USING MENTAL IMAGERY PROCESSES FOR TEACHING/SEARCHING IN MATHEMATICS AND COMPUTER SCIENCE

PIERRE ARNOUX AND ALAIN FINKEL

Abstract. The role of mental representations in mathematics and computer sciences (for teaching or searching) is often downplayed or even completely ignored. Using an ongoing work on the subject, we argue for a more systematic study and use of mental representations, to get an intuition of mathematical concepts, and also to understand and build proofs. We give several examples.

1. introduction

1.1. **Background.** The study of mental representations (specially visual ones, i.e. visual images) has a long history. Philosophers from Platon and Aristote to Heidegger and Merleau-Ponty have explored mental states. The beginning of the scientific exploration of mental states can be dated 150 years ago. The Wursbourg School (Tichener 1850), the introspective psychology (Binet 1900) described the phenomenology of mental states. Wittgenstein(1922) proposes a theory of semantics in which one builds images from facts and combines images. (See the web page about mental imagery: `http://plato.stanford.edu/entries/mental-imagery/bibliography-mental-imagery.html`). More precisely, Pavio (1969) was one of the first to prove the role of double coding (verbal and visual); among a lot of other researchers, Kosslyn, Denis, Mellet, Pinker,...[DMK04a, DMK04b] studied the different modalities of representations in different contexts like itinerary strategies; they also used the neuroimaging for studying the mental imagery, and it now becomes possible to check, using the neurosciences technologies, that different areas of the brain are mobilised for processing verbal processes or visual processes. In particular, some researchers, like Lacey, Campbell, Sathian [LCS07], explore the visuo-haptic cross-modal memory paradigm. We believe that the construction of multimodal representations (visual, symbolic/verbal, haptic,...) is important for understanding mathematics (and for understanding in general).

1.2. **Mental imagery for mathematics and computer science.** Understanding a mathematical definition, a proof, a concept ,.... is strongly correlated with the building and handling of mental objects, (see the book *L'homme neuronal*, Jean-Pierre Changeux, 1983). One of the oldest and more striking examples might be the plane representation of complex numbers, discovered more than two centuries after their invention, which made clear many of their properties; it is today unthinkable to teach complex numbers without teaching the complex plane, and this mental representation is invaluable in many respects (for understanding concepts, or finding proof strategies; for an example, see the book *Visual complex analysis*, Tristan

Date: version June 3, 2008, Accepted for presentation at ICME-11, the International Congress on Mathematical Education, eleventh edition, Confeence TSG 17: Research and development in the teaching and learning of advanced mathematical topics, Monterrey, Mexico, July 6 - 13, 2008.

Needham, 1997). The Working Group on Representations [GJ98] synthetises four different interpretations of the terms "representations". Recently, Carlson, Oehrtman and Thompson studied the mental representations (in terms of mental images and mental actions) associated with functions [COT07]. Thompson studied the relations between mathematical reasonning and imagery [Tho96]. However, the usefulness of building representations is often downplayed as an aside, or ignored; one reason is that good representations are not easy to obtain (it took more than 200 years for the complex plane!), and not easy to use in a profitable way. Another reason is that teachers and researchers do not know, in general, that it is important to communicate about them. We have started a modest team work of study of mental representations, and we would like to argue for a more systematic study, at several levels:

- Classifying different types of representations (visual, kinestesic...); There is already some work in this direction, but not usually applied to mathematical teaching. It would be good to have a set of useful representations of different types. For example, it is unclear to us (maybe because of our own limitations) what can be an auditive representation which is non-verbal .
- Studying the efficiency of various representations in a systematic way. We know, because many students have told us so, that some representations are useful to understand concepts and proofs, but this remains at the moment very unsystematic. It is highly probable that no representation be universal, different people will be helped by different representations. A good lecture would be one that can use representations of various types, so that all audiences can find the adapted one.
- Building a catalogue of representations. Good representations are difficult to come by; one cannot expect that a teacher, even an outstanding one, can find by himself a good model for all the concepts in his lectures. Furthermore, any person will naturally find a particular type of representation, and will be able to speak efficiently only to a part of the audience. A large catalogue of mental representations, with their type and the way they can be used efficiently, would be a very useful teaching tool.

We give a few examples below.

2. Models for graphs and automata

What is a graph G ? It can be *said* as a relation R between elements of a set E, or as a subset of the cartesien product of $R \subseteq E \times E$ (even if mathematicians know that a relation is a subset, this is not the usual meaning of the word relation in natural languages); if the student knows what is a function, a relation may be defined as a function which associates to some element $x \in E$ a subset $E_x \subseteq E$. These definition are said mainly verbal in the sense that they do not explicitely induce a visual image or a special feeling. A graph can also be *seen* as a figure with nodes and arcs $G = (E, \to)$. Concrete examples are maps (nodes are towns and arcs are roads, nodes are metro stations and arcs are rails).

A different view, well-adapted to a computer presentation, is that of the adjacency matrix, a square matrix $M = (m_{i,j})$ indexed by the nodes, where $m_{i,j}$ count the number of arcs joining nodes i and j. Counting the number of paths of lengths 2 from i to j leads in a very natural way to the classical formula $\sum_k m_{i,k} m_{k,j}$, since any path of length 2 must go through an indeterminate node k. This motivates efficiently the formula for the matrix product; a large chunk of algebra can be illustrated in this way.

Now a graph can be labelled by letters belonging to an alphabet Σ and we obtain a *labelled graph*: $G = (E, \Sigma, \to)$ where $\to \subseteq E \times \Sigma \times E$. A finite automaton (a basic and fundamental

object in computer science) can be *said* as a triple $(E, \Sigma, \rightarrow)$, or *shown* as a labelled graph. The function of a finite automaton is to recognize a set of words on the alphabet Σ.

Let A be a finite automaton $A = (E, \Sigma, \rightarrow)$ where $\rightarrow \subseteq E \times \Sigma \times E$. Fix an initial state e_0 and a finite subset $F \subseteq E$ of final states and define the language $L(A)$ of A as the set of finite words $w \in \Sigma^*$ labelling a path from e_0 to a final state $e \in F$: $L(A) = \{w \in \Sigma^*; e_0 \rightarrow^* e, e \in F\}$.

This function can also be represented in different ways:

- a *recognizer* that reads some letters on an input tape, the reading is conditionned by the set $\rightarrow$ of transitions, and it accepts or refuses some words.
- a *producer* that writes some letters on some output tape,
- a *transducer* which reads letters from a tape and writes letters on another tape. This is a combination of both recognizer and producer.

These differents visions are *not cognitively equivalent* even if they are mathematically (almost) equivalent. For instance, the environnement is important for the recognizer view, but does not exist in the producer view. These representations are mainly visual and kinestesic (in the sense that they use some movement and feeling of movement). Every student would probably have his own preference.

Another view of a finite automaton, which is suitable even for children, is that of a labyrinth with rooms (states) and corridors (transitions); the corridors are one-way only, and they have a name taken in the alphabet Σ (several corridors can have the same name). The labyrinth has one entrance, and one or several exits. To recognize a word, the player is given this word as a stack of cards , each one showing a letter of the word. He enters the labyrinth by the entrance; at each step, he chooses a corridor named by the letter of the card on the top of the stack, discards this card, and follows the corridor to the next hall. Recognition fails if at some moment no corridor in the present room is named by the letter of the top card, or if the last room is not an exit; moreover, the exit of the labyrinth is allowed only when the stack of cards becomes empty. This model can be understood at a very basic level, and allows to attack and solve a number of elementary (and not elementary) problems. It is not complete: for example, what happens if two corridors from the same room bear the same name ? But this incompleteness is in fact convenient, because it allows to introduce interesting concepts (deterministic/nondeterministic) in a natural way, often following questions from the students.

If one views word recognition as a walk on a finite graph (random or deterministic, depending on the type of the graph), there are interesting consequences. For example, such a walk, if long enough, must contain loops, and loops can obviously be iterated: one gets here a very simple and intuitive proof of the classical pumping lemma, usually stated in this way: If a finite automaton has k states, any word w recognized by the automaton, of length at least k, can be factorized $w = xuy$, where $x, u, y \in \Sigma^*$ so that all the words $xu^n y$, for all integers n, are recognized by the automaton. If now, one sees set of words (i.e. languages) as defined by (the set of words belonging to) a rational expression then the pumping lemma becomes:

Lemma 1 (pumping lemma)**.** *Let L be a rational language, then there exists an integer k_L such that for every word $w \in L$ such that $|w| \geq k_L$ there exist three words $x, u, y \in \Sigma^*$ such that $u \neq \epsilon$, $|xu| \leq k_L$, $w = xuy$ and for all $n \geq 0, xu^n y \in L$.*

This lemma is not so simple to prove from the rational expression of L.

There is a deep connection between the set of languages generated by finite automata and by rational expressions. The **Kleene** theorem says that these two sets of languages are the same. Due to this theorem, one may use either the finite automata view (with visual and kinestesic representations) or the rational expressions view (more oriented verbal representations).

In theoretical science, and more generally in all sciences, some viewpoints produce some results more easily than other. For instance, knowing that finite automata exactly generate the class of rational languages (which may be also characterised by formula of logics and/or algebraic characterisation by finite quotient of monoid), we may easily prove the classical pumping lemma for rational languages.

There is another version of the pumping lemma for stack automata and context-free grammars. Still, a visual reasoning on the derivation tree associated with a context-free grammar helps a lot to guess how the (double) iteration occurs; and this iteration lemma is very difficult to guess and prove when only considering stack automata. We don't say that visual reasoning is better than verbal/algebraic reasoning ! But to be able to have representations is important and to be able to adapt our representations allows us to better understand and make statement and proofs.

3. A kinetic model for elementary calculus

A simple image can be presented for the presentation of the basic calculus results, Rolle's theorem and mean values theorem.

We first introduce a simple model of a function: a vehicle moving along a line, with the position as a function of time, and we define the derivative as the speed. All students have an intuitive idea of speed, but they usually have never realized that the speed they can read on the dial of a car is something which is difficult to define as soon as it is not constant; trying to define precisely this concept can motivate the usual definition of the derivative as a limit, even if it is certainly too much to expect that the students can recover the definition by themselves.

Rolle's theorem. A basic intuition on Rolle's theorem can be given in the following way: throw an object vertically, at time 0, and catch it back at the same height at some later time. In the mean time, the objects goes up, with slowing speed, stops and goes back down. It is clear that, at the maximum height, the speed is zero.

This is of course not a proof, but it is a good indication: If the object is moving in time, with same initial and final position, consider the moment when it is farthest for the initial position; at that moment, its speed will be zero.

This is again not a proof, but it shows what we need: a continuous function on a bounded interval has extremal values which are obtained; if the boundary values are the same, one of these extremal values must be in the interior of the interval, and for a local extremum, the derivative is 0.

The kinetic image also shows what can go wrong with the theorem: if we throw the object too strongly, it will hurt the ceiling, and there will be a shock. At the time of the shock, the value will be extremal, but the speed will not be zero, it will be undefined. Indeed, Rolle's theorem only works if the function is everywhere derivable inside the interval: if it is undefined in only one point, it is enough to prevent the theorem to apply.

Mean value theorem. The strong version of the mean value theorem asserts that, if a function f defined on an interval $[a, b]$ is everywhere derivable in the interval, there is a point where the derivative is equal to the ratio $\frac{f(b)-f(a)}{b-a}$; it is easily proved by using Rolle's theorem, but it is difficult to make the student's understand the underlying meaning.

The weak version asserts that, if the derivative is everywhere bounded by M, then $|f(b) - f(a)|$ is bounded by $M(b - a)$.

The meaning of the weak version has a clear kinetic interpretation, which is obvious to all students: if a car never exceeds the speed of 100 km/h, in one hour, it cannot travel more than 100 km. In particular, if speed is limited to 100 km/h, and the car enter a highway at some point and leaves it one hour later at an exit distant of 110km, it can be scientifically argued that at some point the car has been driving over the speed limit.

The strong version is more subtle: it asserts, for example, that if a vehicle has always a well defined speed and covers 100 km in one hour, at some moment, its speed will be exactly 100km/h. A geometric picture with a broken affine map shows the necessity of derivability (as for Rolle's theorem); a kinematic example can be more impressive: suppose a motocyle drives on a highway, at a speed of 150 km/h. After 20 minutes, it crashes into a garbage collection truck which drives at 75 km/h; the truck driver discovers the motorcycle stuck in the back ot the truck after 40 minutes. In one hour, the motorcycle has traveled exactly 100 km; however, its speed has never been 100km/h: at one (uncomfortable) moment, it has been undefined.

4. Mechanical model in probability

The first concepts of probability theory (expected value, variance) are often mysterious for the students, and basic results, such as Chebyshev's inequality are usually felt as pure algebraic computation void of meaning. The usual definition of the basic object in probability, a random variable, is itself very abstract, and does not allow an easy intuition. But there is a mechanical model for a random variable which, although not complete, can give a much better grasp on these concepts.

The mechanical model. A random variable is usually defined as a real function, denoted by X, on a space endowed with a measure μ of total mass 1; but in most elementary computations, one only uses the image measure $X_*\mu$ on the real line.

Hence we view a random variable as a distribution of measure, or probability, on the real line; and we represent this probability distribution as a metal bar of variable width and finite weight. One should take some time to play with this representation for many cases: the uniform distribution on an interval is just a metal bar of constant width and finite length, like a usual metal ruler; a Bernoulli distribution, associated to a game of heads and tails, is given by two point masses (atoms) linked with a rigid bar without weight, a limit model of an halter; a bell curve distribution is an infinite bar which becomes extremely thin for large values, and which can be shown explicitly by the graph of the bell curve. We can consider this bar as being made from small atoms; the random variable consists in choosing an atom in the bar, and it is given by the location of the atom. A random choice will go more often to the places where the bar is thick, and there are many atoms. We will always suppose that the total mass of the bar is finite, otherwise there is no meaningful way to make a random choice, and we can normalize the mass to 1, as is usual in probability.

It takes time to explain this analogy, which can be made rather precise. It is not complete, but its deficiencies illustrate in concrete ways fundamental problems of measure theory and distribution theory; depending on the assistance, one can choose to overlook them, or explain the fine details.

Several concepts follow immediately from this representation.

Expected value and center of mass: the non-random part. The probability of a set A, $P(X \in A)$, is the mass of the corresponding part of the bar. Once the representation

is understood, the computation to do is usually clear (although it can require some analytic skills).

The expected value is just the center of mass of the bar, and gives in some sense the "nonrandom part" of the random variable. It always exists for a bounded bar, but its existence for a non-bounded bar is not guaranteed: a bar infinite in one direction, of width $\frac{1}{1+x^2}$ for $x \geq 0$, has a center of mass at infinity (there is too much mass very far), and a bar infinite in both directions, of width $\frac{1}{1+x^2}$ for all x, has no center of mass at all! This means that if one repeats the experience, the mean value will go to infinity in the first case, and will vary in an erratic and non-bounded way in the second case.

As further developments of the theory will show, the expected value is the value around which the random variable fluctuates: under suitable hypothesis, the average of successive independent trials converges in a precise sense to this expected value (laws of large numbers).

Variance and moment of inertia: measuring randomness. The variance is the moment of inertia of the bar with respect to the gravity center, if it exists, and this gives a way to measure the randomness of the distribution: if the moment of inertia is zero, all the mass is concentrated in one point, and the variable, taking only one value, is not random. One can get a feeling of this moment of inertia, by imagining the metal bar pivoting on an axis through the center of mass, and the effort needed to make this bar move around the axis; is is obviously 0 if all the mass is in the axis, and large if most of the mass is very far; it is easy to compute for finite example, like the halter, where is is clearly proportional to the square of the distance separating the two masses.

Simple examples, like the halter (Bernoulli), show that, if the variable is expressed in some unit (for exemple meter), the variance is expressed in the square of this unit; hence it is natural to consider the square root of the variance, which is called the standard deviation, and which is proportional to the variable, and expressed in the same unit.

Playing with different models show that the variance, or preferably the standard deviation, is a first crude measure of the randomness of the variable.

An application: Chebyshev's inequality. This equality is usually stated in the following way:

Theorem 1 (Chebyshev's inequality)**.** *Let X be a random variable with expected value E and variance V. We have: $P(|X - E| \geq d) \leq \frac{V}{d^2}$.*

The proof is generally given as a sequence of inequalities, which are clear, but void of apparent signification. It is rarely understood by the students. Let us precise this last statement. We may distinguish at least two different understanding: the first one consists in being able to verify that a sequence of formulas makes a correct proof. This can be done locally without having a global vision of the proof and without any feeling of understanding (this feeling often indicates something important about the deep understanding of the human; while the feeling is not sufficient, we think that it is necessary). If we make a metaphor with computer science and logics, we observe that a program may verify that a given finite sequence of formula is a correct proof more easily than find such a proof or know whether a proof exists. The second understanding contains the first one and moreover, there exists a mental representation of the complete proof in the mind of the human; this often comes with a positive feeling. We make a connection between this feeling and a somatic-marker (an hypothesis from Antonio Damasio) which is a mechanism by which emotional processes can guide decision-making (in this case to decide whether we have understood).

Using the previous model, we can give an sketch of proof for this theorem, which is intuitive and mathematically sound, as follows:

It is clear that, in a system of masses, if we slide the mass away from the center of mass, the moment of inertia increases. This is immediately felt by all students, and it is easy to prove from the formula. Suppose that we have a system of masses such that there is a mass m at distance at least d form the center of mass, and the rest of the mass at a smaller distance. The minimum of the moment of inertia for such a system will be realized if we have exactly mass m at distance d, and the rest of the mass is at the center of mass (one can make an exact picture: in that case, we will have two atoms of mass $\frac{m}{2}$ on both sides of the center of mass, at distance d, and one atom of mass $1-m$ at the center of mass). It is clear that, in that case, the moment of inertia is exactly md^2; in any other case, it is greater.

Due to the lack of picture associated with our verbal concrete and visual-oriented description, the reader, (you !) has probally taken some time for constructing a visual and kinetic mental representation of the system of masses. If after some time, the reader did not get through to build a satisfying mental representation, he has the choice to continue his reading without real understanding or to stop his reading. This is a strong argument to provide to students (and also to our colleagues) some representations for helping them to understand us.

We have proved that, as soon as there is at least mass m at distance at least d from the center of mass, the variance, or moment of inertia, V satisfies $V \geq md^2$, or $m \leq \frac{V}{d^2}$; this is exactly Chebyshev's inequality; the formalization of this proof is the usual proof, which appears natural in this setting.

Further developments. This model can be extended to several dimensional random variables; one must of course start with two dimensions, since most difficulties are already present in that case, yet one can still give a good visual representation, which is difficult in 3 dimensions, and impossible in higher dimension.

In that case, the center of mass is easy to define in the same way as before, but the moment of inertia is no more a number, it is a tensor. The analogy can only be followed with rather advanced students, but one can then show very interesting phenomena.

A two-dimensional random variable can also be seen as a pair of random variable X, Y on the same underlying space, by taking coordinates, and it defines a distribution of mass in the plane. The tensor of inertia is then defined by a matrix, whose diagonal elements are the variances of the two variables, and the anti-diagonal element (which are equal, since the matrix of inertia is symmetric) are the covariance of the two variables; if this covariance is zero, the variables can be considered as orthogonal, in a sense which can be made precise: spaces of square-integrable function (L^2-space).

A special case is that of 2 independent variables: this means that the distribution of masses is a product distribution, and this easily implies that the matrix of inertia is diagonal and covariance is 0: independence implies orthogonality. It is useful at this point to study a few simple cases: two independent 0-1 variable (head or tail variable), or two independent uniform variables on the unit interval; one can then prove either geometrically or algebraically that the variance of the sum of two independent variables is the sum of the variance.

This is a form of Pythagoras theorem: if we consider the standard deviation as the size, or length, of randomness of the random variable, and the variance as the square of the length, then, for two orthogonal variables, the square of the length of their sum is the sum of the squares of their lengths.

Students often ask why we define the variance by taking the square of the distance to the expected value, and not the absolute value, which might also give an interesting quantity (it is the L^1 norm for integrable functions), and is often easier to compute; this property of additivity is probably the best answer that can be made: no other quantity would satisfy the Pythagoras theorem, and this leads to deep consequences (easy proof of the weak law of large numbers, which justifies the initial definition of the expected value).

A physical model for Cauchy density. Here is a nice physical model for the classical Cauchy density, which is often used as a counter example.

Consider a laser beam which is pivoting on an axis at the origin of the plane, and shoots in a random direction. Take a horizontal wall at unit distance, of equation $y = 1$. The laser beams marks a random point on this wall; it is a small computation (which makes an interesting geometry exercise) to prove that the random variable so obtained has density $\frac{1}{\pi}\frac{1}{1+x^2}$. As we remarked above, this random variable has no expected value; it would be easy to make experiments, by choosing random numbers in the interval $(-\frac{\pi}{2}, \frac{\pi}{2})$ and taking their tangent, for example to test the behaviour of the average.

Limits of the model. This mechanical model has some limitations. First of all, it becomes quickly necessary to give explicit analytic formulas (and this must be one of the aims of a probability course), and to go beyond the pictures.

In that case, the simple model with atoms becomes insufficient, as it can only model correctly finite probability; but this can be a good occasion to introduce a number of questions on the foundations of probability and analysis, up to and including measure theory and distributions, if the assistance allows it.

It is unclear to us how well this model can motivate the important questions of convergence (which could be leaved aside in a first presentation); however, the model for the Cauchy density given above could motivate the question (and some models without variance for the income distribution could be used to show its practical interest).

More importantly, it is clear that this presentation only makes sense for students who have some basic knowledge of mechanics (center of mass, moment of inertia, and their computation in some simple situations). But this is a part of a more general fact: intuition can only be built on the foundation of some previous knowledge, informal or formal, and in this sense, it is wrong to oppose learning and understanding: intuition of sophisticated phenomena generally needs some technical culture.

5. CONCLUSION

We think that understanding a mathematical notion means building an internal (mental) representations which must be sufficiently adapted to the object. We build and internally manipulate these representations (and also the representations from the representations and so on). We believe that the first efficient representations are, in general, concrete (visual, kinetic, auditive,...) and they allow to build from them other more sophisticated symbolic representations. Concrete mental representations (inspired by the real physical world) are interesting for all sciences, including pure and theoretical mathematics and physics. If the thoughts of children are based on a simple and naive physical world (even if we know now that this world does not really exist !), the thoughts of students and colleagues are also based on the internalization of the physical world increased by different knowledges more and

more symbolically sophisticated. Our opinion is that the more we are able to build mental representations, the more we are able to think efficiently as student or searcher.

It is however not trivial to build an effective mental representation. It would be valuable to gather many different mental representations, and to test their efficiency, the different parameters that play a role in their reception by the students, and the conditions under which they can be useful; this is the effort we started at a small scale.

The reader has probably observed that there is no picture in this paper; there will be in the journal version.

References

[COT07] M.P. Carlson, M. Oehrtman, and P.W. Thompson. Foundational reasoning abilities that promote coherence in students understanding of function. In *Making the connection: Research and teaching in undergraduate mathematics*, pages 150–171. MAA, 2007.

[DMK04a] M. Denis, E. Mellet, and SM. Kosslyn. Neuroimaging of mental imagery. *Psychology Press*, 2004.

[DMK04b] M. Denis, E. Mellet, and SM. Kosslyn. Neuroimaging of mental imagery: An introduction. *European Journal of Cognitive Psychology*, 16:625–630, 2004.

[GJ98] Gerald A. Goldin and Claude Janvier. Representations and the psychology of mathematics education. *Journal of Mathematical Behavior*, 17(1), 1998.

[LCS07] S. Lacey, C. Campbell, and K. Sathian. Vision and touch: Multiple or multisensory representations of objects? *Perception*, 36(10):1513– 1521, 2007.

[Tho96] P.W. Thompson. Imagery and the development of mathematical reasoning. In *Theories of learning mathematics*, pages 267–283. Hillsdale, NJ: Erlbaum, 1996.

(Arnoux) Institut de Mathématiques de Luminy, CNRS U.M.R. 6206, 163, avenue de Luminy, Case 907, 13288 Marseille cedex 09, France

E-mail address, Arnoux: arnoux@iml.univ-mrs.fr

(Finkel) LSV, ENS Cachan & CNRS 61, av. du Président Wilson, 94230 Cachan, France

E-mail address, Finkel: finkel@lsv.ens-cachan.fr