Scientific Research Publishing

Advances in Pure Mathematics, 2021, 11, 671-686
https://www.scirp.org/journal/apm
ISSN Online: 2160-0384
ISSN Print: 2160-0368

A Study on Computer Consciousness on Intuitive Geometry Based on Mathematics Experiments and Statistical Analysis

Xiang Sun, Zhenbing Zeng

Department of Mathematics, College of Science, Shanghai University, Shanghai, China
Email: sunx1816@shu.edu.cn, zbzeng@shu.edu.cn

How to cite this paper: Sun, X. and Zeng, Z.B. (2021) A Study on Computer Consciousness on Intuitive Geometry Based on Mathematics Experiments and Statistical Analysis. *Advances in Pure Mathematics*, **11**, 671-686.
https://doi.org/10.4236/apm.2021.118045

Received: July 9, 2021
Accepted: August 8, 2021
Published: August 11, 2021

Abstract

In this paper, we present our research on building computing machines consciousness about intuitive geometry based on mathematics experiments and statistical inference. The investigation consists of the following five steps. At first, we select a set of geometric configurations and for each configuration we construct a large amount of geometric data as observation data using dynamic geometry programs together with the pseudo-random number generator. Secondly, we refer to the geometric predicates in the algebraic method of machine proof of geometric theorems to construct statistics suitable for measuring the approximate geometric relationships in the observation data. In the third step, we propose a geometric relationship detection method based on the similarity of data distribution, where the search space has been reduced into small batches of data by pre-searching for efficiency, and the hypothetical test of the possible geometric relationships in the search results has be performed. In the fourth step, we explore the integer relation of the line segment lengths in the geometric configuration in addition. At the final step, we do numerical experiments for the pre-selected geometric configurations to verify the effectiveness of our method. The results show that computer equipped with the above procedures can find out the hidden geometric relations from the randomly generated data of related geometric configurations, and in this sense, computing machines can actually attain certain consciousness of intuitive geometry as early civilized humans in ancient Mesopotamia.

Keywords

Intuitive Geometry, Distribution Similarity, Wasserstein Distance, Mechanical Geometry Theorem-Proving

1. Introduction

Intuitive geometric knowledge is an origin of human civilization, just as shown by the Plimpton 322 tablet that people in the Old Babylonian period (between −1900 and −1600) already knew the rule of the right triangle *i.e.*, the Pythagorean theorem, through various instances of right triangles, almost one thousand years before proof was given in Greek time. From the analogue view, the machine's consciousness would be better built starting from recognizing geometric configurations, formating of geometric concepts and discovering geometric properties from observing sufficiently many examples of geometric configuration without human interference, and automated verification (or proof) of the observed geometric theorems. Indeed, machine proof of geometric theorems has been regarded as an essential subject of artificial intelligence research during the inception of artificial intelligence. In the past few decades, researchers have made significant progress in using computers to prove geometric theorems. The research work of computer proof of geometric theorems is mainly developed from the following three directions:

1) Algebraic calculation method based on coordinates;

2) Point elimination method based on geometric invariants;

3) Proving theorems by simulating human thinking the reasoning database search method.

The machine proof of geometric theorems originated in the 1950s. Tarski [1] proposed that most of the decision problems in elementary algebra and elementary geometry can be verified using an algebraic method. Among many implementations, great progress was made by Wu Wen-Tsün in the 1970s. Inspired by ancient Chinese mathematics, Wu proposed the algebraic method of geometric theorem machine proof, called the "Wu's method" [2] [3] [4]. Its basic idea is to transform a geometric problem into a system of algebraic equations, and then verify (prove or disprove) the geometric theorem by calculating the relationship between the system of algebraic equations. Wu's method has been successfully used for the mechanized proof of geometric theorems along with the rapid development of computer algebra systems like Reduce, Derive, Mathematica, Maple, and so on. Soon after Wu's initial work, the Gröbner basis method, which was developed by Buchberger for processing polynomial system in the 1960s, has also been widely used in the field of geometric theorem proving [5] [6]. Both Wu's method and Gröbner basis method are essentially verifying algebraic identities with some constrained variables and a set of polynomial constrained equations. Starting from the fact that the lower and upper bounds of a polynomial equation can be determined by its coefficients, Hong [7] proposed the "one-example illustration method" that can verify the correctness of a geometric theorem via a single instance of the related geometry statement. Furthermore, based on the following observation: if a multivariate polynomial has a value equal to zero on a sufficiently large grid, then this polynomial is always equal to zero, Zhang *et al.* proposed the "numerical parallel method" [8], which passed a certain scale of

examples to verify whether the given polynomial is an identity. As Hong's method usually involves constructing a very complicated example and computation of too large objects, Hong's method had never been used in any non-trivial theorem, meanwhile, the parallel numerical method is the first numerical algorithm with practical and feasible significance in the machine proof of geometric theorems as it was easily implemented in portable computer (like CASIO's PB700 or Sharp PC1500) at the end of 1980s.

When the above-mentioned algebraic methods are used to prove geometric theorems, they usually include large-scale complicated calculations involving polynomials, which geometric meaning generally can't be understood by human, and for human it is also too difficult to check the correctness of the machine computation by manual method. Therefore, such proofs are called "human non-readable". Zhang *et al.* [9] proposed to use the area method to prove the geometric theorem and realized the readable proof for the first time. Zhou *et al.* [10] introduced the Pythagorean difference to the proof process of Non-Euclidean geometric theorems. Similar to the area method and the Pythagorean difference method, a generalized vector method was suggested in [11]. These methods are collectively referred to as the "geometric invariant method" [12] [13].

Another category method, the "deductive reasoning method" based on database searching, which simulates the idea of human proof of geometric theorems, namely, using known hypotheses and standard axioms to perform inference searches on geometric propositions, can be traced to 1960. Gelernter *et al.* [14] proposed a method that combined the backstepping method with the depth-first search and implemented a program based on the backstepping method on the computer. Nevins [15] combined the forward and backstepping method to prove the geometric theorem. Zhang *et al.* [16] gave a more effective method based on a geometric deduction database system. Based on the idea of a structured database, the amount of calculation in the inference process was significantly reduced, and it proves that generate geometric propositions are generally readable. It worths indicating that together with dynamic geometry programs (like Geometer's Sketchpad), the deductive reasoning method has been widely used for developing educational software in China. Nevertheless, there has no report on studies to promote computers to obtain graphical intuitive analysis capabilities for elemental geometry yet.

Considering that the intuitive knowledge of geometry played the essential role in the development of human intelligence—in both meaning of humankind and human individuals, it is natural to expect that computing machines that are able to see or understand certain geometry meaning, like three-point collineance, four points lie on the same circle, or square of one edge equals to the sum of squares of other two edges in certain triangles, would eventually lead to a higher stage of machine intelligence—the ASI (Artificial Super Intelligence), Sun studied recently in his Master thesis [17] the problem to train the intelligent agents such as computers to "observe" a large number of intuitive geometric configurations, to combine the powerful algebraic computing capabilities and data storage

capabilities of machine, so to understand and master the intuitive geometrical analysis capabilities of humans in the long-term goal of AI. The work implemented a symbolic computation program with Maple software to mimic dynamic geometry for randomly generating geometric configuration in batch, and designed several statistical formulas to discover latent geometry relationships from suitable amount of graphic data, therefore, exhibited a potential probability of the conscious evolution of the computing machine species.

As an English translation of one part of the thesis, this paper focuses on establishing statistics of geometric relations in graphic data and establishing a quantitative method for comparing the similarities between the distributions of graphic data.

The rest of this paper is organized as follows. In Section 2, we introduce the geometric theorem machine proof methods related to the content of this paper. In Section 3, we propose the geometric relationship detection method based on distribution similarity. In Section 4, we conducted numerical experiments and compared the results under different observation error levels. In the final section, we draw a short conclusion.

2. Related Methods of Mechanical Geometry Theorem-Proving

2.1. Wu's Method

Let F and G be two multivariate polynomials about the variable $\boldsymbol{x}$, the class of F is k, and the highest degree of F and G about x_k are d and s respectively. Arrange F and G in descending order of the variable x_k and write as follows form:

$$\begin{cases} F = f_d x_k^d + f_{d-1} x_k^{d-1} + \cdots + f_0 \\ G = g_s x_k^s + g_{s-1} x_k^{s-1} + \cdots + g_0 \end{cases} \tag{1}$$

Then, there must be a non-negative integer t and polynomials T and R. The highest coefficient of R with respect to x_k is less than d or $R = 0$, which satisfies:

$$f_d^t \times G = T \times F + R \tag{2}$$

In Equation (2), R is the pseudo remainder of polynomial G with respect to polynomial F, denoted as $prem(G, F) = R$.

If the polynomial group $TS = T_1, T_2, \cdots, T_s$ satisfies $s = 1$, $T_1 \neq 0$ or $\forall i < j$, $class(T_i) < class(T_j)$, then the polynomial group TS of the following form is called a triangular polynomial group:

$$\begin{cases} T_1(x_1, x_2, \cdots, x_{i1}) \\ T_2(x_1, x_2, \cdots, x_{i1}, x_{i2}) \\ \cdots \\ T_s(x_1, x_2, \cdots, x_{i1}, \cdots, x_{is}) \end{cases} \tag{3}$$

Assuming that $TS = T_1, T_2, \cdots, T_s$ is a triangular polynomial group, the remainder of polynomial G with respect to TS can be obtained by the following

"continuous pseudo division":

$$\begin{cases} prem(G,T_s)=R_s \\ prem(R_s,T_{s-1})=R_{s-1} \\ \cdots \\ prem(R_2,T_1)=R_1 \end{cases} \tag{4}$$

Let $R=R_1$, and write Equations (4) as $prem(G,TS)=R$. Further, the remainder formula Equation (2) can be extended to the following form:

$$I_1^{t_1}\times\cdots\times I_s^{t_s}\times G=\sum_{i=1}^{s}C_i\times T_i+R \tag{5}$$

Among them, I_i and C_i are the initial formula and polynomial of T_i respectively.

The general procedure of Wu's method to prove geometric theorem is as follows:

1) The geometric theorem is algebraized, the known assumptions are partially transformed into a polynomial group *H*, and the theorem's conclusion is transformed into a polynomial *g*.

2) The polynomial group *H* is sorted according to the Wu-Ritt principle [2] [18], and the ascending $CS=\{f_1=0,f_2=0,\cdots,f_s=0\}$ is obtained.

3) Solve the theorem conclusion polynomial g and the continuous pseudo-division of ascending sequence, get $prem(g,CS)=R$, and judge whether the residue *R* is 0. If $R=0$, according to Equation (5), it is easy to get the equation $I_1^{t_1}\times\cdots\times I_s^{t_s}\times g=C_1f_1+C_2f_2+\cdots+C_sf_s$, and $g=0$ can be derived from the non-degenerate condition $I_i^{t_i}\neq 0$ and $f_i=0$. From this, we can know that the theorem to be proved holds under non-degenerate conditions.

2.2. Numerical Parallel Method

The single-example illustration method has expanded a new idea for the machine proof of geometric theorems, but it has not been realized due to its high computational complexity. Zhang *et al.* [8] proposed a numerical parallel method inspired by Wu's method.

Suppose the polynomial $F(x_1,x_2,\cdots,x_n)\in\mathbb{K}[x_1,x_2,\cdots,x_n]$, the highest degree of the polynomial *F* to the variable x_i is less than or equal to $d_i\,(i=1,2,\cdots,n)$, and $S_i\,(i=1,2,\cdots,n)$ is an arbitrary subset of d_i+1 elements in the domain $\mathbb{K}$. If the following Equation (6) holds, *F* is an identity that is always 0:

$$\forall(x_1^*,x_2^*,\cdots,x_n^*)\in S_1\times S_2\times\cdots\times S_n, F(x_1^*,x_2^*,\cdots,x_n^*)=0 \tag{6}$$

The conclusion can be drawn from the above: To verify whether an n-ary polynomial $F(x_1,x_2,\cdots,x_n)$ with the highest degree of each variable of $d_i\,(i=1,2,\cdots,n)$ is an identity that is always 0, only *N* different numerical examples need to be verified, where $N=(d_1+1)\times(d_2+1)\times\cdots\times(d_n+1)$.

The general procedure of Wu's method to prove geometric theorem is as follows:

1) The geometric theorem is algebraized, the known assumptions are partially transformed into a polynomial group H, and the theorem's conclusion is transformed into a polynomial g.

2) Solve the triangle polynomial set TS reduced by the polynomial set H. Solve the conclusion of the polynomial g for the remainder of TS, $R = prem(g, TS)$. Estimate the maximum degree $d_i\ (i = 1, 2, \cdots, n)$ of the remainder R for the independent variable.

3) According to Equation (6), construct the set of instances to be tested and substitute these instances into TS one by one, solve the specific values of the constraint variables, and then substitute them into g. If $g = 0$, it indicates that the instance is consistent with the theorem; Otherwise, this geometric theorem is generally invalid.

3. Intuitive Geometry Based on Experimental Mathematics and Statistical Analysis

3.1. Data and Statistics

The algebraic methods such as Wu's method, single-example illustration method, and numerical parallel method prove geometric theorems. It is necessary to algebraize the geometric theorems. We propose to calculate the numerical value of the geometric configuration instance without algebraic processing, so it needs to generate a large number of geometric configuration legends. Data can be generated by changing the free points in the geometric configuration. We use Maple to write a dynamic geometry subroutine module similar to the geometric sketchpad and super sketchpad to realize the data generation of the geometric configuration.

The algebraic method proves the geometric theorem, and a polynomial $f(x_1, x_2, \cdots, x_n) = 0$ expresses the geometric relationship by selecting appropriate coordinates. Our Maple program simulates the intelligent subject to observe the geometric configuration intuitively, adding slight disturbances to the data and rounding the coordinates of the points. In this way, it is not possible to directly use the polynomial $f(x_1, x_2, \cdots, x_n) = 0$ to express the geometric relationship. In analogy to the geometric predicate in the algebraic method, we have constructed relevant statistics to express the geometric relationship.

The construction of statistics satisfies the following three principles:

1) $f = 0$, if and only if a particular geometric relationship is strictly valid numerically, the degree of approximate validity of a particular geometric relationship is measured by the degree of deviation from 0.

2) Statistics should eliminate the influence of dimensions.

3) For N samples Equation (7) that satisfy a particular geometric relationship, satisfy Equation (8)

$$A^{(i)} = \left(u_1^{(i)}, v_1^{(i)}\right), B^{(i)} = \left(u_2^{(i)}, v_2^{(i)}\right), C^{(i)} = \left(u_3^{(i)}, v_3^{(i)}\right), \cdots, i = 1, 2, \cdots, N \tag{7}$$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} f\left(A^{(i)}, B^{(i)}, C^{(i)}, \cdots\right) \to 0 \tag{8}$$

We briefly explain the construction of statistics corresponding to commonly used geometric relations.

Measure equilateral triangle: use Equation (9) to measure.

$$f = \arg\max_i \left(|i| - \frac{\pi}{3} \right), i = \angle A, \angle B, \angle C \tag{9}$$

Measurement angle bisector: Measured by the difference between the two angles formed by dividing the angle by a straight line. Measure vertical (or parallel): the value is measured by the difference between the angle of two vectors and $\frac{\pi}{2}$ (or 0), and the outer product of the vector measures the sign. The three points are measured in common: the value is measured by the farthest distance among the three points, and the directed area of a triangle measures the sign. Measure the collinearity of three points: use the smallest angle between the three points in the vector. Measure multiple points in a circle: fit a circle $(x-x_0)^2 + (y-y_0)^2 = r_0^2$ closest to these N points by the least square method, and then use Equation (10) to measure.

$$f = \left(\arg\max_{r_i} \left(|r_i - r_0| \right) \right) \Big/ r_0 - 1 \tag{10}$$

3.2. Distribution Similarity Geometric Relationship Detection

In this section, we propose a geometric relationship detection method based on distribution similarity. Before that, let me introduce the methods of measuring the similarity of distributions and the nonparametric test methods used in this paper.

Considering the similarity of two probability distributions, P and Q, Kullback-Leibler divergence (KL divergence) in Equation (11) and Jensen-Shannon divergence (JS divergence) in Equation (12) can be used.

$$D_{KL}(P \| Q) = \int_x p(x) \log \left(\frac{p(x)}{q(x)} \right) \mathrm{d}x \tag{11}$$

$$D_{JS}(P \| Q) = \frac{1}{2} D_{KL}(P \| M) + \frac{1}{2} D_{KL}(Q \| M), M = \frac{P+Q}{2} \tag{12}$$

When the support sets of the two distributions, P and Q, do not overlap or the overlap is small, it is difficult for KL divergence and JS divergence to quantify the similarity between the distributions. In recent years, the similarity of the Wasserstein distance Equation (13) metric distribution has been widely used in machine learning. In this paper, we use Wasserstein distance to measure the similarity of distributions. In Equation (13), $\gamma(x, y)$ satisfies $\int_{\mathbb{R}^d} \gamma(x, y) \mathrm{d}y = P$ and $\int_{\mathbb{R}^d} \gamma(x, y) \mathrm{d}x = Q$. In general, it is tough to calculate the Wasserstein distance, but when the data dimension $d = 1$, the p-Wasserstein distance can be expressed as Equation (14). Among them, F^{-1} and G^{-1} are the quantile functions of P and Q, respectively.

$$W_p(P,Q)=\left(\inf_{\gamma\in\Gamma(P,Q)}\int_{\mathbb{R}^d\times\mathbb{R}^d}\|x-y\|^p\,\gamma(x,y)\right)^{1/p} \tag{13}$$

$$W_p(P,Q)=\int_0^1\left|F^{-1}(t)-G^{-1}(t)\right|^p\,\mathrm{d}t \tag{14}$$

In this way, the calculation of p-Wasserstein distance is simplified. In the previous section, we constructed the statistics of geometric relations and mapped the observation data to one dimension to use 1-Wasserstein distance Equation (15) to measure similarity.

$$W_1(P,Q)=\int_0^1\left|F^{-1}(t)-G^{-1}(t)\right|\mathrm{d}t \tag{15}$$

In statistics, hypothesis testing is often used in statistical inference, inferring hypotheses about the population based on empirical data. It can also be used to test whether two distributions come from the same distribution. In this paper, we used two non-parametric tests. One is the Kolmogorov-Smirnov (K-S) test, which uses the K-S statistic Equation (16) or Equation (17) to accept or reject the null hypothesis.

$$D_n=\sup_x\left|F_n(x)-F(x)\right| \tag{16}$$

$$D_{n,m}=\sup_x\left|F_{1,n}(x)-F_{2,m}(x)\right| \tag{17}$$

Another method is referred to as the permutation test based on the 2-Wasserstein distance used in Matsui *et al.* [19] and Schefzik *et al.* [20]. Considering the Wasserstein distance when $d=1$ and $p=2$, it can be decomposed into three parts [21] in Equation (18).

$$\begin{aligned}W_2(P,Q)&=\int_0^1\left|F^{-1}(t)-G^{-1}(t)\right|^2\mathrm{d}t\\&=\left(\mu_p-\mu_q\right)^2+\left(\sigma_p-\sigma_q\right)^2+2\sigma_p\sigma_q\left(1-\rho_{p,q}\right)\end{aligned} \tag{18}$$

Among them, the mean, variance and shape of the three items on the right are respectively distributed. $\rho_{p,q}$ is the Pearson correlation coefficient of the corresponding point in the quantile map of F and G. Equation (18) can be approximated by the empirical estimation formula Equation (19).

$$W_2\left(\hat{P},\hat{Q}\right)=\int_0^1\left|\hat{F}^{-1}(t)-\hat{G}^{-1}(t)\right|^2\mathrm{d}t \tag{19}$$

$\hat{P}$ and $\hat{Q}$ are the distribution of the observation data $X(i),Y(i),i=1,2,\cdots,n$, and $\hat{F}^{-1}(t)$ and $\hat{G}^{-1}(t)$ are the quantile functions of the observation data. The null hypothesis is $H_0:P=Q$. Under the condition that the null hypothesis is established, the distribution functions P and Q are obtained by random replacement of samples. Calculate the distance $d_i^*=W_2\left(\hat{P}_i^*,\hat{Q}_i^*\right)$ according to Equation (18), mark $D_i^*=\left(d_{i,1}^*,d_{i,2}^*,\cdots,d_{i,B}^*\right)$ as the distances between the two distributions after all possible permutations. Then the p-value can be calculated according to Equation (20), where $d_i=W_2\left(\hat{P},\hat{Q}\right)$. The subset $D_{i,sub}^*=\left(\tilde{d}_{i,1}^*,\tilde{d}_{i,2}^*,\cdots,\tilde{d}_{i,B_s}^*\right),B_s<B$ of D_i^* can approximate Equation (20) to reduce the amount of calculation, and Equation (21) can be obtained.

$$p = \frac{\sum_{b=1}^{B} I\left(d_{i,b}^{*} \geq d_i\right)}{B} \tag{20}$$

$$p_s = \frac{\sum_{b=1}^{B} I\left(\tilde{d}_{i,b}^{*} \geq d_i\right)}{B_s} \tag{21}$$

The geometric relationship detection method based on distribution similarity mainly includes the following steps:

Step 1: Call the Maple subroutine to generate the corresponding geometric configuration legend, and generate a large sample data according to the dynamic geometry.

Step 2: Randomly select a batch of samples, and measure the similarity according to Equation (15). Under fixed disturbance δ, construct the standard distribution $P_{(\delta,f)}$ of each geometric relationship to measure the observation data distribution $P_{(emp,f)}$. Among them, f is a statistic that measures geometric relations.

Step 3: In Step 2, reduce the search range according to the 1-Wasserstein distance, and perform a significance test on the remaining distributions with high similarity. When $P_{(\delta,f)}$ and $P_{(emp,f)}$ approximately obey the normal distribution, use the T-test and the F-test to test the position and scale parameters of the normal distribution, respectively. Otherwise, the permutation test is used for the non-parametric test. Use the K-S method to test whether $P_{(\delta,f)}$ and $P_{(emp,f)}$ approximately obey a normal distribution.

The complete process can be found in Algorithm 1.

Algorithm 1 Distribution similarity geometric relationship detection

Input: Geometric instance data, D; error, δ.
Output: Approximate geometric relationship, R.

$R = list()$,c is the combination.
Construct $P_{(\delta,f)}$.
while c not exhausted **do**
  $P_{(emp,f)} \leftarrow$ calculate $f(c)$,**repeat**.
  calculate $W_1(P_{(emp,f)}, P_{(\delta,f)})$ according to Equation (15),measure similarity.
  if $P_{(emp,f)}, P_{(\delta,f)}$ approximately normal distribution **then**
    Parameter test
  else
    Nonparameteric test
  end if
  hypothesis test accepted r,$R.append(r)$.
  change c.
end while
Return R.

3.3. Integral Coefficient Invariant Discovery

In Section 3.2, we propose a geometric relationship detection method based on distribution similarity to explore the deterministic vertical and collinear geometric relationships in geometric configurations. In this section, we explore the integer coefficient relationship between the lengths of geometric quantities. This

type of relationship involves uncertain, unknown quantities and uncertain integer coefficients. Since most of the relations between geometric quantities are homogeneous relations of first and second order in plane geometry, we study the first and second integer coefficient relations between the length of line segments in geometric figures. Specifically, there is a vector $\boldsymbol{x}=\left(x_1,x_2,\cdots,x_n\right)\in\mathbb{R}^n$, and a group of integers $a_1,a_2,\cdots,a_n$ that is not all 0 is found to satisfy Equation (24).

$$a_1x_1+a_2x_2+\cdots+a_nx_n=0 \tag{22}$$

At present, the widely used integer relational algorithms are mainly the LLL algorithm proposed by Lenstra *et al.*, and the PSLQ algorithm proposed by Bailey *et al.* In addition, Feng *et al.* [22] researched a PSLQ algorithm with empirical data as input. Our data is observational data, and the accuracy of the data does not meet the requirements of these algorithms, so we try to solve it in the following way.

First, we converted the sample data, and the generated data is uniformly expressed as an array of points P Equation (23),

$$P=\left[p_1:\left(x_1,y_1\right),p_2:\left(x_2,y_2\right),\cdots,p_n:\left(x_n,y_n\right)\right] \tag{23}$$

where the order of points p_i in P is fixed. There is a line between any two points by default, and all the line segment lengths in the geometric figure D Equation (24) are obtained, where d is distance.

$$D=\left[d\left(p_1,p_2\right),d\left(p_1,p_3\right),\cdots,d\left(p_{n-1},p_n\right)\right] \tag{24}$$

Extending it to quadratic and the reciprocal of the geometric quantity get D_2 Equation (25) and D_{-1} Equation (26).

$$D_2=\left[\left(d\left(p_1,p_2\right)\right)^2,d\left(p_1,p_2\right)*d\left(p_1,p_3\right),\cdots,\left(d\left(p_{n-1},p_n\right)\right)^2\right] \tag{25}$$

$$D_{-1}=\left[1/d\left(p_1,p_2\right),1/d\left(p_1,p_3\right),\cdots,1/d\left(p_{n-1},p_n\right)\right] \tag{26}$$

Then, we randomly select no less than m converted samples and write them as $\boldsymbol{A}_{m\times m}$ in the form of a matrix, and write the integer coefficient vector to be solved as $\boldsymbol{x}_{m\times 1}$, $\boldsymbol{x}\in\mathbb{Z}^m,\boldsymbol{x}\neq\boldsymbol{0}$, where m is the number of elements in D or D_2. If the length of the line segment in the geometric configuration has an integral coefficient relationship, then there is such a $\boldsymbol{x}$ that satisfies $\boldsymbol{Ax}=\boldsymbol{0}$. Due to the observation error of the sample data, the problem is transformed into Equation (27).

$$\arg\min_{\boldsymbol{x}}\frac{1}{m}\left\|\boldsymbol{Ax}\right\|_2^2,\boldsymbol{x}\in\mathbb{Z}^m,\boldsymbol{x}\neq\boldsymbol{0} \tag{27}$$

Usually, in plane geometry, the integer coefficients between geometric quantities are relatively small. Due to the existence of such prior knowledge, we add regular term constraints based on Equation (27) to obtain Equation (28).

$$\arg\min_{\boldsymbol{x}}\frac{1}{m}\left\|\boldsymbol{Ax}\right\|_2^2+\lambda\left\|\boldsymbol{x}\right\|_2,\boldsymbol{x}\in\mathbb{Z}^m,\boldsymbol{x}\neq\boldsymbol{0} \tag{28}$$

In addition, in geometric figures, the integer coefficient relationships that exist are not unique. In Equation (28), integer coefficient relationships involving a

small number of geometric quantities will conceal the relationship involving more integer coefficients involving geometric quantities. Consider decomposing Equation (28) into sub-problems Equation (29).

$$\arg\min_{\boldsymbol{x}} \frac{1}{m}\left\|\boldsymbol{A}\boldsymbol{x}\right\|_2^2 + \lambda\left\|\boldsymbol{x}\right\|_2, x_i \geq 1\left(1 \leq i \leq m\right), \boldsymbol{x} \in \mathbb{Z}^m \tag{29}$$

Since the elements in D, D_{-1}, and D_2 are always non-negative, if $\boldsymbol{x}$ satisfies Equation (28), then at least two components of $\boldsymbol{x}$ are not 0 and are integers. We let a component x_i of $\boldsymbol{x}$ add constraints to Equation (28) to get Equation (29) to search sequentially.

The solution of Equation (29) is complex. We use the Cplex solver developed by IBM to solve this problem. The algorithm steps are shown in Algorithm 2, where the ***err*** refers to the sum of the difference between the approximate integral relationship and the strict integral relationship and the regular term.

Algorithm 2 Integral coefficient invariant discovery

Input: Geometric instance data, D; degree d; regular term coefficient λ(default=1).
Output: Approximate integer relationship, $\boldsymbol{x}$; error, $\boldsymbol{err}$.

$D_d \leftarrow d, m = dim(D_d)$.
$\boldsymbol{A}_{m\times m} \leftarrow$ randomly select m instances.
$j = 1$
for $i = 0$ to m **do**
  if $\underset{\boldsymbol{x}}{arg\min} \frac{1}{m}\|\boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_2$ subject to $x_i \geq 1, \boldsymbol{x} \in \mathbb{Z}^m$ solvable **then**
    x_j is the solution in procedure 5, $err_j = \frac{1}{m}\|\boldsymbol{A}\boldsymbol{x_j}\|_2^2 + \lambda\|\boldsymbol{x_j}\|_2$, $j+ = 1$.
  end if
end for
Return $\boldsymbol{x}, \boldsymbol{err}$.

4. Numerical Experiment

In this section, we construct observation data of some geometric theorems and performed numerical experiments on this basis, including 12 geometric theorems such as Orthocenter theorem, Centroid theorem, Incenter theorem, Morley theorem, Euler Line, Five Circles theorem, Nine-point Circle theorem, Ptolemy's theorem, corollary to Ptolemy's theorem, Candy theorem, Pappus theorem, and Desargue theorem. We first carry out numerical experiments under $\delta = N(0,2)$ disturbances and then carry out numerical experiments with different disturbance levels of $\delta = N(0,1)$ and $\delta = N(0,3)$, where $N(0,2)$ is normal distribution.

Take the Orthocenter theorem to illustrate a feasible method of selecting thresholds to reduce the search space. Figure 1 is the Wasserstein distance of the geometric relationship of the three-point collinear relationship in the Orthocenter theorem. The Wasserstein distance is naturally divided into two categories. The right part does not satisfy the three-point collinear relationship, so the data of these combinations can be quickly excluded. In the Orthocenter theorem, the perpendicular, three-point common point relationship test is the same, and the hypothetical test of the result after rapid elimination is performed. The Orthocenter theorem, the empirical distribution of the perpendicular relationship

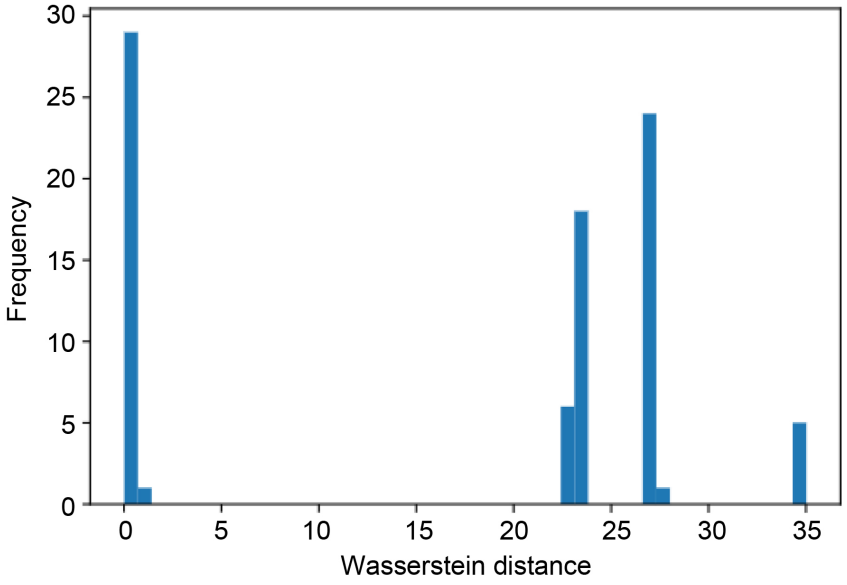

Figure 1. Wasserstein distance of the three-point collinear relationship in the Orthocenter theorem.

statistics of the three vertical lines, and the standard distribution of the perpendicular relationship statistics under $\delta = N(0,2)$ disturbance are shown in **Figure 2**. A further hypothetical test is performed on the perpendicular relationship of the three vertical lines and the significance level a = 005. Because the empirical distribution and the standard distribution are approximately normal distributions, the parameter test is used directly. The results are shown in **Table 1**.

Since there are many collinear and perpendicular relationships, it is too verbose to list them all. Here are examples of each type of geometric relationship. The results are shown in **Table 2**, where 1-Wd means 1-Wasseratein distance.

We take Ptolemy's theorem, the corollary of Ptolemy's theorem, and Candy theorem as examples to carry out numerical experiments on the invariants of integral coefficients. The first is Ptolemy's theorem, 100 samples are randomly selected to solve, and three effective solution vectors Equation (30) are obtained. These three solution vectors correspond to the same geometric relationship, which is the conclusion of Ptolemy's theorem. We re-selected 100 samples for 10 experiments, and the errors obtained were 94.3702, 78.3523, 59.2743, 31.8584, 76.5764, 81.6923, 43.1427, 69.1004, 118.8367, 76.9757.

$$\begin{cases} \boldsymbol{x}_1 = (0,0,0,0,0,1,0,0,0,-1,0,0,1,0,0,0,0,0,0,0,0) \\ \boldsymbol{x}_2 = (0,0,0,0,0,-1,0,0,0,1,0,0,-1,0,0,0,0,0,0,0,0) \\ \boldsymbol{x}_3 = (0,0,0,0,0,1,0,0,0,-1,0,0,1,0,0,0,0,0,0,0,0) \end{cases} \tag{30}$$

The second is the corollary of Ptolemy's theorem, where the result of the corollary is in an order relationship with the size of the geometric value, and the length of the line segment is sorted before starting the solution. Similarly, 100 samples are randomly selected and solved to obtain the solution vector Equation (31).

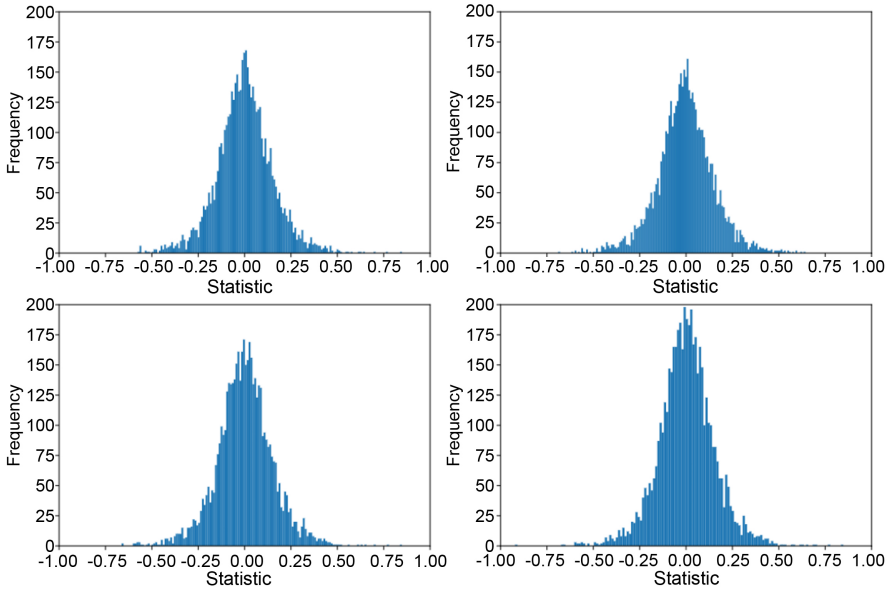

Figure 2. The distribution of the perpendicular relationship statistics of the vertical lines, the lower right corner subfigure is the standard distribution and the rest are the empirical distribution.

Table 1. Hypothesis test of perpendicular relationship ($\alpha = 0.05$).

p-value of T-test	p-value of F-test
0.1097	0.2907
0.7752	0.3685
0.6879	0.4565

Table 2. Geometric relationship detection results based on distribution similarity ($\delta = N(0,2), \alpha = 0.05$).

name	conclusion	1-Wd	hypothesis testing type	p-value
Orthocenter	three points in common	0.1446	Non-parametric test	0.8879
Morley	form an equilateral triangle	0.0104	Non-parametric test	0.8623
Five Circles	five points round	0.1328	Non-parametric test	0.3572
Nine-points Circle	nine points round	0.1475	Non-parametric test	0.4741
Pappus	three points collinear	0.0954	Non-parametric test	0.8186

$$\begin{cases} \boldsymbol{x}_1 = (1,1,0,0,0,-1) \\ \boldsymbol{x}_2 = (1,1,0,0,0,-1) \\ \boldsymbol{x}_3 = (0,0,1,-1,0,0) \\ \boldsymbol{x}_4 = (0,0,0,1,-1,0) \\ \boldsymbol{x}_5 = (0,0,0,-1,1,0) \\ \boldsymbol{x}_6 = (-1,-1,0,0,0,1) \end{cases} \tag{31}$$

Finally, in the numerical experiment of Candy theorem, the theorem's conclu-

sion could not be obtained. If we expand the conclusion of Candy theorem, we get a cubic relationship. It is tough to solve the cubic relationship. The dimension of the solution vector $\boldsymbol{x}$ will be increased very largely, and the error will also accumulate due to the multiplication of each item.

The comparative experimental results of the other two groups of different disturbances can be seen in Table 3 and Table 4. The error of the integral coefficient invariant relationship in Table 4 is taken from the average of the results of 10 experiments, and "–" means that the result of the theorem is not obtained.

5. Conclusion

In this paper, we construct statistics that measure approximate geometric relationships for inaccurate observation data, map the observation data to one dimension through statistics. Using the distribution similarity of the Wasserstein distance metric, we propose a method for detecting geometric relationship similarities. The method has been successfully applied for checking the following geometric relations: 1) three lines intersect at one point; 2) three points lie on one line; 3) three points form one equilateral triangle; 4) two lines are parallel or perpendicular to each other; 5) one line bisects a given angle; 6) four points form a convex quadrilateral; and 7) four or more points lie on the same circle. We have also proposed a searching method to find linear or quadratic equations with integer coefficients between observed geometric quantities under certain prior conditions. The constrained searching method can be used to find linear or quadratic relation with integer coefficients between geometric quantities under a priori conditions. Numerical experiments show that the method proposed in this

Table 3. Geometric relationship detection results ($\delta = N(0,1)$ and $\delta = N(0,3)$, $\alpha = 0.05$).

name	1-Wd ($\delta = N(0,1)$)	1-Wd ($\delta = N(0,3)$)	p-value ($\delta = N(0,1)$)	p-value ($\delta = N(0,3)$)
Orthocenter	0.0946	0.3060	0.6410	0.7904
Morley	0.0118	0.0161	0.7582	0.3920
Five Circles	0.0195	0.0293	0.5596	0.6654
Nine-points Circle	0.1408	0.1846	0.2574	0.1939
Pappus	0.0392	0.0705	0.7406	0.7195

Table 4. Integral coefficient relation error under different disturbance levels.

δ	Ptolemy's theorem	corollary of Ptolemy's theorem	Candy theorem
$\delta = N(0,1)$	6.9082	1.9009	-
$\delta = N(0,2)$	73.0179	1.9890	-
$\delta = N(0,3)$	229.6791	2.0426	-

paper is adequate, which will help the machine to obtain intuitive analysis capabilities for geometric figures, which have practical significance and certain application prospects. Our experiment failed in finding a cubic relation in the CandyTheorem, namely, assume that AB is an arbitrary chord in the circle O, P is a point on AB, C, D are arbitrary two points on the circle O, E, F are intersection points of the circle O and line CP, DP, and G, H are intersection points of the chord AB with CF, DE, respectively, then $1/AP - 1/BP = 1/GP - 1/HP$. We will try to overcome this difficulty by control error accumulation in numerical analysis. An interesting problem is to train computers to find latent inequalities from configuration data. A very simple and famous example of such kind is Euler's Inequality, which states that the distance d between the incenter r and the circumcenter R of a triangle satisfies

$$d^2 = R(R-2r),$$

and therefore $R \geq 2r$. Since almost interesting theorems that involved equalities in Euclidean geometry have been well established in past three thousand years, a prospective application of machine intelligence in the future would be automated discovering of geometric inequalities through analyzing big data of geometric configurations.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Tarski, A. (1951) A Decision Method for Elementary Algebra and Geometry. In: Johannes, K., Ed., *Texts & Monographs in Symbolic Computation*, Springer, Berlin, 24-84. https://doi.org/10.1007/978-3-7091-9459-1_3

[2] Wu, W.-T. (2012) Mechanical Theorem Proving in Geometries: Basic Principles. Springer-Verlag Wien, Wien.

[3] Chou, S.C. (1988) An Introduction to Wu's Method for Mechanical Theorem Proving in Geometry. *Journal of Automated Reasoning*, **4**, 237-267. https://doi.org/10.1007/BF00244942

[4] Chou, S.C. and Gao, X.S. (1990) Ritt-Wu's Decomposition Algorithm and Geometry Theorem Proving. *International Conference on Automated Deduction*, Kaiserslautern, July 1990, 207-220. https://doi.org/10.1007/3-540-52885-7_89

[5] Buchberger, B., Collins, G.E. and Kutzler, B. (1988) Algebraic Methods for Geometric Reasoning. *Annual Review of Computer Science*, **3**, 85-119. https://doi.org/10.1146/annurev.cs.03.060188.000505

[6] Kutzler, B. and Stifter, S. (1986) On the Application of Buchberger's Algorithm to Automated Geometry Theorem Proving. *Journal of Symbolic Computation*, **2**, 389-397. https://doi.org/10.1016/S0747-7171(86)80006-2

[7] Hong, J.W. (1986) Can We Prove Geometric Theorems with Examples? *Science in China Series A-Mathematics, Physics, Astronomy & Technological Science*, **16**, 234-242. https://doi.org/10.1360/za1986-16-3-234

[8] Zhang, J.Z., Yang, L. and Deng, M. (1990) The Parallel Numerical Method of Me-

chanical Theorem Proving. *Theoretical Computer Science*, **74**, 253-271. https://doi.org/10.1016/0304-3975(90)90077-U

[9] Zhang, J.Z., Chou, S.C. and Gao, X.S. (1995) Automated Production of Traditional Proofs for Theorems in Euclidean Geometry I. The Hilbert Intersection Point Theorems. *Annals of Mathematics & Artificial Intelligence*, **13**, 109-137. https://doi.org/10.1007/BF01531326

[10] Chou, S.C., Gao, X.S. and Zhang, J.Z. (1993) Automated Production of Traditional Proofs for Constructive Geometry Theorems. *Proceedings Eighth Annual IEEE Symposium on Logic in Computer Science*, Montreal, 19-23 June 1993, 48-56. https://doi.org/10.1109/LICS.1993.287601

[11] Chou, S.C., Gao, X.S. and Zhang, J.Z. (1993) Mechanical Geometry Theorem Proving by Vector Calculation. *Proceedings of the* 1993 *International Symposium on Symbolic and Algebraic Computation*, Kiev, Ukraine, August 1993, 284-291. https://doi.org/10.1145/164081.164142

[12] Chou, S.C., Gao, X.S. and Zhang, J.Z. (1996) Automated Generation of Readable Proofs with Geometric Invariants. *Journal of Automated Reasoning*, **17**, 349-370. https://doi.org/10.1007/BF00283134

[13] Zhang, J.Z., Gao, X.S. and Chou, S.C. (2015) Invariance Method of Machine Proof for Geometric Theorems. Science Press, Beijing.

[14] Gelernter, H., Hansen, J., Loveland, D. (1960) Empirical Explorations of the Geometry-Theorem Proving Machine. *Western Joint IRE-AIEE-ACM Computer Conference*, San Francisco, California, 3-5 May 1960, 143-149. https://doi.org/10.1145/1460361.1460381

[15] Nevins, A.J. (1975) Plane Geometry Theorem Proving Using Forward Chaining. *Artificial Intelligence*, **6**, 1-23. https://doi.org/10.1016/0004-3702(75)90013-2

[16] Chou, S.C., Gao, X.S. and Zhang, J.Z. (2000) A Deductive Database Approach to Automated Geometry Theorem Proving and Discovering. *Journal of Automated Reasoning*, **25**, 219-246. https://doi.org/10.1023/A:1006171315513

[17] Sun, X. (2021) Research on Intuitive Geometry Based on Computer Experimental Mathematics and Statistical Analysis. Master Thesis, Shanghai University, Shanghai.

[18] Wu, W.-T. (2003) Mathematics Mechanization. Science Press, Beijing.

[19] Matsui, Y., Mizuta, M., Ito, S., Miyano, S. and Shimamura, T. (2016) D3M: Detection of Differential Distributions of Methylation Levels. *Bioinformatics*, **32**, 2248-2255. https://doi.org/10.1093/bioinformatics/btw138

[20] Schefzik, R., Flesch, H. and Goncalves, A. (2021) Fast Identification of Differential Distributions in Single-Cell RNA-Sequencing Data with waddR. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btab226

[21] Irpino, A. and Verde, R. (2015) Basic Statistics for Distributional Symbolic Variables: A New Metric-Based Approach. *Advances in Data Analysis and Classification*, **9**, 143-175. https://doi.org/10.1007/s11634-014-0176-4

[22] Feng, Y., Chen, J.W. and Wu, W.Y. (2019) The PSLQ Algorithm for Empirical Data. *Mathematics of Computation*, **88**, 1479-1501. https://doi.org/10.1090/mcom/3356