Snapshots of modern mathematics
from Oberwolfach

№ 15/2018

Estimating the volume of a convex body.

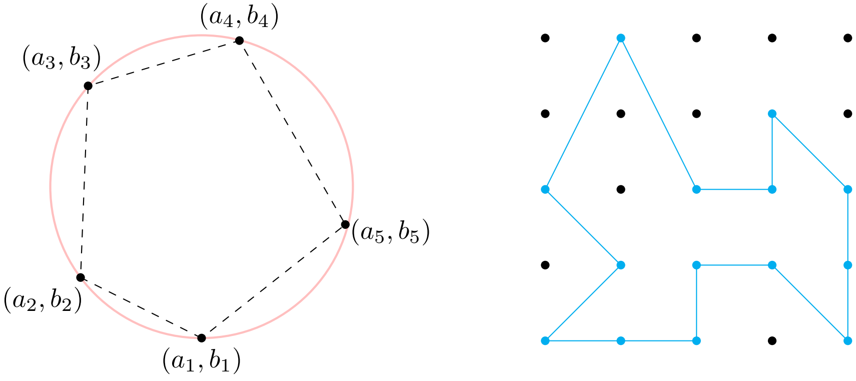
Nicolai Baldin

Sometimes the volume of a convex body needs to be estimated, if we cannot calculate it analytically. We explain how statistics can be used not only to approximate the *volume* of the convex body, but also its *shape*.

1 Calculating the volume in analytic geometry

Calculating the volume of geometric objects is a major topic in analytic geometry. In high school, we humbly touch it when we study two- and three-dimensional shapes. We learn various formulas and techniques of calculating areas and volumes of simple geometric objects like a polygon and ellipse in planimetrics; and a polyhedron, cone and sphere in solid geometry. Sometimes, we cover even some high-dimensional objects, but this is usually already a university topic. Often advanced geometry problems appear in international olympiads in mathematics for high school students or even in different university-level mathematical competitions, see for example [3, Problem 1].

In the two-dimensional case, there are plenty of formulas for calculating areas of various geometric objects. For example, in order to calculate the area enclosed by a circle with radius $r > 0$ we use the formula $S = \pi r^2$ that was discovered by the ancient Greeks. Perhaps, we may want to calculate the area S_p of a simple polygon, for example an irregular pentagon, inscribed in a circle, see Figure 1(a). In this case the area, of course, depends only on the vertices of the polygon. Due to the *shoelace formula* discovered by the prominent German



(a) Shoelace formula.

(b) Pick's theorem.

Figure 1: Two ways of calculating the volume.

mathematician Carl Friedrich Gauß (1777–1855), we have

$$S_p = \frac{|(a_1b_2 + a_2b_3 + a_3b_4 + a_4b_5 + a_5b_1) - (b_1a_2 + b_2a_3 + b_3a_4 + b_5a_5 + b_5a_1)|}{2}.$$

A proof of this formula can be found in [7]. Another interesting result is related to calculating the area of a polygon constructed on a square-grid of points with integer coordinates such that all the polygon's vertices are grid points, see Figure 1(b). Pick's theorem, described by Georg Alexander Pick (1859–1942), provides a simple formula for calculating the area S of this polygon in terms of the number $n_\circ$ of grid points located in the interior of the polygon and the number n_∂ of grid points (blue) lying on the polygon's boundary:

$$S = n_\circ + \frac{n_\partial}{2} - 1\,.$$

Already in the three-dimensional case some problems appear to be quite challenging for arbitrary *convex bodies*.[1] To calculate the volume enclosed in a sphere with radius $r > 0$ we use the well-known formula $V = 4\pi r^3/3$.

[1] A set $C \subset \mathbb{R}^3$ is said to be *convex* if, for all x and y in C and all $t \in [0,1]$, the point $(1-t)x + ty$ also belongs to C. In other words, every point on the line segment connecting x and y is in C. A convex set is called a *convex body* if it satisfies a certain pretty generic topological property, namely, for the experts, if it has non-empty interior. One example of a convex body is the *convex hull* of a finite set of points $x_1, \ldots, x_n$: it is defined as the set $\widehat{C} := \{\sum_i \lambda_i x_i | \sum_i \lambda_i = 1, \lambda_i \geq 0\}$. A *polyhedron* is a set that can be written as the convex hull of a finite set of points. For example, the cube is a polyhedron: it is the convex hull of its eight corners.

Calculating the volume of a polyhedron with given vertices inscribed in a sphere is already an involved task. Let us assume that the boundary of a polyhedron P is given by a union of triangles $A_i, i = 1, \ldots, n$, (general faces can be divided into triangles) with vertices $(\vec{a}_i, \vec{b}_i, \vec{c}_i)$ which are assumed to be ordered counter-clockwise on A_i, when looking at them from the outside of the polyhedron. This means that on each A_i we can define the outer normal vector $\vec{n}_i = (\vec{b}_i - \vec{a}_i) \times (\vec{c}_i - \vec{a}_i)$, which is a vector that is perpendicular to the triangle. Then the volume of P is given by

$$V_P = \frac{1}{6} \sum_{i=1}^{n} \vec{a}_i \cdot \vec{n}_i \, . \tag{1}$$

A proof of this result is based on the "divergence theorem" and can be found in [6]. As in the two-dimensional case, this result (as well as the divergence theorem) is due to Gauß.

Generalizing the 3-dimensional case, for $d > 3$, a d-dimensional sphere is defined as the set of all points $\vec{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ such that

$$(x_1 - c_1)^2 + (x_2 - c_2)^2 + \ldots + (x_{d-1} - c_{d-1})^2 + (x_d - c_d)^2 = r^2 \, ,$$

where $\vec{c} = (c_1, ..., c_d)$ is the centre point and $r > 0$ is the radius. The enclosed volume is given by

$$V_d = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2} + 1)} r^d \, , \tag{2}$$

where Γ is the *gamma function*, which satisfies

$$\Gamma(\tfrac{1}{2}) = \sqrt{\pi} \qquad \Gamma(1) = 1 \qquad \Gamma(x + 1) = x\Gamma(x).$$

The formula (2) can be obtained in different ways, for example using recursion and integration in spherical coordinates. The analytical expression for the volume of a polyhedron becomes even more complicated in dimensions $d > 3$. We refer to [4] for a comprehensive summary of existing results in calculating the volume in analytic geometry. *What about other convex bodies which have an arbitrary boundary?* There is no unique recipe that allows to calculate the volume of an arbitrary convex body exactly, but, as we shall see later, there are several techniques that allow to *approximate* the volume of an arbitrary convex body with good precision.

2 Estimating the volume in statistics

There can be no doubt that the origin of analytic geometry in antiquity was empirical. However, when you think about calculating the volumes of some natural objects that arise nowadays in biology like a patient's tumour or in astronomy like star clusters, the objects themselves are not accessible, in the sense that we do not know the true shape of a studied object. We have access to only some information, or the *data*, often imprecise, and we want to recover the true shape of the body, its volume and possibly other characteristics. The data we have are some sort of measurements like the detection of presence of a body in a certain region. Extracting the information from the data about the true body is one object of study in *statistics*.

2.1 A brief introduction to probability theory

In order to start conducting a statistical inference analysis, we need to introduce several notions from the beautiful subject of *probability theory*. We are mostly interested in describing the following experiment. Let us pick a random point X in the interval I_{AB} with end points A and B. It is intuitively clear that the *probability* that the point lies in any subinterval $I_{A_1B_1}$ is

$$\mathbb{P}(X \in I_{A_1B_1}) = \frac{|I_{A_1B_1}|}{|I_{AB}|} = \frac{B_1 - A_1}{B - A}, \text{ see Figure 2.} \tag{3}$$

Thus, we also have $\mathbb{P}(X \in I_{AB}) = 1$. We say that X is *uniformly distributed* on the interval I_{AB} if the *law* (3) holds. This is the so-called uniform distribution, the simplest continuous distribution in probability theory. If we randomly draw the point X sufficiently many times, we expect its average position to be close to the midpoint $(A + B)/2$ from this interval. In probability theory, this idea of the long-run average of repetitions is incorporated in the notion of the *expected value* (sometimes also called the *expectation* or *mean*). For instance, the expected position or *value* of the variable X is

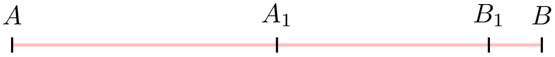


Figure 2: The interval I_{AB} with subinterval $I_{A_1B_1}$.

$$\mathbb{E}[X] = \frac{A + B}{2}\,.$$

Now let us draw $n \in \mathbb{N}$ points $X_1, \ldots, X_n$ uniformly from this interval independently of each other, that is, for each of the points the *law* (3) holds, see Figure 3.

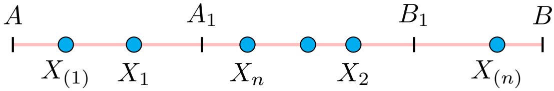


Figure 3: A sample of n points $X_1, \dots, X_n$ drawn uniformly from I_{AB}.

Then the probability to see *all the points* in the subinterval $I_{A_1B_1}$ is

$$\mathbb{P}(X_1, \dots, X_n \in I_{A_1B_1}) = \frac{(B_1 - A_1)^n}{(B - A)^n}. \tag{4}$$

It is quite interesting that whatever the length of the subinterval $I_{A_1B_1} \subsetneq I_{AB}$ is, the probability to see all the points lying in this interval tends to zero as the number of points n tends to infinity, because $(B_1 - A_1)/(B - A) < 1$. Probability theory also tells us that the event that two points coincide has probability zero.

Exercise. *What is the probability to observe at least one point X_1 from the sample $X_1, \dots, X_n$ in the subinterval $I_{A_1B_1}$?*

Probability theory studies different distributional characteristics of the points $X_1, \dots, X_n$ assuming that the locations of the interval end points A and B are *known*. Let us sort the values $X_1, \dots, X_n$ in increasing order by defining the points $X_{(1)}, \dots, X_{(n)}$ such that $X_{(1)} = \min(X_1, \dots, X_n)$, $X_{(2)} = \min(\{X_1, \dots, X_n\} \setminus \{X_{(1)}\}), \dots, X_{(n)} = \max(X_1, \dots, X_n)$. The standard questions that we usually elaborate on in a first course in probability theory at university include:

- What is the probability of observing the largest point $X_{(n)}$ in a certain subinterval $I_{A_1B_1}$ or, in the formal probability theory language, what is the *probability distribution* of $X_{(n)}$?
- What is the length of the interval $I_{X_{(1)}X_{(n)}}$ likely to be equal to, or in formal language, what is the *expectation* of the length?

The answers clearly depend on the locations of the points A and B. To answer the first question, note that the event that $X_{(n)} \in I_{AB_1}$ implies that *all* the points $X_1, ..., X_n$ lie in the subinterval I_{AB_1}. Using equation (4), we compute

$$\begin{aligned}\mathbb{P}(X_{(n)} \in I_{A_1B_1}) &= \mathbb{P}(X_{(n)} \in I_{AB_1}) - \mathbb{P}(X_{(n)} \in I_{AA_1}) \\ &= \mathbb{P}(X_1, ..., X_n \in I_{AB_1}) - \mathbb{P}(X_1, ..., X_n \in I_{AA_1}) \\ &= \frac{(B_1 - A)^n - (A_1 - A)^n}{(B - A)^n}.\end{aligned}$$

In particular, the probability to observe the rightmost point $X_{(n)}$ in the subinterval I_{AM}, where $M = (A + B)/2$ is equal to $(1/2)^n$ which quickly tends to zero when n goes to infinity.

To answer the second question about the expected length of the interval $I_{X_{(1)}X_{(n)}}$, you may first want to calculate the expected value of $X_{(n)}$ and so the expected length of the subinterval $I_{X_{(n)}B}$. Using symmetry, we can also calculate the length of the interval $I_{AX_{(1)}}$ and thereby of $I_{X_{(1)}X_{(n)}}$ using linearity of expectation. We just say that

$$\mathbb{E}[|I_{AX_{(1)}}|] = \mathbb{E}[X_{(1)}] - A = \frac{B-A}{n+1}\,, \tag{5}$$

which implies $\mathbb{E}[|I_{X_{(1)}X_{(n)}}|] = (B-A) - 2\frac{B-A}{n+1} = \frac{(n-1)(B-A)}{n+1}$, and leave details of the proof as an exercise for ambitious readers.

2.2 Statistical inference in the one-dimensional case

In contrast to probability theory, statistics deals with the inverse problem: we do not know the locations of the points A and B and we observe only the points $X_1, \ldots, X_n$ which lie uniformly over the interval. The problem then is to conduct statistical inference about the true interval. The questions include

- What is the length L_{AB} of the interval I_{AB}?
- What are the locations of the points A and B?

If we know the length of the interval I_{AB} then we can always estimate the locations of the points A and B somehow with a better precision. In practice, it is never possible to determine the desired quantities precisely, but we might hope to find some functions of the data, so-called *estimators*, that approximate these quantities “well”. We comment on the estimation quality in Section 3.

A naive estimator for the length L_{AB} of the interval, which is simply

$$\widehat{L}_{naive} = X_{(n)} - X_{(1)}$$

performs rather poorly, although it seems to be a good starting point. A more attractive idea is to somehow dilate, that is to say stretch, the interval $I_{X_{(1)}X_{(n)}}$ and take the length of the dilated interval as an estimator for L_{AB}. There are at least two appealing ways how we can dilate the interval: 1) one can just add and subtract some fixed vectors from the end points $X_{(n)}$ and $X_{(1)}$ (additive dilation) and 2) one can dilate the interval $I_{X_{(1)}X_{(n)}}$ from its centre $(X_{(n)} + X_{(1)})/2$ with some scaling factor (multiplicative dilation). However, in the one-dimensional case both types of dilations are equivalent.

Let us sketch the idea of a possible dilation. If we repeat our experiment sufficiently many times the average locations of the ordered points $X_{(1)}, ..., X_{(n)}$ will tend to $\mathbb{E}[X_{(1)}], ..., \mathbb{E}[X_{(n)}]$ and will lie equidistantly over the interval I_{AB}, see Figure 4. As we have seen in (5), the distance between the points

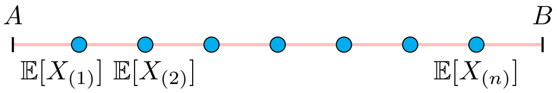


Figure 4: Average locations of the points after repeating the experiment sufficiently many times.

A and $X_{(1)}$, as well as the distance between the points B and $X_{(n)}$, will tend to $\mathbb{E}[X_{(n)} - X_{(1)}]/(n-1)$:

$$\mathbb{E}[|I_{AX_{(1)}}|] = \mathbb{E}[|I_{X_{(n)}B}|] = \frac{B-A}{n+1} = \frac{\mathbb{E}[X_{(n)} - X_{(1)}]}{n-1}.$$

Therefore, a reasonable additive dilation factor is $2(X_{(n)} - X_{(1)})/(n-1)$ and so our dilated estimator for the length is

$$\widehat{L}_1 = (X_{(n)} - X_{(1)}) + 2\frac{X_{(n)} - X_{(1)}}{n-1} = \frac{(n+1)}{(n-1)}(X_{(n)} - X_{(1)})\,. \tag{6}$$

This estimator is not only *unbiased*, which means that $\mathbb{E}[\widehat{L}_1] = B - A$, but also, as we shall see in Section 3, is optimal in a certain statistical sense and therefore it outperforms the estimator $\widehat{L}_{naive}$.

2.3 Higher dimensions

Although the one-dimensional model is very useful to grasp the main ideas of estimating the volume, it is not widely used in real world applications. The two-dimensional model already covers several important applications, for example in geology and medicine. Here, we observe the points $X_1, ..., X_n$ lying in a set $C \subset \mathbb{R}^2$ and we would like to recover the volume V_C of the set and a description of the set itself. Let us assume that the set C is *convex*. On the one hand, this assumption is quite restrictive, but, on the other hand, it allows to develop a nice theory and it still covers many interesting phenomena.

First, let us focus on estimating the volume of the set C (keep in mind that if we know the volume of the set we can estimate the shape of the set with a better precision). As in the one-dimensional case, we would like to start with a simple estimator. What do you think would be an analogue of the simple estimator $\widehat{L}_{naive} = X_{(n)} - X_{(1)}$ in the two-dimensional case? It is natural to take the volume $|\widehat{C}|$ of the convex hull as a starting estimator for the volume V_C of the set C. It is intuitive that this estimator performs quite poorly because it always underestimates the true volume and so it should be dilated as in the

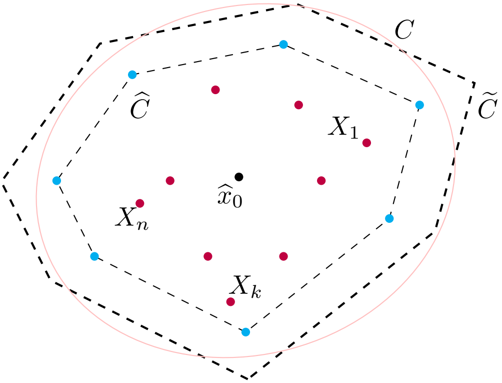


Figure 5: The points $X_1, ..., X_n$ drawn uniformly over a set C, the convex hull of the points $\widehat{C} = \mathrm{conv}(X_1, ..., X_n)$ and the dilated hull estimator $\widetilde{C}$.

one-dimensional case. Without details, we claim that the optimal estimator is

$$\widehat{V}_{opt} = \frac{n+1}{n_\circ + 1} |\widehat{C}| \,, \tag{7}$$

where $n_\circ$ is the number of points that lie in the interior of the convex hull $\widehat{C}$. These points are coloured purple in Figure 5. Note that $\widehat{V}_{opt}$ is the volume of the "dilated" hull $\widetilde{C}$, the set obtained by dilating the convex hull with the same factor from the centre $\widehat{x}_0$ of the convex hull:

$$\widetilde{C} = \left\{ \widehat{x}_0 + \left(\frac{n+1}{n_\circ + 1} \right)^{1/2} (x - \widehat{x}_0) \,\middle|\, x \in \widehat{C} \right\},$$

which in fact can be used to estimate the shape of the set C itself. Similarly, the same estimators for the volume and the set itself can be used in higher dimensions. Do you see that the estimator (6) is a special case of the estimator (7)?

3 Estimation quality

We have already seen that there are different estimators for the length L_{AB} of the interval I_{AB} based on observations of the points lying in the interval. One could even estimate the length by $\widehat{L} = 1$ always and independently of the data. *How can we rank different estimators?* How can we say that one estimator is better than another and according to what criteria? We need to find a function that measures the performance of our estimators so that we could rank the estimators according to their performance. Let us focus on estimating the length of the interval I_{AB}. It is desirable to have the estimator of the length close to the true length of the interval. In statistics, to measure this proximity it is common to use a quadratic *loss function*,

$$l(L_{AB}, \widehat{L}) = (L_{AB} - \widehat{L})^2 \,,$$

essentially because it is differentiable and symmetric: an error above the target yields the same loss as the same magnitude of error below the target. Given one particular *sample* $X = (X_1, ..., X_n)$, we can rank all possible estimators according to this function. Unfortunately, the ranking of estimators based on the loss function l can be different for different samples! See Figure 6 for an example when the ranking of $l\big(L_{AB}, \widehat{L}_1\big)$ and $l\big(L_{AB}, \widehat{L}_{naive}\big)$ is different for two different samples $X = (X_1, ..., X_n)$ and $X' = (X'_1, ..., X'_n)$. That is why we would like to rank all estimators according to their "average" loss, which is the expected loss $\mathbb{E}[l(L_{AB}, \widehat{L})]$, also referred to as the *risk function*. It may sound quite ambitious, but in fact it is not hard to compare different estimators according to the risk function. In particular, we can calculate that

$$\mathbb{E}[l(L_{AB}, \widehat{L}_1)] \leq \mathbb{E}[l(L_{AB}, \widehat{L}_{naive})] \,.$$

Similarly in higher dimensions, we measure the performance of an estimator of the volume of the convex body according to its risk function $\mathbb{E}[l(V_C, \widehat{V})]$.

The risk can be seen also as a function of the true parameter (the interval or the body). Note that our estimator $\widehat{L} = 1$ for the length L_{AB} estimates it perfectly when the length L_{AB} indeed equals 1. Since we do not have access to the true shape of the body, we want our preferred estimator perform well

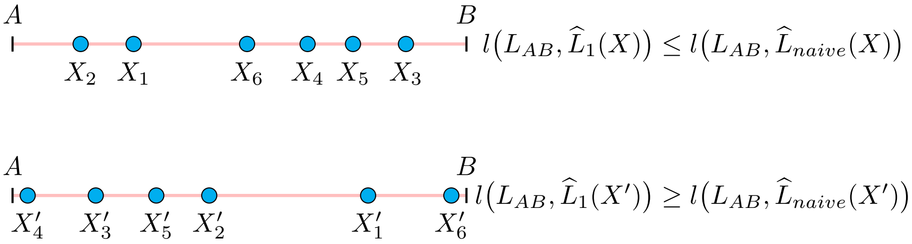

A
B
$l\big(L_{AB}, \widehat{L}_1(X)\big) \leq l\big(L_{AB}, \widehat{L}_{naive}(X)\big)$
X_2
X_1
X_6
X_4
X_5
X_3
A
B
$l\big(L_{AB}, \widehat{L}_1(X')\big) \geq l\big(L_{AB}, \widehat{L}_{naive}(X')\big)$
X'_4
X'_3
X'_5
X'_2
X'_1
X'_6

Figure 6: Two samples $X = (X_1, ..., X_6)$ and $X' = (X'_1, ..., X'_6)$ that yield different rankings of the loss function. In the first case, it is obvious that $X_3 - X_2$ underestimates the length $B - A$ and needs to be enlarged while in the second case $X'_6 - X'_4$ is already pretty close to $B - A$ and the dilation can only increase the loss.

for all possible shapes. That is why in statistics we often analyse estimators according to their *worst-case risk*, $\sup_C \mathbb{E}[l(V_C, \widehat{V})]$. Another desirable uniform criteria is the so-called UMVU property of an estimator which means that the estimator is Unbiased and has Minimum Variance[2] among all unbiased estimators Uniformly over all possible true parameters. One can show that the estimator for the volume $\widehat{V}_{opt}$ in (7) is in fact nearly UMVU in all dimensions, see [2] and [1] for the proof of this result.

4 Looking further. Computational geometry

Fast algorithms for calculating volumes of arbitrary convex bodies are needed in computer science and computational geometry. As the dimension grows, the studied objects become more and more complicated and it is no longer possible to apply some nice analytical formula like (1) even if we know the location of the object. Different numerical methods based on partitioning the initial body into simpler sets like cubes serve to estimate the volumes with a good precision. However, many such numerical methods have been found to be computationally inefficient and therefore different fast randomised methods to estimate the volume are used. Often a trade-off between running time of an algorithm and the quality of the estimate has to be made. We refer to [9] for a survey of the existing fast randomised algorithms for calculating the volume.

One of such randomised algorithms, although definitely not the fastest, is exactly to follow the experiment above. Given a body that we want to find the volume of, we can draw the points uniformly over it, calculate the volume of the convex hull of the points and then make a necessary dilation. Since it is computationally easier to calculate the volume of a polytope than of an arbitrary convex body, this procedure can save expensive running time, although computing the volume of the convex hull is still an involved task. Nevertheless, it is fascinating that once the volume of the convex hull is computed the dilation (7) involving the number of points should be employed to estimate the volume in an optimal way. This is a beautiful example of achieving a substantial gain combining efficient algorithms with advanced probability theory and statistics.

To summarise, we have only touched on some of the topics of probability theory and statistics. For those who would like to explore these subjects in more detail we refer to introductory books [5] and [8] which we find both entertaining and rigorous.

[2] In probability theory, the variance of a random variable X is defined by $\mathrm{Var}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2]$ and it measures how far the random variable is spread out.

Solution to the exercise:

$$1 - \frac{(B-B_1)^n+(A_1-A)^n}{(B-A)^n}$$

Acknowledgements

The author would like to thank Markus Reiß and Martin Wahl for valuable discussions.

References

[1] N. Baldin, *The wrapping hull and a unified framework for volume estimation*, arxiv:1703.01658, 2017.

[2] N. Baldin and M. Reiß, *Unbiased estimation of the volume of a convex body*, Stochastic Processes and their Applications **126** (2016), no. 12, 3716–3732, http://www.sciencedirect.com/science/article/pii/S0304414916300369.

[3] IMC Advisory board, *International mathematics competition for university students*, 2015, http://www.imc-math.org.uk/imc2009/imc2009-day2-solutions.pdf, visited on 28th July 2017.

[4] O. Bretscher, *Linear algebra with applications*, Pearson Education, 2013.

[5] R. Durrett, *Probability: Theory and examples*, Cambridge University Press, 2010.

[6] R. Nürnberg, *Notes on calculating the volume and centroid of a polyhedron*, 2015, http://wwwf.imperial.ac.uk/~rn/centroid.pdf, visited on 28th July 2017.

[7] Art of Problem Solving Wiki, *Shoelace theorem*, http://www.artofproblemsolving.com/wiki/index.php/Shoelace_Theorem, visited on 28th July 2017.

[8] Y. Suhov and M. Kelbert, *Probability and statistics by example: Volume 1, basic probability and statistics*, Cambridge University Press, 2005.

[9] S. Vempala, *Recent Progress and Open Problems in Algorithmic Convex Geometry*, IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2010), Leibniz International Proceedings in Informatics (LIPIcs), vol. 8, pp. 42–64.

Nicolai Baldin *is a PhD student at the University of Cambridge and previously at Humboldt University of Berlin*

Mathematical subjects
Probability Theory and Statistics, Geometry and Topology

Connections to other fields
Computer Science

DOI
10.14760/SNAP-2018-015-EN

Snapshots of modern mathematics from Oberwolfach are written by participants in the scientific program of the Mathematisches Forschungsinstitut Oberwolfach (MFO). The snapshot project is designed to promote the understanding and appreciation of modern mathematics and mathematical research in the interested public worldwide. It started as part of the project "Oberwolfach meets IMAGINARY" in 2013 with a grant by the Klaus Tschira Foundation. The project has also been supported by the Oberwolfach Foundation and the MFO. All snapshots can be found on www.imaginary.org/snapshots and on www.mfo.de/snapshots.

Mathematisches Forschungsinstitut
Oberwolfach gGmbH
Schwarzwaldstr. 9 – 11
77709 Oberwolfach
Germany

Director
Gerhard Huisken

Mathematisches Forschungsinstitut Oberwolfach
Member of the Leibniz Leibniz Association
Klaus Tschira Stiftung gemeinnützige GmbH
KTS
oberwolfach FOUNDATION
IMAGINARY open mathematics