passed to Verification Phase.

The round complexity of Conjecture Phase varies from different implementations. Here, we denote the complexity of Conjecture Phase as $O(conj)$. When adapting *HD-PI* as Conjecture Phase, $O(conj)$ varies from $O(\log n')$ to $O(n')$ depending on the data distribution, where n' is the size of set C'. Even though the worst case has a linear complexity, in our experiment, Conjecture Phase is very efficient.

5.3 Verification Phase

When there is no user error, p_c found in Conjecture Phase would be the true best point. However, since the user may make mistakes, the result from Conjecture Phase may no longer be the true best point. Therefore, Verification Phase is applied to make sure that the true best point is not pruned for further consideration with high probability. In the following 2 sections, we introduce Verification Phases of *Verify-Point* and *Verify-Space* respectively.

5.3.1 Verification Phase of Verify-Point. During Conjecture Phase, the set of halfspaces selected by the user is stored in set S. Observe that for each halfspace $s \in S$, s contains the utility vector u with probability at least $1-\theta$ and its counterpart s^- contains u with probability at most θ (since a user makes a mistake with probability at most θ). Based on this observation, it would be more efficient to *verify* the correctness of those halfspaces that are expected to help eliminate many points from further consideration. Therefore, we develop the algorithm called *Verify-Point*.

Before we present *Verify-Point*, we first introduce some preliminaries and data structures that will be used in this algorithm. Recall that the input of *Verify-Point* is $conv(D)$. For each point $p_i \in conv(D)$, we define its *best partition* P_i, or *partition* for short, as $P_i = \{u | p_i = \arg\max_{p \in D} \ u \cdot p\}$, which corresponds to the region in the utility space where p_i is the max-utility point. In another perspective, p_i being the max-utility point means that it ranks higher than any other point. Therefore, P_i is also equal to the intersection of a set of halfspaces and the utility space R_0, i.e., $(\bigcap_{p_j \in conv(D)/\{p_i\}} h^+_{i,j}) \cap R_0$, which is a $(d-1)$-dimensional convex polytope. Given a partition P and a hyperplane $h_{i,j}$, there are 3 cases: (1) P is in $h^+_{i,j}$, (2) P is in $h^-_{i,j}$ and (3) P intersects $h_{i,j}$. For example, as shown in Figure 3, (1) P_1 is in $h^+_{1,2}$, (2) P_2 and P_3 are in $h^-_{1,2}$, and (3) both P_4 and P_5 (marked with shaded regions) intersect $h^+_{1,2}$. Suppose that the true utility vector is verified to lie in halfspace $s_{i,j}$ and this could be done in Verification Phase (to be described later) w.h.p. If a partition P_i is disjoint from this *verified* halfspace $s_{i,j}$, we can safely prune p_i from further consideration because for any $u \in R$, there is always some point p_j that ranks higher than p_i. To find the relation between P and the hyperplane $h_{i,j}$, it is sufficient to check P's vertices with $h_{i,j}$ in $O(|V_P|)$ time, where V_P is the set of vertices of a convex polytope P.

In Verification Phase of *Verify-Point*, given a set S of halfspaces returned from Conjecture Phase, we want to find the "best" halfspace that is expected to help eliminate the largest fraction of partitions and their corresponding points from further consideration, so that we can minimize the number of questions. To efficiently find this halfspace, we maintain a table L, where each row records the relation between a hyperplane $h_{i,j}$ and a set X of all partitions that are intersected or contained by the utility range R. Specifically, L consists of 3 columns, which are named (1) $h^+_{i,j}$, which stores a set of all partitions in X that are entirely in $h^+_{i,j}$, (2) $h^-_{i,j}$, which stores a set of all partitions in X that are entirely in $h^-_{i,j}$, and (3) intersects, which stores a set of all partitions in X that intersect with $h_{i,j}$. An example of L corresponding to Figure 3 is shown in Table 2. By maintaining table L, we can efficiently find the best halfspace using a concept called $Num(\cdot)$. Specifically, given a halfspace s, we define $Num(s)$ to be the number of partitions that lie entirely outside s. For example, from Table 2, we could compute $Num(h^+_{2,5}) = 2$ since there are two partitions that are completely outside halfspace $h^+_{2,5}$ (because P_3 and P_5 are in $h^-_{2,5}$). As described before, we want to find the best halfspace that is expected to prune the *largest* number of partitions. Thus, the best halfspace is defined to be the halfspace in S with the greatest value of $Num(\cdot)$ (i.e., $\arg\max_{s \in S} Num(s)$).

We are now ready to introduce *Verify-Point* which involves two steps. The first step is the initialization step. Initially, we set the utility range R to be the entire utility space R_0 and the set C containing all possible best points to be $conv(D)$. We also set L to record the relation between the set of all hyperplanes $\{h_{i,j} | p_i, p_j \in conv(D)\}$ and the set of all partitions $\{P_i | p_i \in conv(D)\}$. The second step is the iterative step which involves a number of iterations where at each iteration, Conjecture Phase is first performed and Verification is then performed until some stopping conditions are satisfied. Thus, Conjecture Phase and Verification Phase is performed in an interleaving way for this iterative step.

- **Conjecture Phase:** We create variable R' and variable C', denoting the current content of the utility range R and the current content of set C, respectively, just before Conjecture Phase is performed. The intuition of why we maintain these copies can

be understood as follows. In Conjecture Phase, R' and C' is *updated* based on "conjectures" (which could be regarded as the information which has a lower confidence) according to the steps described in Section 5.2 (which includes how to update S, initialized to an empty set each time we re-perform Conjecture Phase). It is worth mentioning that variable R, variable C and variable L does *not* change in Conjecture Phase. But, they will be updated in the next Verification Phase based on "verified" information, and thus, these data structures are correct with high confidence.

- **Verification Phase:** Verification Phase runs for several rounds.
  - In each round, it selects the best halfspace in S and verifies with the user using the checking subroutine developed in Section 4.2. Formally, we select $s_{i,j} = \arg\max_{s \in S} Num(s)$ and then perform $Check(p_i, p_j, k)$ for checking. Based on the checking result, the correct halfspace may be $h^+_{i,j}$ or its counterpart $h^-_{i,j}$. For the ease of illustration, assume that $h^+_{i,j}$ is correct. We then update the data structures R, C, S and L as follows: (1) update the utility range R to $R \cap h^+_{i,j}$; (2) update the set S of halfspaces: remove $s_{i,j}$ from S; then, for each halfspace $s \in S$, if R is contained completely in s, or if R is completely outside s (which may happen due to user errors), remove s from S because no useful information can be obtained from s (because the question generated from s cannot help us further reduce the size of R); (3) for each partition P_l in $h^-_{i,j}$, we delete P_l from all rows of L and remove the corresponding point p_l from the set C; (4) for each partition P_l which intersects $h^+_{i,j}$, we update P_l to $P_l \cap h^+_{i,j}$, and update each row in L (corresponding to a hyperplane) by recomputing the relation of the updated P_l with the hyperplane; and (5) for each row in L and its corresponding hyperplane h, if R lies completely on one side of h (i.e., h^+ or h^-), we remove this row from L because we already know that all remaining partitions in R lies on one side of h (i.e., h^+ or h^-), so maintaining the relationship between h and partitions is no longer needed. We can easily verify that after the above steps, the updated table L correctly stores the relations between hyperplanes and partitions within the updated utility range R.
  - Verification Phase keeps selecting the next halfspace for checking until (1) there is only 1 point left in C, which is returned as the final best point, or (2) after running for at least 1 round, all halfspaces $s \in S$ cannot prune at least β_1 portion of the remaining partitions in R, where β_1 is a non-negative real number and a user parameter. When the first stopping condition is satisfied, we could terminate the whole algorithm (since we find the answer already). When the second stopping condition is satisfied, we need to enter the next iteration and re-set variable S to an empty set.

As a running example, consider the example in Figure 3. The utility range R is the outer triangle and the partitions are polygons bounded by solid lines. The corresponding set C containing possible best points is $\{p_1, p_2, p_3, p_4, p_5\}$ and table L is shown in Table 2. Assume that the algorithm enters Verification Phase with $S = \{h^+_{2,5}, h^+_{1,4}, h^+_{2,4}\}$. Verification Phase chooses to check $h^+_{2,5}$ first, because $Num(h^+_{2,5})$ is the largest (i.e., 2 in this case). Assume that after the user is checked via the checking subroutine, $h^+_{2,5}$ is verified to be correct. Then, the algorithm will update the utility range R to $R \cap h^+_{2,5}$, remove p_3 and p_5 from C (since P_3 and P_5 are in $h^-_{2,5}$) and remove $h^+_{2,5}$ from S. The updated L is shown in Table 3.

The main theorem of *Verify-Point* is presented in Theorem 3.

Theorem 3. *Given the input size n, the error rate θ and a failure probability δ, Verify-Point returns the best point with probability at least $1 - \delta$, using an expected number of $O(\frac{c}{1-\theta} \log n(conj + \log\log n + \log \frac{1}{\delta}))$ rounds where c is a data-dependent parameter denoting the pruning power and $conj$ is the round complexity of Conjecture Phase.*

Proof sketch. We first show that in the worst case, the number of iterations is $O(\frac{c}{1-\theta} \log n)$ and each iteration consists of $O(conj + k)$ rounds where k is the number of rounds in a checking subroutine. Therefore, the total number of rounds is $O(\frac{c}{1-\theta} \log n) \cdot O(conj + k)$ and the total success probability is at least $1 - O(\frac{cP_k}{1-\theta} \log n)$. Lemma 1 shows that to let $\delta = O(\frac{cP_k}{1-\theta} \log n)$, it suffices to set $k = O(\log\log n + \log \frac{1}{\delta})$, which completes the proof. □

5.3.2 Verification Phase of Verify-Space. Our second algorithm, *Verify-Space*, is closely related to *Verify-Point*: Its Verification Phase also runs for several rounds and in each round, it selects the best halfspace in S for checking (i.e., calling the checking subroutine), where S stores all halfspaces indicated by the user in Conjecture Phase. However, the key difference is that instead of finding the halfspace that prunes the largest *number* of partitions, *Verify-Space* directly find the halfspace that prunes the largest *space* of utility range R, which makes it less data-dependent.

Verify-Space shares the same structure with *Verify-Point*. The main difference is that in *Verify-Space*, whenever we want to decide the next halfspace for checking, we find the halfspace $s \in S$ that is expected to remove the *largest* space of the utility space (i.e., the halfspace $s \in S$ that *minimizes* the resulting utility space R). To find this "best" halfspace, one major step is to compute the volume of the polytope formed by intersecting the halfspace s with the current utility range R (i.e., $s \cap R$). Let $vol(P)$ denote the volume of a $(d-1)$-dimensional polytope P. Our task is to find $s_{i,j} = \arg\min_{s \in S} vol(s \cap R)$. However, computing the volume of a polytope in a high-dimensional space is time-consuming. There are several approximation algorithms that can estimate the volume of polytope, but to the best of our knowledge, even the fastest algorithm among them [6] takes $O(d^3)$ time, which is very time-consuming if we compute the volume of $s \cap R$ for each $s \in S$. Fortunately, what we really want is to minimize the ratio between the new utility range and the old one (i.e., $\frac{vol(s \cap R)}{vol(R)}$), and thus, it is not necessary to compute the exact volume. Therefore, we apply a random sampling technique called *Billiard Walk* [7] to sample N points in R and use the sampled points to compute this ratio. Lemma 3 shows that sampling a small number of points suffices to compute all ratios with high accuracy.

Lemma 3. *Let $R \subseteq \mathbb{R}^d$ be the utility range. Let T be a set of points randomly sampled from R. Then, given a non-negative real number $\rho \in [0, 1]$ and a non-negative real number ε, if sample size*

$|T| = O(\frac{1}{\varepsilon^2}(\log\frac{1}{\rho}+d))$, *then with probability* $1-\rho$, *for any halfspace* $s \subseteq \mathbb{R}^d$, $\frac{vol(s\cap R)}{vol(R)} - \frac{|\{t\in T|t\in s\cap R\}|}{|\{t\in T|t\in R\}|} \leq \varepsilon$.

The above lemma is derived directly from Theorem 5 in [16]. In order to make the relative errors of all intersections smaller than ε, we need only to sample $|T| = O(\frac{1}{\varepsilon^2}(\log\frac{|S|}{\rho}+d))$ points. The failing probability can be easily proved to be less than ρ using the union bound. As a convention, in our experiments, we set $\rho = 0.1$.

To illustrate Verification Phase of *Verify-Space*, consider the example shown in Figure 4. Assume that now the utility range is R and after Conjecture Phase, we record $S = \{s_1, s_2, s_3\}$ from the user. After sampling, the algorithm first chooses to verify s_2 because it has the smallest intersection with R. If the user confirms that s_2 is correct, then the utility range is updated to $s_2 \cap R$, and s_2 is removed from S. The algorithm then continues to select from S the next halfspace that has the smallest intersection with $s_2 \cap R$.

The main conclusion of *Verify-Space* is presented in Theorem 4.

Theorem 4. *Given the input size n, the error rate* θ *and a failure probability* δ, *Verify-Space returns the best point with probability at least* $1-\delta$, *using an expected number of* $O(\frac{d}{1-2\theta}\log n(conj + \log\log n + \log\frac{1}{\delta})$ *rounds, where d is the dimensionality and conj is the round complexity of Conjecture Phase.*

Proof sketch. We prove a special case where we set $\beta_2 = 0.5$, although in practice, β_2 can be set to another value to further improve the performance. We first prove that in the worst case, the expected number of iterations is $O(\frac{d}{1-2\theta}\log n)$, and each iteration uses $O(conj + k)$ rounds where k is the number of comparisons in the checking subroutine. The total success probability is $1 - O(\frac{d}{1-2\theta}\log nP_k)$ using union bound. Lemma 1 shows that to let $\delta = O(\frac{d}{1-2\theta}\log nP_k)$, it suffices to set $k = O(\log\log n + \log\frac{1}{\delta})$. □

6 EXPERIMENT

We conducted experiments on a computer with 1.80 GHz CPU and 12 GB RAM. All programs were implemented in C/C++.

Datasets. The experiments were conducted on synthetic and real datasets that are used in [19, 27, 29]. Specifically, we generated *anti-correlated* datasets by a dataset generator developed for skyline operators [5]. Besides, we used 4 real datasets, namely *Island*, *Weather*, *Car* and *NBA*. *Island* is two-dimensional and contains 63,383 geographic locations. *Weather* includes 178,080 tuples described by 4 attributes. *Car* is a 4-d dataset consisting of 68,005 used cars after cars whose attribute values fall out of a normal range are filtered out. *NBA* involves six attributes and contains 16,916 players after deleting records with missing values. Each dimension is normalized into range $[0, 1]$. Following the existing studies [27, 29], we preprocessed all the datasets to contain only the skyline points (which are all the possible best points for any utility function), since we are only interested in the best point.

Algorithms. We evaluated our 2-d algorithm: *2RI* and two multi-dimensional algorithms: *Verify-Space* and *Verify-Point*. The competitor algorithms are: *Median* [29], *Hull* [29], *2D-PI* [27], *Active-Ranking* [12], *Preference-Learning* [22], *UtilityApprox* [19], *UH-Simplex* [29], *HD-PI* [27] and *RH* [27]. Since some of them cannot return the best point directly, when comparing them with our algorithms, we made the following adaptions.

(1) Algorithms *Median*, *Hull* and *2D-PI* are designed for 2 dimensional tasks. *Median* and *Hull* already aim at returning the user's best point so they are left unchanged. *2D-PI* returns one of top-k points where k is a user parameter, so we set k to 1 such that it returns the best point. (2) Algorithm *Active-ranking* aims at learning the entire ranking of all points by interacting with the user. We return the best point after the full ranking is determined. (3) Algorithm *Preference-learning* focuses on learning the utility vector of the user. According to the experimental result in [22], the utility vector learnt is very close to the theoretical optimum if the error threshold ϵ of the learnt utility vector is set to a value below 10^{-5} (e.g., 10^{-6}). In our experiment, we set ϵ to 10^{-6} (since the learnt utility vector could achieve the optimum), and return the best points w.r.t. the learnt utility vector. (4) Algorithm *UtilityApprox* and *UH-Simplex* focus on reducing the regret ratio below a given threshold ϵ. Following [27], we set $\epsilon = 1 - f(p_2)/f(p_1)$, where p_1 and p_2 are the top-1 and top-2 points w.r.t. the user's utility vector, respectively. In this way, if no user error is made, the returned point is guaranteed to be the best point. (5) Algorithms *HD-PI* and *RH*, similar to the 2-d algorithm *2DPI*, aim at returning one of the top-k points. We ensure they return the best point by setting $k = 1$.

Parameter Setting. We evaluate the performance of each algorithm by varying different parameters: (1) the dataset size N; (2) the dimensionality d, (3) the user error rate upper bound θ, (4) the stopping threshold β_1 in *Verify-Point* and β_2 in *Verify-Space*, (5) the variable ε in Lemma 3, and (6) the parameter k in the checking subroutine. Unless stated explicitly, for each synthetic dataset, we set $N = 100,000$ and $d = 4$. We set $\theta = 0.05$, which is a reasonable error rate upper bound according to the human reliability assessment data in [14]. According to the experimental results in Section 6.1, we set the default value of $\beta_1 = 0.2$, $\beta_2 = 0.2$, $k = 3$ and $\varepsilon = 0.1$.

Performance Measurement. We evaluate the performance of each algorithm with the following measurements: (1) *Accuracy* which is the probability of retrieving the best point. Formally, we define the accuracy to be $\frac{N_c}{N_e}$ where N_e is the total number of experiments and N_c is the number of times the best point is returned. (2) *Number of questions* required to return the result, and (3) *Execution time* which is the average processing time to determine the question asked in each round. For each setting, we repeat the algorithm 100 times and the average value is reported.

In Section 6.1, we study different configurations of our algorithms. The performance of all algorithms on synthetic and real datasets are discussed in Section 6.2 and 6.3, respectively. We conducted a user study with user errors in Section 6.4. Finally, we summarize the experiments in Section 6.5.

6.1 Study on Configuration of Our Algorithms

In this section, We study the effect of different settings of parameters β_1, β_2, ε and k on our algorithms *2RI*, *Verify-Point* and *Verify-Space*. We also study the value of parameter c in Theorem 3. Finally, we test the flexibility of our algorithms.

We studied the effect of ε on *Verify-Space* and the experimental result can be found in the Appendix. When varying ε between 0.05 and 0.25, we observe no significant change in terms of the number of questions and the accuracies. However, when $\varepsilon > 0.15$, *Verify-Space* becomes less stable since the standard deviation of the

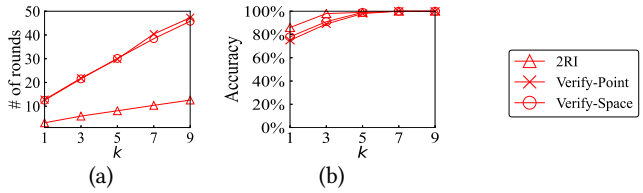


Figure 5: Effect of k

number of questions asked by the algorithm increases, which can be attributed to the less accurate volume ratio estimation. Moreover, when $\varepsilon < 0.1$, the average processing time increases due to the increase in sample size. We eventually decided to set $\varepsilon = 0.1$.

We studied the effect of β_1 on *Verify-Point* and β_2 on *Verify-Space* when ranging them from 0 to 0.5. The related figures can be found in the Appendix. Both algorithms use the least number of rounds when $\beta_1 = \beta_2 = 0.2$. On the other hand, changing β does not have obvious impact on the accuracy of finding the best point. Therefore, in the rest of our experiments, we set $\beta_1 = \beta_2 = 0.2$.

In Figure 5, we evaluated the effect of different choices of k. The experiments of *2RI* were conducted on 2-d synthetic datasets and the experiments of *Verify-Point* and *Verify-Space* were conducted on 4-d synthetic datasets. Since the checking subroutine needs to make decisions based on the majority of answers, k only takes odd values. We observe that the number of rounds and the accuracy both increase with k. Specifically, when $k \geq 5$, the accuracies of all algorithms are very close to 1. But, the trade-off is that an additional number of questions needs to be asked. Therefore, we choose $k = 3$ for later experiments because it asks a small number of questions while achieving a satisfactory level of accuracy (i.e., 90%).

We studied on the value of c shown in Theorem 3 on different datasets. Specifically, we evaluated c on 3 real datasets, namely *Car*, *NBA* and *Weather*, and 3 synthetic datasets, namely *Anti*, *Corr* and *Indep*, which were generated using the dataset generator developed in [5], consisting of anti-correlated data, correlated data and independent data with size 100,000, respectively. The results can be found in the Appendix. For all the datasets we tested, the value of c never exceeds 3. Therefore, c can be regarded as a small constant.

Experiments about Flexibility: To show the flexibility of our proposed framework, we conducted experiments for the following 2 variants of our framework: (1) [Output Size Flexibility] Instead of returning only the best point, we adapted *Verify-Point* and *Verify-Space* to (a) return a set of at most k points containing the best point; (b) return a set of k points with the highest utility scores. (2) [User Input Form Flexibility] Instead of performing pairwise comparisons, one user input form, in Conjecture Phase of *Verify-Point* and *Verify-Space*, we adopt to (a) display $s \geq 2$ points and ask the user to select the favorite point, which is a type of interaction [19, 29]; (b) display $s \geq 2$ points and ask the user to partition them into two groups, namely the *superior* group and the *inferior* group, which is another type of interaction adopted in [18]. For the sake of space, the results of these experiments can be found in the Appendix.

6.2 Performance on Synthetic Datasets

We compared our 2-d algorithm *2RI* against *2DPI*, *Median* and *Hull* on a 2-d synthetic dataset. For completeness, we also record the performance of all d-dimensional algorithms. Figure 6 presents the performance of all the algorithms when N varies from 100 to 1,000,000. According to Figure 6 (c), all algorithms determine the next question to be asked within 10^{-2} second and their execution times do not increase significantly when the input size grows. As shown in Figure 6 (b), *2RI* achieves the highest accuracy among all the algorithms under all settings, and finishes within only 5-6 questions as shown in Figure 6 (a). On the other hand, algorithm *2DPI*, *Median* and *Hull* cannot efficiently handle user errors and their best point retrieval accuracies is at least 10 percentage smaller than *2RI*. It is worth mentioning that a lower accuracy could lead to unforgettable and unchangeable consequences as described in Section 1.

In Figure 7, we evaluated the performance of our algorithms *Verify-Point* and *Verify-Space* against other existing d-dimensional algorithms on 4-d synthetic datasets. We observe that *Verify-Point* and *Verify-Space* are the most accurate on finding the best point, and their accuracies decrease with the slowest rate along with the increase in the input size. Algorithm *UtilityApprox* has the highest accuracy among the remaining algorithms, but, it is still 10% lower than ours and it asks around 10 more questions. *HDPI* and *UH-Simplex* ask slightly fewer questions than our algorithms. But, their accuracies are more than 10 percent and 30 percent lower than ours, respectively. Algorithm *Active-Ranking* aims at learning the entire ranking and algorithm *Preference-Learning* aims at learning the user's utility vector. Therefore, although they claim that they can handle user errors, their performance is poor in terms of both the number of questions asked and the best point retrieval accuracy. According to Figure 7 (c), for all algorithms, the running time on deciding the next question to be asked increase along with the input size. But, *Verify-Point* and *Verify-Space* finish within 1 second on 1M datasets, which is acceptable for real-time interaction.

We studied the effect of different values of the user error rate θ on d-dimensional algorithms. Figure 8 shows the result. When θ grows, the degeneration of accuracy can be observed on all the algorithms. However, the performance of *Verify-Point* and *Verify-Space* degenerates at the slowest rate and their accuracies remain the highest for all settings of θ. In particular, when θ is high (i.e., 0.15), our algorithms can still reach 70% accuracy while all other algorithms are lower than 50%. We notice that the number of questions asked by *Verify-Point* and *Verify-Space* gradually increases when θ increases, which can be attributed to the extra checking rounds required to resolve conflicts caused by extra user errors. The number of questions grows in a slow rate and both algorithms finish within 30 questions even if the error rate is high (i.e., 0.15). Figure 9 presents our experiment on evaluating the scalability on the dimensionality d. Compared with the existing algorithms, *Verify-Point* and *Verify-Space* constantly achieve higher accuracies for all dimensional settings, and the gap gets even larger when d increases. Meanwhile, the number of questions asked by our algorithms grows only by 3-4 questions for each dimensionality increase. As for the running time, all algorithms require more time when the dimensionality is larger. However, for $d = 5$, our algorithms still finish within 1 second, which is an acceptable interactive speed.

We measured the percentage of executions of *Verify-Point* and *Verify-Space* that finish with different numbers of iterations on 3d, 4d and 5d synthetic datasets. For the sake of space, those figures are put in the Appendix. All executions finish within 3 iterations

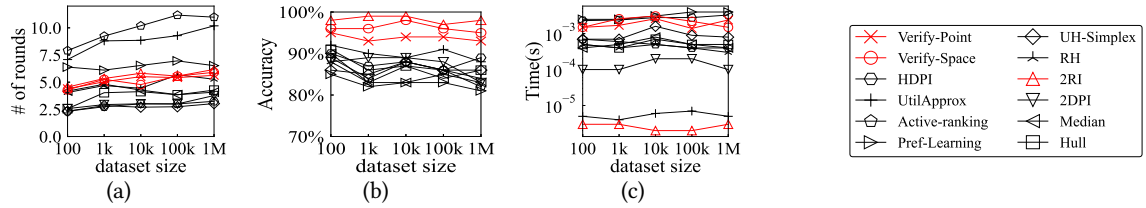


Figure 6: 2d dataset

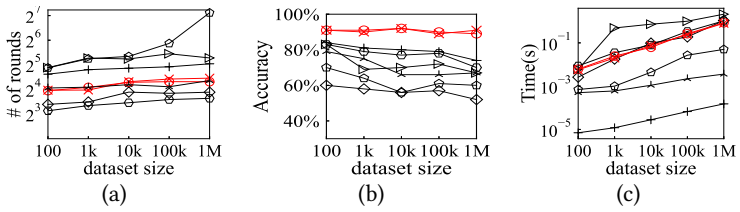


Figure 7: 4d dataset

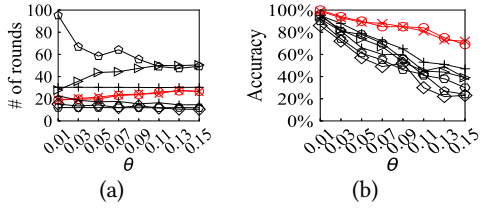


Figure 8: Effect of θ

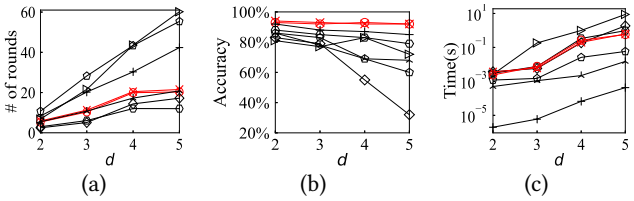


Figure 9: Effect of d

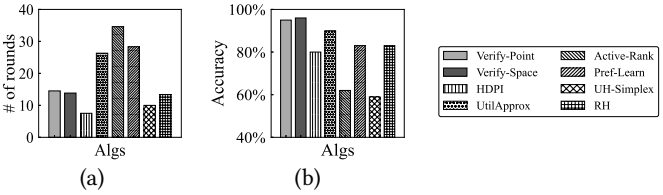


Figure 10: Results on dataset *Weather*

and there is no significant difference between these two algorithms. When dimensionality increases, the percentage of executions that finish with more than 1 iteration also slightly increases. For example, when $d = 3$, about 85% of executions finish within 1 iteration, only 15% use 2 iterations, and less than 1% use 3 iterations to finish. On the other hand, when $d = 5$, the percentage of executions that take 1, 2 and 3 iterations are about 73%, 24% and 3%, respectively. One possible explanation is that when the dimensionality increases, an error in Conjecture Phase will have a higher impact due to more complex spatial structure, making the algorithm more likely to focus on the wrong region. As a consequence, an additional iteration must take place to perform a new round of searching.

We also studied the average number of questions in each iteration and each phase (i.e., Conjecture Phase and Verification Phase) on these 3 synthetic datasets. The results can be found in the Appendix. When d increases, the ratio between the number of questions asked in Verification Phase and the number of questions asked in Conjecture Phase also increases. For example, when $d = 3$, this ratio is around 0.47, but when $d = 5$, this ratio grows to around 0.54. This is explainable since in Verification Phase, we verify the correctness of each subspace $s \in S$, where S contains the subspaces indicated by the user in Conjecture Phase. Recall that after checking some subspaces $s \in S$ for in the first several rounds, some other subspaces $s \in S$ may be removed from S and need not to be verified again since they contain the entire utility range and are no long useful for updating it. However, it may be harder to remove the subspaces in higher dimensions, resulting in more subspaces needed to be checked in Verification Phase.

6.3 Performance on Real Datasets

We conducted experiments on 4 real datasets, namely *Island*, *Car*, *Weather* and *NBA*. For four 2-d algorithms *2RI*, *2DPI*, *Median* and *Hull*, their performance on the 2-d dataset *Island* is reported. For completeness, the performance of all 8 d-dimensional approaches are also reported. Due to the space constraint, the experimental results can be found in the Appendix. Among all 2-d algorithms, *2RI* requires slightly more than 5 questions, and obtains the highest accuracy which is around 97%. Other competitors take 1-2 less questions compared with *2RI*, but their accuracies are much lower. Specifically, *2DPI* and *Median* obtain accuracies below 90%, and *Hull* is less than 80%.

For d-dimensional algorithms, we studied their performance on all 4 datasets. Due to the lack of space, we only show the results on *Weather* in Figure 10. The performance on *Island*, *Car* and *NBA* can be found in the Appendix. *Verify-Point* and *Verify-Space* obtain the highest accuracy using a small number of questions on all datasets, which is consistent with their performance on synthetic datasets.

We also studied the percentage of executions of *Verify-Point* and *Verify-Space* that finish with different numbers of iterations, as well as the number of questions in each phase, on all 4 real datasets. The performance of our algorithms on real datasets is consistent with synthetic datasets, and the results of these experiments could be found in the Appendix due to space constraint.

6.4 User Study

To see the impact of user errors and how our algorithms can help improve the quality of the returned point, we conducted a user study on the *Car* dataset. Following the same settings in [22, 27, 29], we

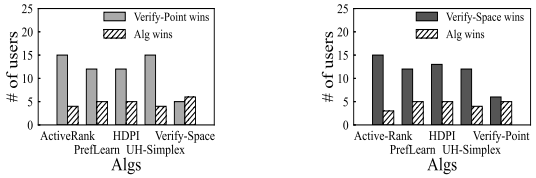
randomly selected 1000 candidate cars from the database and each record is described by 4 attributes, namely price, year of purchase, power and used kilometers. We compared our algorithms *Verify-Point* and *Verify-Space* against 4 existing algorithms, namely *HD-PI*, *UH-Simplex*, *Preference-learning* and *Active-Ranking*. For each algorithm, the users were asked to select the preferred car from a pair of cars for several rounds until a car is returned. 25 participants were recruited and their average results were reported. Since we cannot directly obtain the user's utility vector, *UH-Simplex* and *Preference-Learning* were re-adapted (different from the way described previously): (1) Algorithm *Preference-Learning* maintains an estimated user's utility vector u during the interaction. We compared the user's answers of some randomly selected questions with the prediction w.r.t. u. If 75% questions [22] can be correctly predicted, we stop and return the best point w.r.t. u. (2) For *UH-Simplex*, we set the threshold $\epsilon = 0$ and this guarantees that the returned car is believed to be the best point in the algorithm's view.

Each algorithm was measured via the following metrics: (1) *the number of questions asked*; and (2) *the dissatisfactory level*, which is an integer score ranging from 0 to 10 given by each participant. It indicates how dissatisfied the participant feels about the returned car, where 0 indicates the least dissatisfied and 10 the most dissatisfied.

We also measured how frequently user errors occur during interactions. Since we cannot directly verify the correctness of each answer, we created a new version of *Verify-Point*, called *Verify-Point-Adapt*, and changed the checking subroutine as follows: Instead of stopping immediately once the majority is determined, it always checks a pair of points for exactly k times before returning the answer. Modified in this way, let n_c denote the number of checking subroutines invoked, and denote w_i the number of questions asked and m_i the number of majority answers of the user in the i-th checking, $1 - \frac{\sum_{i=1}^{n_c} m_i}{\sum_{i=1}^{n_c} w_i}$ will be an unbiased estimator of the user error rate. Among all the 220 checking questions asked by *Verify-Point-Adapt*, users made 10 errors, yielding an overall error rate of 4.5%.

In the experiment, we observe that *Verify-Point* and *Verify-Space* obtain the lowest dissatisfaction score, which is slightly above 1 (out of 10). On the other hand, the dissatisfaction scores of all other competitors are at least 3. As for the number of questions asked, *HDPI* and *UH-Simplex* use the least number of rounds which is around 8. *Verify-Point* and *Verify-Space* also finish with slightly over 10 rounds. On the other hand, *Active-Ranking* and *Preference-Learning* require a lot more rounds. They use on average 23 and 32 rounds, respectively. The related figures can be found in the Appendix.

In Figure 11 (a) and (b), we compared the user's preference on the recommendations of our algorithms *Verify-Point* and *Verify-Space*, respectively, against *HDPI*, *UH-Simplex*, *Preference-Learning* and *Active-Ranking*. Specifically, for each user, if the dissatisfaction score of algorithm A is lower than algorithm B, then the recommendation of A is better than B. For example, in Figure 11 (a), when *Verify-Point* is compared with *HDPI*, 12 users favor the car recommended by *Verify-Point*, while only 4 think the recommendation of *HDPI* is more preferred. Overall, the recommendation made by our algorithms are much more preferred than the existing algorithms. On the other hand, there is no significant gap between the performance of *Verify-Point* and *Verify-Space*.



(a) *Verify-Point* (b) *Verify-Space*

Figure 11: User study comparing whether Verify-Point/Verify-Space is better than each of the competitors

We studied the percentage of *Verify-Point* and *Verify-Space* that finished on each iteration in user study. The figures for these experiments can be found in the Appendix. In particular, almost 90% of users finished within only 1 iteration on *Verify-Point* and *Verify-Space*, and 10% of users finished in 2 iterations. Only 4% of users (i.e., one user) ended up with 3 iterations when using *Verify-Point*. Compared with the experimental results on multi-dimensional synthetic and real datasets, the percentage of users who need more than 1 iteration is smaller in our user study, probably because the user error rate in real world (as found by us as 4.5%) is even smaller than our default setting of θ (i.e., 5%).

6.5 Summary

The experiments demonstrated the superiority of our 2-d algorithm *2RI* and our d-dimensional algorithms, *Verify-Point* and *Verify-Space*, over the existing approaches: (1) we are efficient and effective. We achieve nearly 100% accuracy in most experiments within a small number of rounds. In particular, for all 2-d experiments, *2RI* consistently outperforms all other algorithms and obtains an accuracy close to 100% even when the input size is large (e.g., 1,000,000). In d-dimensional experiments, *Verify-Point* and *Verify-Space* maintain the highest accuracy among all competitors. (2) The scalability of *Verify-Point* and *Verify-Space* is demonstrated. Specifically, they are scalable to the input size and the dimensionality. For example, on the 6-d dataset *NBA*, our algorithms obtain over 90% accuracies using only slightly over 20 questions, while *UtilitApprox*, *Preference-Learning* and *Active-Ranking* are significantly lower than ours using over 40 questions. (3) Our algorithms are capable to handling many user errors. When the user error rate is very large (e.g., 0.15), only our algorithms can keep the accuracies above 70%, while all other algorithms drop under 50%.

7 CONCLUSION

In this paper, we propose a more robust interactive model that returns the best point in dataset under the setting of random errors. Our model is more practical than existing algorithms in a sense that ours can return the best point with high confidence even if the user makes mistake when interacting with the system. Specifically, we propose a 2-d algorithm that is asymptotically optimal in terms of the number of rounds required and two multi-dimensional algorithms with provable guarantee and superior empirical performance. We conducted extensive experiments to show that our algorithms is both efficient and effective in determining the best point facing user errors compared with existing algorithms. In the future, we consider the case where user makes persistent errors when answering questions.

REFERENCES

[1] Yongkil Ahn. 2019. The economic cost of a fat finger mistake: a comparative case study from Samsung Securities's ghost stock blunder. *Journal of Operational Risk* 16, 2 (2019).

[2] Ilaria Bartolini, Paolo Ciaccia, Vincent Oria, and M Tamer Özsu. 2007. Flexible integration of multimedia sub-queries with qualitative preferences. *Multimedia Tools and Applications* 33, 3 (2007), 275–300.

[3] Ilaria Bartolini, Paolo Ciaccia, and Marco Patella. 2014. Domination in the probabilistic world: Computing skylines for arbitrary correlations and ranking semantics. *ACM Transactions on Database Systems (TODS)* 39, 2 (2014), 1–45.

[4] Ilaria Bartolini, Paolo Ciaccia, and Florian Waas. 2001. FeedbackBypass: A new approach to interactive similarity query processing. In *VLDB*. 201–210.

[5] Stephan Borzsony, Donald Kossmann, and Konrad Stocker. 2001. The skyline operator. In *Proceedings 17th international conference on data engineering*. IEEE, 421–430.

[6] Apostolos Chalkis, Ioannis Z Emiris, and Vissarion Fisikopoulos. 2019. A practical algorithm for volume estimation based on billiard trajectories and simulated annealing. *arXiv preprint arXiv:1905.05494* (2019).

[7] Apostolos Chalkis and Vissarion Fisikopoulos. 2020. volesti: Volume approximation and sampling for convex polytopes in r. *arXiv preprint arXiv:2007.01578* (2020).

[8] Sean Chester, Alex Thomo, S Venkatesh, and Sue Whitesides. 2014. Computing k-regret minimizing sets. *Proceedings of the VLDB Endowment* 7, 5 (2014), 389–400.

[9] Mark Theodoor De Berg, Marc Van Kreveld, Mark Overmars, and Otfried Schwarzkopf. 2000. *Computational geometry: algorithms and applications*. Springer Science & Business Media.

[10] Brian Eriksson. 2013. Learning to top-k search using pairwise comparisons. In *Artificial Intelligence and Statistics*. PMLR, 265–273.

[11] Kilem L Gwet. 2008. Intrarater reliability. *Wiley encyclopedia of clinical trials* 4 (2008).

[12] Kevin G Jamieson and Robert Nowak. 2011. Active ranking using pairwise comparisons. *Advances in neural information processing systems* 24 (2011).

[13] Yiling Jia, Huazheng Wang, Stephen Guo, and Hongning Wang. 2021. Pairrank: Online pairwise learning to rank by divide-and-conquer. In *Proceedings of the Web Conference 2021*. 146–157.

[14] Barry Kirwan. 2017. *A guide to practical human reliability assessment*. CRC press.

[15] Jongwuk Lee, Gae-won You, Seung-won Hwang, Joachim Selke, and Wolf-Tilo Balke. 2012. Interactive skyline queries. *Information Sciences* 211 (2012), 18–35.

[16] Yi Li, Philip M Long, and Aravind Srinivasan. 2001. Improved bounds on the sample complexity of learning. *J. Comput. System Sci.* 62, 3 (2001), 516–527.

[17] Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval* 3, 3 (2009), 225–331.

[18] Denis Mindolin and Jan Chomicki. 2009. Discovering relative importance of skyline attributes. *Proceedings of the VLDB Endowment* 2, 1 (2009), 610–621.

[19] Danupon Nanongkai, Ashwin Lall, Atish Das Sarma, and Kazuhisa Makino. 2012. Interactive regret minimization. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. 109–120.

[20] Danupon Nanongkai, Atish Das Sarma, Ashwin Lall, Richard J Lipton, and Jun Xu. 2010. Regret-minimizing representative databases. *Proceedings of the VLDB Endowment* 3, 1-2 (2010), 1114–1124.

[21] Peng Peng and Raymond Chi-Wing Wong. 2014. Geometry approach for k-regret query. In *2014 IEEE 30th International Conference on Data Engineering*. IEEE, 772–783.

[22] Li Qian, Jinyang Gao, and HV Jagadish. 2015. Learning user preferences by adaptive pairwise comparison. *Proceedings of the VLDB Endowment* 8, 11 (2015), 1322–1333.

[23] Assessing Questionnaire Reliability. 2022. https://select-statistics.co.uk/blog/assessing-questionnaire-reliability/

[24] Gerard Salton. 1989. Automatic text processing: The transformation, analysis, and retrieval of. *Reading: Addison-Wesley* 169 (1989).

[25] Thomas Seidl and Hans-Peter Kriegel. 1997. Efficient user-adaptable similarity search in large multimedia databases. In *VLDB*, Vol. 97. 506–515.

[26] Christian Walck et al. 2007. Hand-book on statistical distributions for experimentalists. *University of Stockholm* 10 (2007), 96–01.

[27] Weicheng Wang, Raymond Chi-Wing Wong, and Min Xie. 2021. Interactive Search for One of the Top-k. In *Proceedings of the 2021 International Conference on Management of Data*. 1920–1932.

[28] A Student with Top-tier Score Admitted by mediocre University (Chinese version only). 2020. https://news.southcn.com/node_6854f1135c/4357641930.shtml

[29] Min Xie, Raymond Chi-Wing Wong, and Ashwin Lall. 2019. Strongly truthful interactive regret minimization. In *Proceedings of the 2019 International Conference on Management of Data*. 281–298.

[30] Jiping Zheng and Chen Chen. 2020. Sorting-based interactive regret minimization. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*. Springer, 473–490.

A ADDITIONAL EXPERIMENTS

Here, we supplement our experimental results that are not shown in the main paper due to space constraint. Section A.1, A.2, A.3, A.4 and A.5 shows the supplementary results on parameter settings, experiments on synthetic datasets, experiments on real-world datasets, experiments on flexibility and user study, respectively.

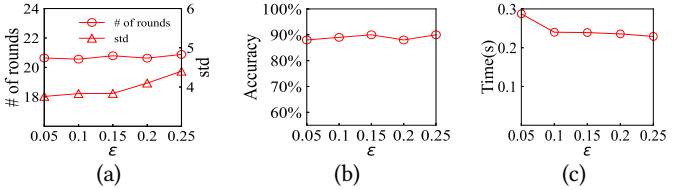


Figure 12: Effect of ε

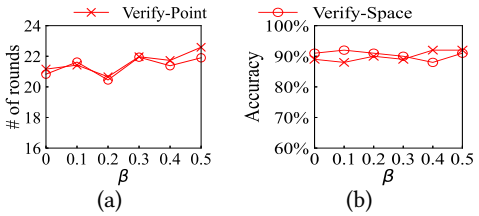


Figure 13: Effect of β

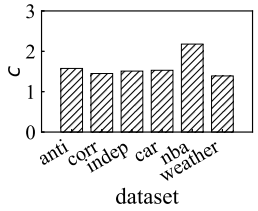


Figure 14: Effect of c

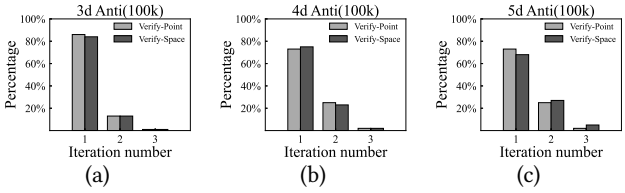


Figure 15: Percentage of iterations (synthetic datasets)

A.1 Additional experiments on parameter settings

In this section, we show the figures about parameter settings that are not shown in the main paper.

In Figure 12, we studied the effect of ε on *Verify-Space*. From Figure 12 (a) and (b), we observe no significant change in terms of the number of questions and the accuracies when varying ε between

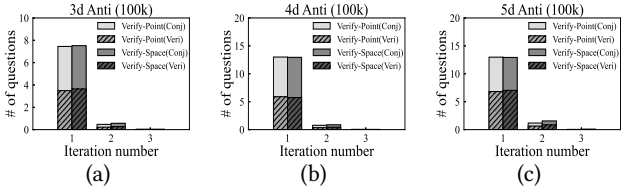


Figure 16: Average no. of questions in each phase (synthetic datasets)

0.05 and 0.25. But, in Figure 12 (a), when ε is set to a value greater than 0.15, *Verify-Space* becomes less stable since the standard deviation of the number of questions asked by the algorithm increases, which can be attributed to the less accurate volume ratio estimation. Moreover, when $\varepsilon < 0.1$, the average processing time increases due to the increase in sample size, as can be observed from Figure 12 (c). We decided to set $\varepsilon = 0.1$ since compared with $\varepsilon = 0.15$, it asks fewer questions.

Figure 13 shows the effect of β_1 on *Verify-Point* and the effect of β_2 on *Verify-Space* when ranging them from 0 to 0.5. As shown in Figure 13 (b), varying β between 0 to 0.5 does not have a significant impact on the accuracy. Since from Figure 13 (a), both algorithms take the least number of questions when $\beta_1 = \beta_2 = 0.2$, we set $\beta_1 = \beta_2 = 0.2$ in our other experiments.

Figure 14 shows the value of c in Theroem 3 on 3 real datasets, namely *Car*, *NBA* and *Weather*, and 3 real datasets, namely *Anti*, *Corr* and *Indep*, as described in Section 6.1. Since the value of c never exceeds 3 in all 6 datasets, it can be regarded as a small constant in practice.

A.2 Additional experiments on synthetic datasets

We supplement the figures about the experimental results on synthetic datasets that are not shown in the main paper in this section. Figure 15 measures the percentage of executions of *Verify-Point* and *Verify-Space* that finish with different numbers of iterations on $3d$, $4d$ and $5d$ synthetic datasets. In this figure, the bar at iteration number i shows how many percentage of executions terminate with i iterations. We only show the results on iteration number 1, 2 and 3 since all executions finish within 3 iterations. We observe that when dimensionality increases, the percentage of executions that finish with more than 1 iteration also slightly increases. For example, when $d = 3$, only 15% use 2 iterations, and less than 1% use 3 iterations to finish. On the other hand, when $d = 5$, the percentage of executions that take 2 iterations and 3 iterations are about 24% and 3%, respectively. A possible explanation of this behavior is give in the main paper.

Figure 16 shows the average number of questions in each iteration and each phase (i.e., Conjecture Phase and Verification Phase) on $3d$, $4d$ and $5d$ synthetic datasets. Similarly, we only show the results in iteration 1, 2 and 3 since all executions finish within 3 iterations. In these figures, "Conj" represents Conjecture Phase and "Veri" represents Verification Phase. As described in Section 6.2, we observe that the number of questions asked in each iteration decreases exponentially. The majority of questions are asked in iteration 1, which takes around 90% of total questions on average. On the other hand, we notice that when d increases, the ratio between the number of questions asked in Verification Phase and the number of questions asked in Conjecture Phase also increase. This trend is explainable and we give a possible explanation in the main paper.

A.3 Additional experiments on real datasets

In this section, we present the experiment results on real datasets that are not shown in the main body of the paper. In Figure 17, we show the results of our 2-d algorithms *2RI* with other 2-d competitors on 2-d real dataset *Island*. For completeness, the results of other multi-dimensional algorithms are also reported. In particular, *2RI* achieves 97% accuracy, outperforming all other competitors. In Figure 18 and Figure 19, we compare our proposed algorithms, namely *Verify-Point* and *Verify-Space*, with other d-dimensional algorithms on real datasets *Car* and *NBA*, respectively. The results on these datasets are consistent with the results on *Weather* that is shown in the main body. Specifically, our algorithms obtain the highest accuracy among all competitors by only asking a small number of questions. Algorithm *HDPI* and *UH-Simplex* uses less number of questions than our algorithms. However, from Figure 18 (b) and Figure 19 (b), their accuracy is at least 10% lower than ours on both datasets.

Figure 20 shows the percentage of executions of *Verify-Point* and *Verify-Space* that finish with different numbers of iterations on all 4 real datasets, namely *Island*, *Car*, *Weather* and *NBA*. The bar at iteration i indicates the percentage of executions that terminate using i iterations. Since all the executions finished within 3 iterations, we only show the iteration number 1, 2 and 3 here. On all these datasets, the the number of questions asked in each iteration decreases exponentially. The overall performance on these real dataset agrees with the performance on synthetic datasets and our analysis.

Figure 21 measures the average number of questions in each iteration and each phase (i.e., Conjecture Phase and Verification Phase) on all 4 real datasets. Similar to synthetic datasets, in these figures, "Conj" represents Conjecture Phase and "Veri" represents Verification Phase. Since all the executions finished within 3 iterations, we only show the results in iteration 1, 2 and 3. Consistent with the performance on synthetic datasets, the ratio between the questions asked in Verification Phase and the questions asked in Conjecture Phase slightly increases along with d.

A.4 Experiments about flexibility

We experimented on the following variants of *Verify-Point* and *Verify-Space* to test the flexibility of our proposed framework:

(1) [Output Size Flexibility] Instead of returning only the best point, we adapted *Verify-Point* and *Verify-Space* to (a) return a set of point with size at most k which contains the best point (Section A.4.1); (b) return top-k point which are k points with the highest utility score (Section A.4.2).

(2) [Output Size Flexibility] Instead of performing pairwise comparison in Conjecture Phase, we adapted Conjecture Phase of *Verify-Point* and *Verify-Space* to (a) display $s \geq 2$ points and ask the user to select the favorite point (Section A.4.3); (b) display $s \geq 2$ points and ask the user to partition them into two groups, namely the *superior*

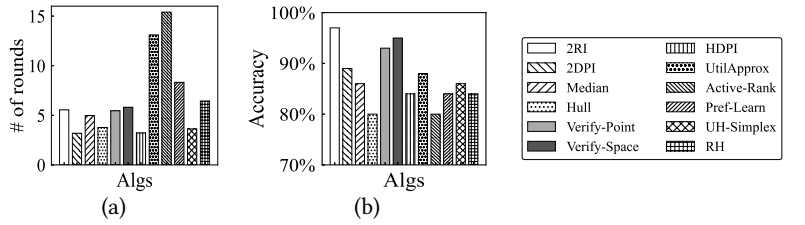


Figure 17: Results on dataset *Island*

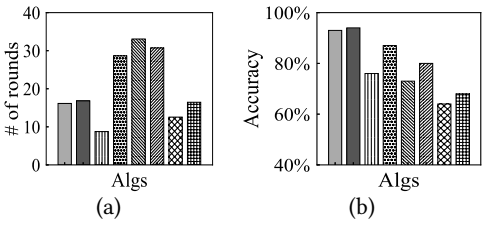


Figure 18: Results on dataset *Car*

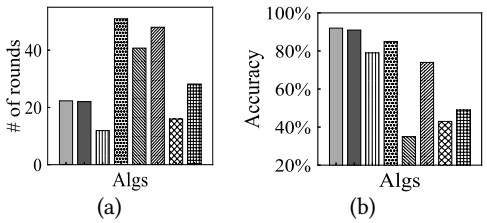


Figure 19: Results on dataset *NBA*

group and the *inferior* group, which is a type of interaction adopted in [18] (Section A.4.4).

We now describe the results of these experiments in the following sections.

A.4.1 Returning a set containing the best point. To return a set containing the best point with size at most k, we adapted the Conjecture Phase (resp. Verification Phase) of *Verify-Point* and *Verify-Space* in the following way. Instead of stopping when there is only 1 point in set C' (resp. C), it stops when there are at most k points in C' (resp. C). Upon termination, the algorithms report all points in C. Besides *Verify-Point* and *Verify-Space*, we also adapted *HDPI* and *UH-Simplex* for comparison. Specifically, *HDPI* and *UH-Simplex* maintain a set storing all possible best point, we adapted these algorithms to stop when there are at most k points in this set, so that the best point must be returned in these algorithms' view by returning this set. On the other hand, since algorithm *RH*, *UtilityApprox*, *Active-Ranking* and *Preference-Learning* do not maintain a data structure to record the possible best points, they can only guarantee returning the best point after it is determined, and thus, cannot be naturally extended to this variant.

Figure 22 shows the experimental results on this variant. From Figure 22 (a), all algorithms tend to ask fewer questions when the size of returned set increases, which is as expected since the algorithms will stop earlier when k grows. We also observe from Figure 22 (b) that *Verify-Point* and *Verify-Space* outperform existing algorithms by a large margin in terms of the best point retrieval accuracies. In particular, their accuracies are close to 100% when $k \geq 6$. The average processing time increases along with the size of the returned set, which is because the data structures become simpler as more questions are asked, and thus, during one execution, the processing time for latter questions are faster compared with former questions.

A.4.2 Returning top-k points. We adapted *Verify-Point* and *Verify-Space* to return top-k points. Specifically, instead of terminating when the best point is found, we keep on asking questions and further reducing the size of utility space until the top-k can be determined. Algorithm *HDPI*, *RH*, *Active-Ranking* and *Preference-Learning* are also adapted to return top-k for comparison. For *HDPI* and *RH*, after the best point is returned, we set them to keep on searching the next best point among the points that are not returned. This process repeats $k-1$ times until all top-k points are returned. Algorithm *Active-Ranking* focuses on learning the ranking of all points, so we return the top-k points after the entire ranking is determined. Algorithm *Preference-Learning* aims at learning user's utility vector. Similar to what we did in Section 6, it first learn the utility vector with an error threshold ϵ set to 10^{-6}, then return the top-k points w.r.t the learnt utility vector. As for algorithm *UH-Simplex* and *UtilityApprox*, since these algorithms requires a correctly-set regret ratio threshold, which is hard to obtain, to guarantee the correctness of the top-k, they cannot be naturally adapted to this variant.

The results of the top-k related experiments are presented in Figure 23. We observe from Figure 23 (a) that except *Preference-Learning*, all algorithm tend to ask more questions when k grows. The reason why *Preference-Learning* is not affected by k is because it aims at learning the utility vector instead of finding top-k. Even when k is large (i.e., $k = 20$), *Verify-Point* and *Verify-Space* can still finish within 60-70 questions. According to Figure 23 (b), the accuracies of all algorithms decreases when k increases. However, *Verify-Point* and *Verify-Space* outperforms all competitors by a large margin in terms of accuracies for all values of k. Specifically, when $k = 20$, all existing algorithms have lower than 20% accuracies, but our algorithms can still maintain around 60% accuracies. No significant change on processing time of our algorithms can be found in Figure 22 (c).

A.4.3 Selecting favorite point among s displayed points. To further test the flexibility of our framework, we changed the type of question asked in Conjecture Phase of *Verify-Point* and *Verify-Space*. The Verification Phase is left unchanged. Instead of displaying only 2 points in each round, we adapted the Conjecture Phase to select $s \geq 2$ points from C' in each round, and the user is required to select the favorite point (i.e., the point with the highest utility) among s displayed points. To simulate user errors, we assume that the user chooses the favorite point with probability at least $1-\theta$, and randomly choose one of the non-favorite point with probability at most θ. Note that we can create a halfspace between each favorite/non-favorite pair using the technique described in Section 5.1.

Figure 24 presents the results. In Figure 24 (a), both algorithms require less questions when s increases, which may credit to a faster reduction rate of utility range in Conjecture Phase since more halfspaces can be obtained in each round. The accuracies of

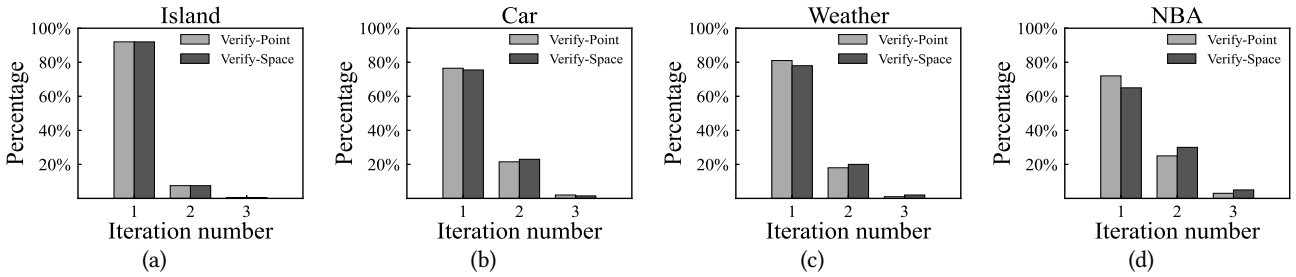


Figure 20: Percentage of iterations (real datasets)

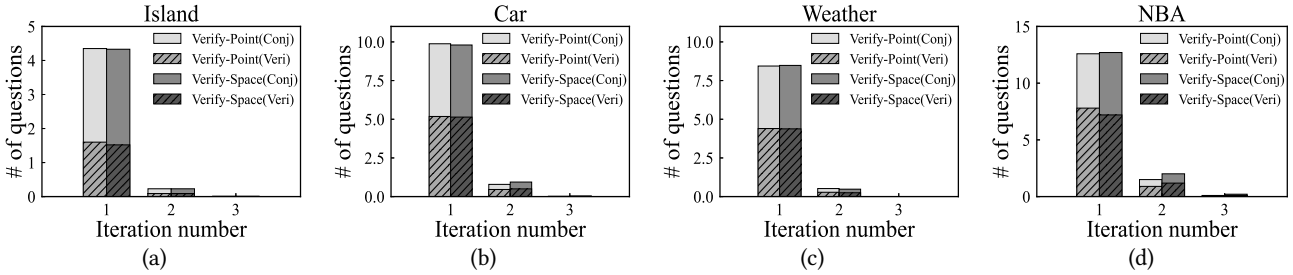


Figure 21: Average no. of questions in each phase (real datasets)

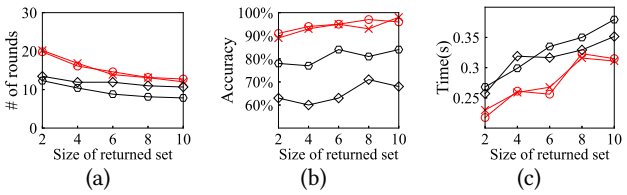


Figure 22: Varying target: returning a set of point which contains the best point

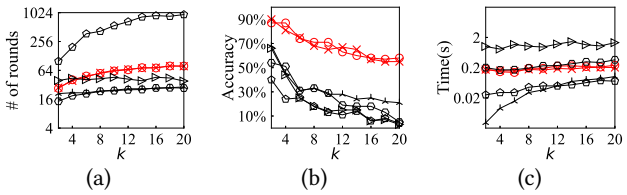


Figure 23: Varying target: returning top-k

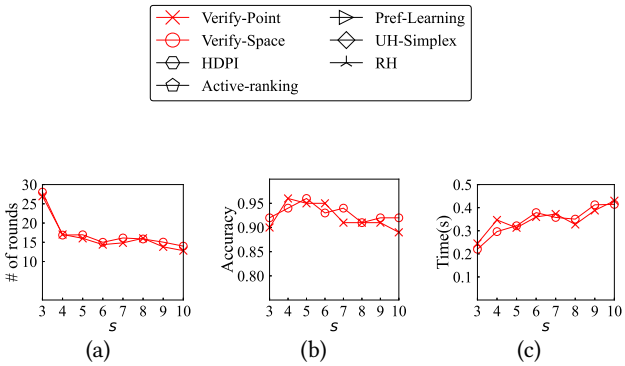


Figure 24: Varying question type: selecting favorite point from s points

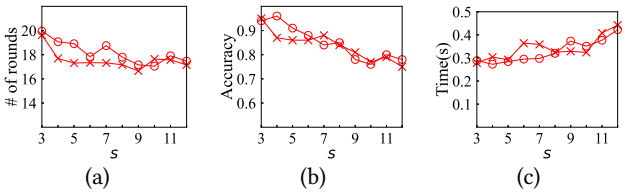


Figure 25: Varying question type: partition s points into superior and inferior group

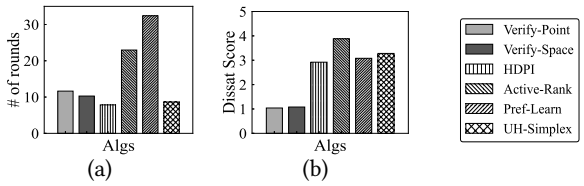


Figure 26: User study showing the no. of rounds and dissatisfaction scores

both algorithms first increase then decrease when s grows as can be observed from Figure 24 (b), which may be the composite effect of multiple factors: When s is relatively small (e.g., $s = 4$), increasing s also increases the number of halfspaces that can be used in Verification Phase, which makes the checking and pruning in Verification Phase more efficient. On the other hand, when s increases, each error will cause more serious impact since the expected number of wrong halfspaces will also increase. Therefore, when s is large (e.g., $s > 6$), the accuracies will decrease. Moreover, when the number of

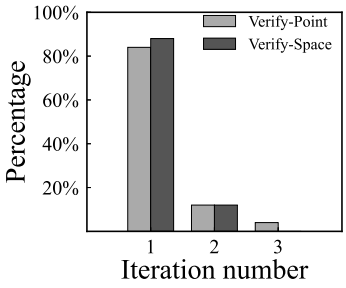


Figure 27: Percentage of iterations (user study)

halfspaces of each round increases, more time is required to update the data structures. Consequently, the average processing time also increases, as can be found in Figure 24 (c).

A.4.4 Partitioning s displayed points into superior and inferior groups. We changed the type of question asked in Conjecture Phase to the following. In each round, $s \geq 2$ points are selected from C', and the user is required to partition the displayed points into two groups, namely the superior group (i.e., those points with higher utility score) and the inferior group (i.e., those with lower utility score). To simulate this process, we first sort the selected points with decreasing utility score, then put the first l points into superior group and the rest into inferior group, where l is a random integer between 1 and $s-1$. We assume that there is at least 1 point in each group since if all points are in the same group, no information can be learnt from this question. To simulate user errors, each point is put into the wrong group with probability θ. Note that we can create a halfspace for each distinct pair of superior/inferior points using the technique described in Section 5.1.

The experimental results are shown in Figure 25. According to Figure 25 (a), when s increases, the number of questions used by *Verify-Point* and *Verify-Space* decreases since with more halfspaces derived from each question, the reduction of utility range in Conjecture Phase can be accelerated. We also observe from Figure 25 (b) that the best point retrieval accuracies decrease when s grows. When s increases, the expected percentage of wrong halfspaces also increases, which leads to the drop in accuracy. From Figure 25 (c), the average processing time increases along with s, due to the similar reason explained in Section A.4.3.

A.5 Additional experiments on user study

We show the figures related to our user study that are not presented in the main paper in this section.

Figure 26 (a) and (b) shows the average number of questions asked and the average dissatisfaction score of each algorithm involved in user study. *Verify-Point* and *Verify-Space* obtain the dissatisfaction score slightly above 1 (out of 10), which is the lowest among all algorithms. On the other hand, the dissatisfaction scores of all other competitors are at least 3. As for the number of questions asked, *HDPI* and *UH-Simplex* use the least number of rounds which is around 8. *Verify-Point* and *Verify-Space* also finish with slightly over 10 rounds. The other 2 algorithms, namely *Active-Ranking* and *Preference-Learning*, require a lot more rounds. They use on average 23 and 32 rounds, respectively.

Figure 27 presents the percentage of *Verify-Point* and *Verify-Space* that finished on each iteration in user study. Here, the bar at iteration number i shows how many percentage of user finish with i iterations. As described in Section 6.4, almost 90% of users finished within only 1 iteration on *Verify-Point* and *Verify-Space*. Only 10% and 3% of users finished in 2 and 3 iterations, respectively. Compared with the experimental results on synthetic and real datasets, the percentage of user who need more than 1 iteration is smaller in our user study, probably because the user error rate in real world (4.5% as what we found) is smaller than our setting (5%).

B RELATED PROOFS

In this section, we show the related proofs of theorems in the main paper. In addition, we also present some supporting lemmas that will be used in the proofs.

B.1 Proof of Theorem 1

We prove the correctness of Theorem 1 in this section.

Consider a 2-dimensional dataset D. The search of the best point using questions in the form " Is point p_i preferred to point p_j?" can be modeled by a binary tree with each leaf corresponding to one point in $conv(D)$. Based on p_i or p_j is selected, we follow different branches of the tree. Note that when an user error occurs, we fall into the wrong subtree and cannot return to the correct track unless the error is handled. In other words, the best point is found if and only if all errors occurred are handled sequentially.

The total number of questions required to find the best point can be divided into 2 parts. The first part is used to traverse the tree, and the second part is used to handle the user errors. Let n denote the number of points in $conv(D)$ (i.e., $n = |conv(D)|$) and m be the random variable denoting the number of errors made by the user. For the first part, in order to find the best point, $\Omega(\log n)$ questions must be asked since we must traverse $\Omega(\log n)$ steps in the tree.

We now bound the number of questions required for the second part. First, we bound the number of errors needed to be handled. Since each question has probability θ to receive a wrong answer, the expected number of errors that will occur along the correct path of the entire search is $\mu_m = \theta \log n$ and the variance is $\sigma_m^2 = \theta(1-\theta)\log n$. By applying Chebyshev inequality, $P(m < \frac{\mu_m}{2}) \leq \frac{4\theta(1-\theta)\log n}{\theta^2 \log^2 n} \leq \frac{4}{\theta \log n}$. Since θ is a constant, when n is large enough, $P(m \geq \frac{\mu_m}{2}) > C$ for some constant $C < 1$. That is, with probability C, we will encounter at least $\frac{\mu_m}{2} = \Omega(\theta \log n)$ errors.

As a running example, assume now the relation between two points, namely p_i and p_j, is $u \cdot p_i > u \cdot p_j$, that is, p_i should be preferred to p_j. However, due to a random user error, the user indicates that p_j is preferred to p_i. Since we cannot tell if the user just made an error by only looking at the answer of this question (i.e., we cannot deduce from other sources if p_i or p_j is preferred), we can only spend k questions on these two points, namely p_i and p_j, to verify their relationships and try to correct this error. Note that in order to fix this error, the majority of the answer must be correct (i.e., more than half of the answers must say that p_i is more preferred), that is, at least $\frac{k}{2}$ of the questions must be answered correctly. When k is large, we can approximate the number of correct answers in all the k answers by a Gaussian distribution

with mean $\mu = (1-\theta)k$ and variance $\sigma^2 = \theta(1-\theta)k$. Now assume that we want this error is corrected with probability at least $1-\alpha$, where α is a number to be determined later. Let z_α denote the α-quantile of $N(0,1)$, by solving $\frac{\frac{k}{2}-\mu}{\sigma} \leq z_\alpha$, we obtain $k \geq \frac{4\theta(1-\theta)}{(1-2\theta)^2} z_\alpha^2$. Since $z_\alpha = \Theta(\sqrt{\log(\frac{1}{\alpha})})$, $k = \Theta(\frac{4\theta(1-\theta)}{(1-2\theta)^2}\log(\frac{1}{\alpha}))$. (Note that this also leads to the conclusion of Lemma 1).

If we want the total failure probability to be less than δ (i.e. the probability of failing to return the best point is less than δ) with the smallest number of rounds, we need to distribute the failure probability evenly to each error. Recall that the number of errors is m. Since all errors must be handled correctly, it requires that $(1-\alpha)^m \geq 1-\delta$, which yields $\frac{\delta}{m+1} \leq \alpha \leq \frac{\delta}{m}$. Let k_m denote the value of k when there are m errors. Then, $k_m = \Theta(\frac{4\theta(1-\theta)}{(1-2\theta)^2}\log(\frac{m}{\delta}))$. Summing these up, we obtain the expected amount of questions required is

$$\begin{aligned}
\Omega(\log n) + E[mk_m] &\geq \Omega(\log n) + C(\frac{\mu_m}{2} \cdot k_{\frac{\mu_m}{2}}) \\
&= \Omega(\log n) + \Omega(\frac{4\theta^2(1-\theta)}{(1-2\theta)^2}(\log n)\log(\frac{\log n}{\delta})) \\
&= \Omega(\frac{4\theta^2(1-\theta)}{(1-2\theta)^2}(\log n)\log(\frac{\log n}{\delta}))
\end{aligned}$$

, which completes the proof.

B.2 Proof of Lemma 1

Given a user error rate θ and a desired failure probability α, Lemma 1 bounds the value of k so that the checking subroutine fails with probability at most α. Note that in the proof of Theorem 1, we already show that in order to get the correct relation between a pair of points in terms of their utility scores with probability at least $1-\alpha$, it suffices to set $k = \Theta(\frac{4\theta(1-\theta)}{(1-2\theta)^2}\log(\frac{1}{\alpha}))$. The conclusion of Lemma 1 then follows directly.

B.3 Lemma 4

In this section, we introduce Lemma 4, which will be used later to prove Theorem 2. But, before presenting this lemma, we first introduce an important concept called $Dist(\cdot,\cdot)$ that will be used later in the lemma.

Definition 1. *Given a range $R = [L,U]$ where L (resp. U) is the lower (resp. upper) bound of the range and a point p, we define the distance between the point and the range, denoted by $Dist(p,R)$, to be 0 if $p \in [L,U]$, and $\min(|p-L|, |p-U|)$ if $p \notin [L,U]$.* □

We are now ready to present Lemma 4. Lemma 4 will be used later in the proof of Theorem 2.

Lemma 4. *Given the range size $m = U - L + 1$ and the user error rate upper bound θ. Let R' be the range obtained by the binary search just after Step 1 of 2RI and p_h be the real best point. Then, $P(Dist(p_h,R') \geq \lceil 2m\theta \rceil) \leq \frac{1}{m\theta}$.* □

We now prove the correctness of Lemma 4.

Let $m = U - L + 1$ denote the size of the search range $R = [L,U]$, and θ denote the upper bound of user error rate. Further, denote p_h the real best point and $R' = [L',U']$ the search range obtained just after Step 1 of *2RI*. We first prove that $E[Dist(p_h,R')] \leq m\theta$.

The binary search terminates in $x = O(\log \frac{1}{2\theta})$ rounds, regardless of the correctness of the user's choice in each round (since when binary search terminates the size of the search range is $\lceil 2m\theta \rceil$). Therefore, we can model an execution of the algorithm as a sequence $S \in \{0,1\}^x$, where $S_i = 1$ means that the i-th question is answered correctly and $S_i = 0$ otherwise. Now consider that we attribute $Dist(p_h,R')$ to each of the x questions. Initially, $Dist(p_h,R') = 0$. Let D_i denote the increase of $Dist(p_h,R')$ "caused" by answering the i-th question, i.e., $Dist(p_h,R') = \sum_{i=0}^{x-1} D_i$. Note that if $S_i = 1$, i.e., the i-the questions is answered correctly, then $D_i = 0$. If $S_i = 0$, indicating an error when answering the i-th questions, $Dist(p_h,R')$ can increase by at most $m/2^{i+1}$. For example, if the user is wrong when answering the first question but correctly answers the remaining questions, $Dist(p_h,R')$ is at most $\frac{m}{2}$. For each $i \in [0, x-1]$, Since $P(S_i = 0) \leq \theta$, $E[D_i] \leq \frac{m\theta}{2^{i+1}}$, therefore, $E[Dist(p_h,R')] = \sum_{i=0}^{x-1} E[D_i] \leq m\theta$.

Since D_is are independent and $\sigma_{D_i}^2 = \frac{m}{2^{i+1}}\theta(1-\theta)$, we have $\sigma_{Dist(p_h,R')}^2 = \sum \sigma_{D_i}^2 = m\theta(1-\theta)$. Applying Chebyshev inequality, we obtain $P(|Dist(p_h,R') - E[Dist(p_h,R')]| > m\theta) \leq \frac{1-\theta}{m\theta} \leq \frac{1}{m\theta}$, which leads directly to the Lemma.

B.4 Proof of Theorem 2

We prove the correctness of Theorem 2 in this section.

When the context is clear, we use the index of point to denote the actual point (e.g., we use i to denote p_i which is the point at index i of the input array). Denote n the size of input (i.e., $n = |conv(D)|$) and m the size of search range at the beginning of an iteration, i.e., $m = U - L + 1$ where U(resp. L) is the upper (resp, lower) bound of the search range. We run binary search on range $[L,U]$ until the range is reduced to $R' = [L',U']$ such that $U' - L' + 1 \leq \lceil 2m\theta \rceil$. We then check if the real best point $p_h \in R'$. If not, we check if $Dist(p_h,R') \leq \lceil 2m\theta \rceil$. Based on the two checking results, there are 3 cases: (1), $p_h \in R'$ (or $Dist(p_h,R') = 0$), (2) $0 < Dist(p_h,R') \leq \lceil 2m\theta \rceil$, and (3) $Dist(p_h,R') > \lceil 2m\theta \rceil$. Note that the special case (i.e., Case (2) described in Step (2) of *2RI*) also belongs to the second case here. We say that an iteration is "good" if the checking result is in case (1) or (2) since in these two cases, we can update the search range with size $\leq \lceil 2m\theta \rceil$.

We now consider two scenarios: Scenario (1), $m\theta \geq 2$; and Scenario (2), $m\theta < 2$.

In Scenario (1), when $m\theta \geq 2$, by applying Lemma 4, the expected number of iterations from the current good iteration to the next good iterations is

$$\begin{aligned}
\frac{1}{P(Dist(p_h,R') \leq \lceil 2m\theta \rceil)} &\leq \frac{1}{P(Dist(p_h,R') \leq 2m\theta)} \\
&\leq \frac{1}{1-\frac{1}{m\theta}} \\
&\leq 1 + \frac{2}{m\theta}.
\end{aligned}$$

.

In Scenario (2), when $m\theta < 2$, note that since $E[Dist(p_h,R')] \leq m\theta$, $P(Dist(p_h,R') > \lceil 2m\theta \rceil) \leq P(Dist(p_h,R') > 2m\theta) \leq \frac{1}{2}$ by

using Markov inequality, and the expected number of iterations from the current good iteration to the next good iterations is at most 2.

We now bound the expected total iterations required. During the execution of the algorithm, $m\theta$ decreases since m decreases. Let the first good iterations where $m\theta < 2$ be the g-th good iteration (i.e., at good iteration $1, 2, \ldots, g-1$, $m\theta \geq 2$, and starting at good iteration g, $m\theta < 2$). Note that since $\theta < 0.5$, when $m\theta \geq 2$, $m \geq 4$. Since each good iteration reduces the search range size by at least a factor of $\Theta(2\theta)$, $g \leq \Theta(\log_{\frac{1}{2\theta}} \frac{n}{4})$. Denote Y the total number of iterations used by the algorithm and Y_i the number of iterations between the $(i-1)$-th and the i-th good iterations, then

$$E[Y_i] \leq \begin{cases} 1 + \frac{2}{m_i\theta} & i < g \\ 2 & i \geq g \end{cases}$$

where m_i is the size of search range at the i-th good iteration. We know that $m_i\theta \geq 2 \cdot (\frac{1}{2\theta})^{g-i-1}$ since $m_{g-1}\theta \geq 2$ and $m_i \geq (\frac{1}{2\theta})^{g-i-1} m_{g-1}$. So when $i < g$, $E[Y_i] = 1 + \frac{2}{m_i\theta} \leq 1 + (2\theta)^{g-i-1}$. Therefore,

$$\begin{aligned} E[Y] &= \sum_{i<g} E[Y_i] + \sum_{i \geq g} E[Y_i] \\ &\leq O(\log_{\frac{1}{2\theta}} n) + O(\log_{\frac{1}{2\theta}} 1) \\ &= O(\log_{\frac{1}{2\theta}} n) \end{aligned}$$

Since Step 1 of *2RI* stops when the search range size shrinks to $\lceil 2m\theta \rceil$, each iteration uses at most $O(\log \frac{1}{2\theta})$ questions (Step 1) and $3k$ extra questions for checking (Step 2). Therefore, the expected number of rounds is $O(\log_{\frac{1}{2\theta}} n) \cdot (\log \frac{1}{2\theta} + 3k) = O(\frac{k}{-\log 2\theta} \log n)$. The failure probability is upper bounded by taking union bound on the expected total number of checkings multiplied by P_k (where P_k is the failure probability of a checking subroutine with k rounds), which is $O(\frac{P_k}{-\log 2\theta} \log n)$. Since the desired failure probability is δ, it suffices to set $P_k = O(\frac{(-\log 2\theta)\delta}{\log n})$. Applying Lemma 1 yields the theorem.

B.5 Proof of Theorem 3

In this section, we show the proof of Theorem 3.

We first prove that, in the worst case, the number of iterations is $O(\frac{c}{1-\theta} \log n)$, where c is the constant appears in Theorem 3. In each round of Verification Phase, we pick the halfspace $s_{i,j} \in S$ that is expected to prune the largest number of partitions. Here, S is the output of Conjecture Phase containing all halfspaces indicated by the user. Let λ denote the average percentage of partitions pruned by the best halfspace in S (i.e., the halfspace with the highest $Num(\cdot)$). Since the utility vector u lies in this halfspace with probability at least $1-\theta$, in each round of Verification Phase, we prune at least λ portion of partitions with probability at least $1-\theta$. Since each iteration has 1 Verification Phase, and there is at least 1 round in each Verification Phase, by setting $c = \frac{1}{\log \frac{1}{1-\lambda}}$, the expected number of iterations is $O(\frac{c}{1-\theta} \log n)$.

Verification Phase ends when S is empty, or all the halfspaces in S cannot prune at least β_1 portion of partitions that are in the utility range R or intersect with R. In the worst case, Verification Phase only checks one halfspace before it is stopped, and the size of the utility range R shrinks at the slowest rate. In such a case, each iteration uses $O(conj)$ rounds (Conjecture Phase) plus k rounds (k questions need for the checking performed in Verification Phase), so the total number of rounds is $O(\frac{c}{1-\theta} \log n) \cdot O(conj + k)$. The failure probability is upper bounded by taking union bound on the total number of checkings multiplied by P_k (where P_k is the failure probability of a checking subroutine with k rounds), which is $O(\frac{cP_k}{1-\theta} \log n)$. Lemma 1 shows that by setting the total failure probability $\delta = O(\frac{cP_k}{1-\theta} \log n)$, it suffices to set $k = O(\log \frac{c \log n}{(1-\theta)\delta})$, which completes the proof.

B.6 Lemma 5

Here, we present Lemma 5, which will be used later to prove Theorem 4.

Lemma 5. *Let R_s be the utility range at the beginning of Conjecture Phase (i.e., R' just at the beginning of Conjecture Phase) and R_e be the utility range after Conjecture Phase (i.e., R' just at the end of Conjecture Phase). Let θ be the upper bound of the user error rate. Then, just after Conjecture Phase, the probability that no halfspace in S (i.e., the set containing all halfspaces indicated by the user) can prune at least half of the size of R_s is at most $\frac{vol(R_e)}{vol(R_s)} + \theta$, where $vol(R)$ denotes the size of a utility range R. Formally, $P(\frac{\min_{s \in S} vol(s \cap R_s)}{vol(R_s)} > \frac{1}{2}) \leq \frac{vol(R_e)}{vol(R_s)} + \theta$.* □

Proof. Let R_s be the utility range just at the beginning of Conjecture Phase and R_e be the utility range just at the end of Conjecture Phase. For any hyperplane h selected to ask the user in Conjecture Phase, it divides R_s into two halfspaces and at least one halfspace s satisfies $vol(s \cap R_s) \leq \frac{vol(R_s)}{2}$. If at any round in Conjecture Phase the user choose the halfspace s that satisfies $vol(s \cap R_s) \leq \frac{vol(R_s)}{2}$, then after Conjecture Phase, the best halfspace in S can reduce the size of R_s by at least half. The only possibility that after Conjecture Phase, no halfspace in S can prune p_s by at least half is when the user always picks the larger halfspace until the end of Conjecture Phase.

Consider Figure 28 as a running example. Assume that in Conjecture Phase we asked 3 questions, which are related to h_1, h_2 and h_3. Each hyperplane is related to two halfspaces (e.g., h_1 is related to two halfspaces, namely h_1^+ and h_1^-), where at least one of them can prune the utility range R_s by at least half. These 3 hyperplanes divides R_s into a number of divisions (6 in this example), and each division is the intersection of one of the halfspaces related to each hyperplane. For example, $t_1 = h_1^+ \cap h_2^- \cap h_3^-$. Among these divisions, the only case we cannot find a halfspace that prune the utility range by at least half is when the utility vector u lies in t_c (marked by gray in the figure). On the other hand, when u falls in t_1, t_2, t_3, t_4 or t_5, at least one of the halfspaces associated with that division can prune R_s by at least half.

Let B_S denote the set of divisions partitioned by the halfspaces in S on the utility range R_s. Let X_t denote the event that the utility vector u lies in division t, and Y_t denote the event that after Conjecture Phase, the algorithm decides that u lies in division t (i.e.,

$R_e = t$). Note that since there will be user errors, even if $R_e = t$ for some division t, the real utility vector u may not lie in that division. Let the only division resulting from always picking the larger subspace be t_c. Our target is to bound $P(Y_{t_c})$.

Note that

$$P(Y_{t_c}) = P(X_{t_c} \cap Y_{t_c}) + \sum_{t \in B_S, t \neq t_c} P(X_t \cap Y_{t_c})$$

Further notice that $X_{t_c} \cap Y_{t_c}$ means that the utility vector u lies in t_c and the resulting utility space $R_e = t_c$, without any distribution information on u, we can assume u is uniformly distributed in the utility range, so $P(X_{t_c} \cap Y_{t_c}) \leq P(X_{t_c}) = \frac{vol(R_e)}{vol(R_s)}$. Also notice that for any $t \neq t_c$, in order to let $X_t \cap Y_{t_c}$ happen, a necessary condition is that the user must make an error on at least 1 specific question, which is the question associated to the hyperplane dividing t and t_c. Therefore, for any $t \neq t_c, P(X_t \cap Y_{t_c}) \leq \theta \cdot P(X_t)$. We conclude that

$$\begin{aligned} P(Y_{t_c}) &\leq P(X_{t_c}) + \sum_{t \in B_S, t \neq t_c} P(X_t \cap Y_{t_c}) \\ &\leq P(X_{t_c}) + \sum_{t \in B_S, t \neq t_c} \theta \cdot P(X_t) \\ &\leq P(X_{t_c}) + \theta \sum_{t \in B_S, t \neq t_c} \cdot P(X_t) \\ &\leq \frac{vol(R_e)}{vol(R_s)} + \theta \end{aligned}$$

□

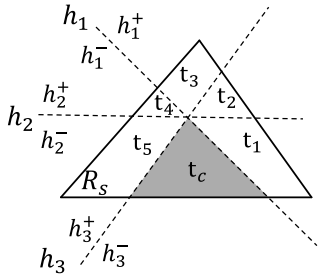


Figure 28: Divisions induced by hyperplanes

B.7 Proof of Theorem 4

In this section, we show the proof of Theorem 4. Here, we prove a special case of Theorem 4 where we set $\beta_2 = 0.5$, meaning that Verification Phase stops when the "best" halfspace in S cannot prune the utility range R by at least half. Note that in practice, β_2 can be set to another value to further improve the performance.

For any two points p_i and p_j in $conv(D)$, the related hyperplane $h_{i,j}$ cuts the utility space into two subspaces. The collection of all hyperplanes partition the utility space R_0 into many cells, where each cell corresponds to a unique ranking of the points in $conv(D)$. Recall that each iteration consists of one Conjecture Phase and one Verification Phase. Clearly, Conjecture Phase should stop before the utility range copy R' contains 1 cell, and Verification Phase should stop before the utility range R contains only 1 cell, since the best point must be already determined by then. Here, we consider the case where Conjecture Phase ends when there is only one cell left in R', and Verification Phase ends when there is only one cell left in R, or none of the halfspace in S is good enough. Note that in practice, when Conjecture Phase and Verification Phase ends, R' and R may contain multiple cells, and thus, the algorithm will terminate even faster.

[12] proved that the number of possible rankings of n points in a d dimensional space is $\Theta(\frac{n^{2d}}{2^d d!})$, which means there are $O(\frac{n^{2d}}{2^d d!})$ cells in the utility space R_0. We use Σ to denote the collection of all cells in R_0. Since *Verify-Space* ends when R contains only 1 cell, the expected size of R when the algorithm terminates is $\frac{\alpha r_0}{|\Sigma|}$ where r_0 is the size of the utility space R_0 and α is a constant depending on the size distribution of cells. Note that when $\frac{\alpha r_0}{|\Sigma|}$ is smaller, a larger portion of the utility space is expected to be pruned, and thus, more rounds are required. The worst case happens when all the cells have equal size, or equivalently, $\alpha = 1$. Consider the time when we enter the first round of a Verification Phase. let r denote the size of R and denote $P(r)$ the probability that the best hyperplane in S, the set containing all halfspaces indicated by the user in Conjecture Phase, fails to cut R of size r by at least half. Formally, $P(r) = P(\min_{s \in S} vol(s \cap R) > \frac{r}{2})$. Since each cell has the same size, there are $\frac{r}{r_0}|\Sigma|$ cells in R and Conjecture Phase ends with only 1 cell left. By Lemma 5, in the first round of this Verification Phase, $P(r) \leq \theta + \frac{r_0}{|\Sigma| r}$.

Assume that at the start of a Verification Phase, the utility range R has size $\frac{r_0}{2^{i+1}} \leq r \leq \frac{r_0}{2^i}$ where $i \in [0, \log|\Sigma|]$. Then, $P(r) = \theta + \frac{r_0}{|\Sigma|/r} \leq \theta + \frac{1}{|\Sigma|/2^{i+1}}$. Denote $T(i)$ the expected number of iterations required to reduce r from $(\frac{r_0}{2^{i+1}}, \frac{r_0}{2^i}]$ to $(\frac{r_0}{2^{i+2}}, \frac{r_0}{2^{i+1}}]$. Note that with probability $1 - P(r)$, the best halfspace in S can reduce the size of R by at least half, meaning that if this best halfspace is correct (which happens with probability at least $1 - \theta$), this single iteration will suffice to reduce r from $(\frac{r_0}{2^{i+1}}, \frac{r_0}{2^i}]$ to $(\frac{r_0}{2^{i+2}}, \frac{r_0}{2^{i+1}}]$ (or even less than $\frac{r_0}{2^{i+2}}$). Therefore, $T(i) \leq \frac{1}{1-P(r)} \cdot \frac{1}{1-\theta} \leq \frac{1}{1-(\theta + \frac{1}{|\Sigma|/2^{i+1}})} \cdot \frac{1}{1-\theta}$. There are 2 cases. Case (1), $\theta + \frac{1}{|\Sigma|/2^{i+1}} \leq \frac{1}{2}$; and Case (2), $\theta + \frac{1}{|\Sigma|/2^{i+1}} > \frac{1}{2}$.

In Case (1), when $\theta + \frac{1}{|\Sigma|/2^{i+1}} \leq \frac{1}{2}$, $T(i) \leq (1 + 2(\theta + \frac{1}{|\Sigma|/2^{i+1}})) \frac{1}{1-\theta}$. Since $i \in [0, \log|\Sigma|]$, the number of rounds required for this case is at most

$$\begin{aligned} \sum_{i=0}^{\log|\Sigma|} T(i) &\leq \frac{1}{1-\theta} \sum_{i=0}^{\log|\Sigma|} (1 + 2\theta + \frac{1}{|\Sigma|/2^{i+2}}) \\ &= \frac{1}{1-\theta} \sum_{i=0}^{\log|\Sigma|} (1 + 2\theta) + \frac{1}{1-\theta} \sum_{i=0}^{\log|\Sigma|} \frac{1}{|\Sigma|/2^{i+2}} \\ &\leq \frac{1+2\theta}{1-\theta} 2d \log n + O(1) \end{aligned}$$

In Case (2), when $\theta + \frac{1}{|\Sigma|/2^{i+1}} > \frac{1}{2}$, this implies that $r \leq \frac{r_0}{2^i} \leq \frac{4r_0}{(1-2\theta)|\Sigma|}$. Since each iteration prunes at least 1 cell from R (since

each iteration has 1 Verification Phase, each Verification Phase runs for at least 1 round and each round prunes at least 1 cell), the size of R is reduced by at least $\frac{r_0}{|\Sigma|}$. Therefore, for this case, we need at most $O(\frac{1}{(1-2\theta)})$ iterations. Summing both cases, the total expected number of iterations is $O(\frac{d}{1-2\theta}\log n)$.

Typically, Verification Phase performs several rounds of checkings before the algorithm returns to Conjecture Phase. In the worst case, it only performs one checking and returns to Conjecture Phase since none of the remaining hyperplanes is "good" enough. Note that this is the worst case since R decreases with the slowest rate. In this case, each Conjecture Phase uses $O(conj)$ rounds and each Verification Phase uses k rounds. Since the expected number of iterations is $O(\frac{d}{1-2\theta}\log n)$, the total number of rounds is thus $O(\frac{d}{1-2\theta}\log n(conj+k))$. The failure probability is upper bounded by taking union bound on the total number of checkings multiplied by P_k (where P_k is the failure probability of a checking subroutine with k rounds), which is $O(\frac{dP_k}{1-2\theta}\log n)$ using union bound. Lemma 1 shows that by setting the total failure probability $\delta = O(\frac{dP_k}{1-2\theta}\log n)$, it suffices to set $k = O(\log\frac{d\log n}{(1-2\theta)\delta})$, which completes the proof.