

Optimization Techniques under Uncertain Criteria, and Their Possible Use in Computerized Education

Richard Aló
Center for Computational Sciences and
Advanced Distributed Simulation
University of Houston-Downtown
One Main Street
Houston, TX 77002, USA
RALo@uh.edu

Olga Kosheleva
Department of Teacher Education
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
olgak@utep.edu

Abstract—The existing successful automated computerized systems more or less simulate the way successful human teachers teach. However, computerized systems provide more individualized options than traditional classroom education, and it is desirable to use this additional freedom to further improve the education success rate. In this papers, we briefly overview the experience of a successful Russian training system, and explain how general techniques of optimization under uncertainty can be used to optimize the content development.

I. RUSSIAN TRAINING SYSTEM: BRIEF OVERVIEW

A. Main Idea Behind the System

The training system that we are describing here was designed by A. M. Zimichev et al. in the 1980s; see, e.g., [10], [11]. It started with an attempt to resolve the following seeming contradiction.

On the one hand, the vast majority of pedagogues strongly supports the humanitarian belief that (almost) students are equal in their ability to learn. Some students may require different learning styles, different support environments, but the history of education has consistently shown that any group originally perceived as inferior in learning – be it by gender or by race or by country of origin – turns out to be as successful as everyone else. Standardized tests consistently prove that kids from different groups have, on average, the same ability to learn. There are big discussions on whether 10% differences between scores are meaningful or not, but the very size of these disputed differences confirms the big pictures: we are all born equally skilled in learning.

Based on the fact that we are all equally skilled in learning, one should expect that with advanced learning strategies, all kids should succeed equally in school. Alas, no matter how advanced is the pedagogical technique that we use in the actual classroom, usually, after each class, there is a big difference between how much different kids learned, difference often in the orders of magnitude.

Since this difference cannot be explained by the difference in the kid's ability to learn, then why is it there? and, most importantly, how can we overcome this difference and enable each kid to learn according to his or her full potential?

To understand the empirically observed difference, researchers decided to quantify how much the students recalled. They asked each student, after the class, to spend an hour or so writing down what this student remembered after the class, be it the material from the class, what was happening outside, etc. Then, the researchers tried their best to quantify this amount of information by counting it in bits.

When they counted the number of bits *relevant* to the material presented in the class (e.g., math), they got exactly the same order of magnitude difference as was expected. Surprisingly, however, when they counted the *total* amount of information, including both the information presented in the class and the irrelevant information (what was the teacher dressed in, what birds were signing, what kids at neighboring desks were saying, etc.), then the total amounts of information recalled by all the kids became approximately the same.

In other words, both the best learners and the worst learners remembered approximately the same amount of information, but the best learners remembered the material presented in the class, while the worst learners remembered a lot of irrelevant details but not what was taught.

In view of this discovery, how can we improve the learning situation? The two natural strategies are to decrease the amount of irrelevant information, and to better motivate the students to learn relevant information as opposed to the irrelevant one. How can we implement these two strategies?

B. How to Reduce Irrelevant Information and Better Motivate Students

In order to find out how to reduce the amount of irrelevant information, let us enumerate where this information comes from, and eliminate these sources one by one.

Some irrelevant information comes from outside the classroom. To eliminate this information, we performed teaching in classroom without windows.

Some irrelevant information comes from the classroom walls. Portraits of famous scientists, periodic tables, copies of famous paintings, etc., are useful tools, but they turn out to be one of the sources of distraction. So, we performed education in rooms with empty walls.

Moreover, since a student remembers the exact same amount of information anyway, we decrease the amount of irrelevant information by using all the channels – vision, hearing, etc. – to transmit useful information. In other words, we use multi-media delivery of useful information.

Some irrelevant information comes from the teacher: the way he or she dresses, the tone of the teacher’s voice, etc. To eliminate this source of distraction, we tried to avoid using teachers as much as possible. Instead, we used video and audio recordings in most classes.

Some information comes from the other students in the class. The need to exchange this information comes from the fact that in each class, there are students with different learning curves and learning styles, so at any given moment of time, some students have grasped the material while the others are still struggling.

- Students who are struggling to understand tend to ask questions to those who already understand;
- students who are ahead get bored and start absorbing irrelevant information.

To eliminate this source of irrelevant information, we group students by similarity of their learning styles, so that all students in each group learn approximately at the same pace.

Finally, some information comes from the student’s body: what a student ate, whether the student is sleepy or not, etc. Whether this information affects the student or not depends on the student’s state of mind:

- if a student is focused on learning he mostly ignores this information;
- if a student is not focused on learning, this information may be all he learns.

Therefore, to decrease this information, we must move the student to a state of mind which is most appropriate for learning. A state of the human body can be characterized by the corresponding frequencies: e.g., slow heartbeat and slow breathing correspond to a quiet state, while faster heartbeat and breathing describe a more active state; an α -rhythm on an EEG indicates sleep, while other rhythms indicate active brain, etc. It is difficult to control the state directly, but it is well known that we can affect these frequencies by using the resonance phenomenon:

- we start with music-like tones whose frequencies are close to the current frequencies of a student; as a result, the physiological characteristics of the body get into resonance with these outside frequencies;
- we then slowly change the frequencies of the music in the direction in which we want the physiological frequencies to change so as to achieve the state which is most useful for learning; as a result, the physiological frequencies also change in the desired direction.

This resonance control phenomenon is well known both in engineering [4] and in psychology: e.g., it is well known that a slow music calms down, while a music that gets faster and faster excites. Our experience shows that this phenomenon works perfectly for learning as well.

C. Additional Features of the Systems

We have enumerated all the ways how we decreased the amount of irrelevant information. Since we decreased the amount of irrelevant information, we can thus afford to increase the amount of useful information. In other words, we can afford to deliver the useful information at a breathtaking speed.

With such speeds, the students get tired fast. So, after a brief intense lesson, we had intense rest.

- Periodically, it was pure rest, where students would be induced – by a similar resonance techniques – into a quiet, almost sleepy state.
- Periodically, it was active rest, again easily inducible by resonance.

D. The System Was Successfully Used

This technique was successfully used for training basic non-creative skills such as learning to type fast, learning words from a foreign language, cross-country driving (e.g., driving a tank), etc.

II. CASE STUDY: OPTIMIZING CONTENT DEVELOPMENT

A. Why the Need for Optimization

We have already mentioned that in our system, the speed of information delivery was much higher than in normal learning. This is not just a specific feature of our system: the increase in learning speed is one of the main reasons why automated learning systems are developed in the first place.

It is well known that in human learning, successful intensive learning courses do not simply consist of teaching the same material but faster, usually, the order of presentation and the amount of time spent on different repetitions are different in an intensive course than in a regular course. It is therefore reasonable, instead of simply following the usual learning order, to look for a new order that would be best for intensive automated teaching.

B. Basic Assumptions About Training

In order to find the optimal training schedule, let us make some (simplifying but realistic) assumptions about training.

In principle, there are several different types of items that we want a student to learn; for example:

- when we teach typing, we want the student to acquire the motor skills of typing all the symbols on the keyboard and all pairs of consequent symbols;
- when we teach words from a foreign language, we want the student to learn all these words;
- when we teach cross-country driving, we want the student to develop motor skills corresponding to different types of terrain: flat surface, rugged terrain, uphill, downhill, narrow bridge, etc.

Let us denote the total number of types to learn by T . For simplicity, we assume that acquiring skills necessary for each of these types takes the same number of training situations s . So, to learn all necessary types, a student needs at least

$T \cdot s$ repetitions. If we denote, by T_0 , the time necessary for handling each repetition, then the total time for training a student for all necessary types is equal to $T_0 \cdot T \cdot s$.

In many learning situations, the total number T of necessary types is large, so the above time of total training is unrealistically large. Therefore, we cannot expect every single student to be 100% skilled in every possible situation type.

Since we cannot train a student to be skilled in every possible thing, it is therefore necessary to train a student in such a way that this doctor will be able to handle *the largest possible number* of these types.

In future applications, some of these types are more frequent, some are less frequent. So, if we know that a student can only learn, say, t different words, and we have to choose which of these words the student will learn perfectly well, we should choose t most frequent ones.

A *skill* of a student can be thus characterized by the number t of the types of items in which this student is well skilled.

To estimate frequencies of different types, we can use a general (semi-empirical) law discovered by G. K. Zipf (see, e.g., [7], [12]), according to which, if we order types from the most frequent to the least frequent one, then the frequency f_i of i -th type is proportional to $1/i$: $f_i = c/i$ for some constant c . The value of this constant can be determined from the fact that the sum of all these frequencies should be equal to 1: $f_1 + \dots + f_T = 1$. Since $1 + 1/2 + \dots + 1/T \approx \ln(T)$, we thus conclude that $c \cdot \ln(T) = 1$, $c = 1/\ln(T)$, and

$$f_i = \frac{1}{\ln(T) \cdot i}. \quad (1)$$

C. Traditional Learning

In traditional learning of a language, a student is trained on texts from real language. In traditional typing lessons, student learn to type by typing real-life texts. In all these cases, a student is trained on a real-life flow of items.

Let us denote by I the time allocated for training. Since handling each repetition takes time T_0 , during this training time, the trainee will see $N = I/T_0$ repetitions. According to our assumption about the training time, the student will be trained only in those types i for which he or she has seen at least s repetitions. Out of the total of N repetitions, the student will see $N \cdot f_i$ repetitions of i -th type; so, the student will be trained in all the types for which $N \cdot f_i \geq s$. Substituting Zipf's expression (1) for f_i , we conclude that the student will learn all the types i for which

$$\frac{I}{T_0} \cdot \frac{1}{\ln(T) \cdot i} \geq s,$$

i.e., for which

$$i \leq \frac{I}{T_0 \cdot \ln(T) \cdot s}.$$

Therefore, the resulting student's skill level t (i.e., the total number of types in which this student will be skilled), will be equal to

$$t = \frac{I}{T_0 \cdot \ln(T) \cdot s}. \quad (2)$$

This formula describes the skill level acquired during a given training time I .

We can also consider the inverse problem: we want a student to be trained for a certain skill level t , and we need to know the time I required for this training. From the formula (2), we can conclude that

$$I = t \cdot T_0 \cdot \ln(T) \cdot s. \quad (3)$$

D. Optimal Training

In automated training, we can generate repetitions in arbitrary order, not necessarily with real-life frequencies. If we want a student to be trained on t different types, then we need to generate exactly s repetitions of this type.

If we fix the total training time I , then during this time, we can generate $N = I/T_0$ repetitions. Since learning each type requires s repetitions, the total amount of different types in which a student can get skilled is equal to $t = N/s = (I/T_0) \cdot s$. Thus, after this training, the student will acquire the skill level

$$t = \frac{I}{T_0 \cdot s}. \quad (4)$$

This formula describes the skill level acquired during a given training time I .

We can also consider the inverse problem: we want a doctor to be trained for a certain skill level t , and we need to know the time I required for this training. From the formula (4), we can conclude that

$$I = t \cdot T_0 \cdot s. \quad (5)$$

E. Conclusion: Automated Training is Faster and Better

By comparing the formulas (2) and (4), we conclude that during the same training time, the skill level acquired during the automated training can be much higher ($\ln(T)$ times higher) than the skill level acquired in traditional training.

Similarly, by comparing the formulas (3) and (5), we conclude that the training time necessary to acquire a given skill can be much shorter ($\ln(T)$ times shorter) for the automated training than for traditional training.

F. How to Optimally Combine Automated and Traditional Training

Designing an automated learning system requires a lot of computer work and a lot of programming. At first, therefore, automated training will not be available for the whole training; realistically, we should expect that only a part of the training is done on an automated system, and after this basic training, a trainee goes into a traditional training. How can we best organize this combined training?

Let us denote the time that we can allocate for automated training by I_{au} , and the training time for the follow-up traditional training by I_{tr} . During the follow-up training, the student encounters $N_{\text{tr}} = I_{\text{tr}}/T_0$ repetitions. Of these repetitions, $N_{\text{tr}} \cdot f_i$ are of type i .

If this number of repetitions is $\geq s$, then for this type, the student acquires necessary skills during the follow-up training,

so there is no need to simulate patients of this type during the automated training. Thus, we get all types from 1 to

$$t_{\text{tr}} = \frac{I_{\text{tr}}}{T_0 \cdot \ln(T) \cdot s} \quad (6)$$

covered.

For each type $i > t_{\text{tr}}$, we get

$$f_i \cdot N_{\text{tr}} = \frac{I_{\text{tr}}}{T_0 \cdot \ln(T) \cdot i} < s$$

repetitions covered during traditional training. So, if we want the student to get the necessary skills, we must generate the remaining number of repetitions

$$n_i = s - \frac{I_{\text{tr}}}{T_0 \cdot \ln(T) \cdot i} \quad (7)$$

during the automated training.

We want to learn as many new types as possible. How many situation types can we thus learn? During the time I_{au} , we can only generate $N_{\text{au}} = I_{\text{au}}/T_0$ repetitions. Since learning type i requires n_i repetitions, the skill level t acquired by a doctor can be determined by the formula

$$\frac{I_{\text{au}}}{T_0} = N_{\text{au}} = \sum_{i=t_{\text{tr}}}^t n_i.$$

Substituting the above expression for n_i , we conclude that

$$\frac{I_{\text{au}}}{T_0} = s \cdot (t - t_{\text{tr}}) - \frac{I_{\text{tr}}}{T_0 \cdot \ln(T)} \cdot \sum_{i=t_{\text{tr}}}^t \frac{1}{i}.$$

Since $1 + 1/2 + \dots + 1/i \approx \ln(i)$, we can rewrite this equation as

$$\frac{I_{\text{au}}}{T_0} = s \cdot (t - t_{\text{tr}}) - \frac{I_{\text{tr}} \cdot (\ln(t) - \ln(t_{\text{tr}}))}{T_0 \cdot \ln(T)}. \quad (8)$$

So, we can make two conclusions:

- If the training times I_{au} and I_{tr} are given, then the resulting acquired skill t can be determined from the equation (8), where t_{tr} is determined from the equation (6).
- Vice versa, if we know the training time I_{au} for the automated training, and the required skill level t , then we must find t_{tr} for the equation (8), and then use the formula (6) to determine the necessary traditional training period as

$$I_{\text{tr}} = t_{\text{tr}} \cdot T_0 \cdot \ln(T) \cdot s. \quad (9)$$

In both cases, the number of repetitions of different types $i = t_{\text{tr}}, t_{\text{tr}} + 1, \dots, t$ generated during the automated training is determined by the formula (7).

G. Other Application

In [1], we used a similar idea to optimize the types of virtual patients used by doctors during medical training – specifically, during a training of surgeons for spinal cord stimulation procedures; see, e.g., [2], [3], [5], [9].

III. CASE STUDY CONTINUED: OPTIMAL ORDER OF PRESENTING THE MATERIAL

A. Formulation of the Problem

In the above section, we described the optimal frequencies with which we repeat each of the items that a student has to learn. Once we know the number of repetitions of each item, the next natural question is: in what order should we present these repetitions? Should we first present all the repetitions of item 1, then all the repetitions of item 2, etc., or should we randomly mix these repetitions?

B. Towards Mathematical Formulation of the Corresponding Optimization Problem

Each item is characterized by several (n) numerical characteristics, so we can geometrically represent each item as a point in the corresponding n -dimensional space.

- Similar items have close values of these characteristics, so the distance between the points corresponding to similar items is small.
- Vice versa, when the items are different, they at least some of these characteristics have different values on these items, so the resulting distance is large.

Thus, the distance between the corresponding points in a multi-D space can be viewed as a measure of similarity between the items.

In terms of multi-D space, an order in which we present repetitions is described as a function $x(t)$, where x is a multi-D point corresponding to the item presented at moment $t = k \cdot \Delta t$, where Δt is the time between repetitions.

As we have mentioned, when we have a few items to learn, we can easily learn them all, so there is no need for sophisticated optimization. Optimization becomes necessary when there are many items – and thus, many repetitions. In this case, similar to the way we simplify the physical problems if we approximate a collection of atoms by a continuous medium, we can approximate the discrete dependence $x(t)$ on discrete time t by a continuous function $x(t)$ of continuous time t .

What is the optimal trajectory $x(t)$? The experience of learning shows that often, presenting the items in random order is beneficial. To allow for this possibility, instead of looking for a deterministic function $x(t)$, we look for *random* processes $x(t)$. Since a deterministic function is a particular case of a random process, we are thus not restricting ourselves.

Let us consider Gaussian random processes. A Gaussian random process can be uniquely characterized by its mean $m(t) \stackrel{\text{def}}{=} E[x(t)]$ and autocorrelation function $A(t, s) \stackrel{\text{def}}{=} E[(x(t) - x(s))^2]$.

Students come with different levels of preparation. Therefore, a good learning strategy should work not only for a student that comes from 0, but also for a student that comes at moment t_0 with the knowledge that other students have already acquired by this time. From this viewpoint, a student's education starts at the moment t_0 . It is therefore natural to require that the random process should look the same whether

we start with a point $t = 0$ or with some later point t_0 . Hence, the characteristics of the process should be the same, i.e., $m(t) = m(t + t_0)$ and $A(t, s) = A(t + t_0, s + t_0)$ for every t , s , and t_0 .

From the first condition, we conclude that $m(t) = \text{const}$. Thus, by changing the origin of the coordinate system, we can safely assume that $m(t) = 0$.

From the second condition, for $t_0 = -s$, we conclude that $A(t, s) = A(t - s, 0)$, i.e., that the autocorrelation function depends only on the difference between the times: $A(t, s) = a(t - s)$, where we denoted $a(t) \stackrel{\text{def}}{=} A(t, 0)$. In other words, the random process must be *stationary*.

The final question is: what autocorrelation function $a(t)$ should we use?

C. We Must Choose a Family of Functions, Not a Single Function

The function $a(t)$ depends on how intensely we train. In more intensive training, we present the material faster, and thus, within the same time interval t , we can cover more diverse topics. More diverse topics means that the average change $a(t)$ can be larger. A natural way to describe this increase is by proportionally enlarging all the distances, which leads from $a(t)$ to $C \cdot a(t)$. In other words, if $a(t)$ is a reasonable function for some training, then a new function $C \cdot a(t)$ should also be reasonable.

We can say that the functions $a(t)$ and $C \cdot a(t)$ describe exactly the same learning strategy, but with different intensities. Since intensity can be different, we cannot select a unique function $a(t)$ and claim it to be the best, because for every function $a(t)$, the function $C \cdot a(t)$ describes exactly the same learning strategy. In view of this, instead of formulating a problem of choosing the best autocorrelation *function*, it is more natural to formulate a problem of choosing the best *family* $\{C \cdot a(t)\}_C$ of autocorrelation functions.

D. Which Family Is the Best? We May Need Non-Numerical Optimality Criteria

Among all the families $\{C \cdot a(t)\}_C$, we want to choose the best one.

In mathematical optimization problems, numerical criteria are most frequently used, when to every alternative (in our case, to each family) we assign some value expressing its performance, and we choose an alternative (in our case, a family) for which this value is the largest. In our problem, as such a numerical criterion, we can select, e.g., the average grade on some standardized test A .

However, it is not necessary to restrict ourselves to such numerical criteria only. For example, if we have several different families that have the same average average grade A , we can choose between them the one that has the minimal level of uncomfortableness U . In this case, the actual criterion that we use to compare two families is not numerical, but more complicated: *a family F_1 is better than the family F_2 if and only if either $A(F_1) < A(F_2)$, or $A(F_1) = A(F_2)$ and $U(F_1) < U(F_2)$* . A criterion can be even more complicated.

What a criterion *must* do is to allow us, for every pair of families, to tell whether the first family is better with respect to this criterion (we'll denote it by $F_1 \succ F_2$), or the second is better ($F_1 \prec F_2$), or these families have the same quality in the sense of this criterion (we'll denote it by $F_1 \sim F_2$). Of course, it is necessary to demand that these choices be consistent, e.g., if $F_1 \prec F_2$ and $F_2 \prec F_3$ then $F_1 \prec F_3$.

E. The Optimality Criterion Must Select a Unique Optimal Family

Another natural demand is that this criterion must choose a *unique* optimal family (i.e., a family that is better with respect to this criterion than any other family). The reason for this demand is very simple.

If a criterion does not choose a family at all, then it is of no use.

If several different families are “the best” according to this criterion, then we still have a problem to choose among those “best”. Therefore, we need some additional criterion for that choice. For example, if several families turn out to have the same average grade, we can choose among them a with the minimal uncomfortableness.

So what we actually do in this case is abandon that criterion for which there were several “best” families, and consider a new “composite” criterion instead: F_1 is better than F_2 according to this new criterion if either it was better according to the old criterion or according to the old criterion they had the same quality and F_1 is better than F_2 according to the additional criterion.

In other words, if a criterion does not allow us to choose a unique best family, it means that this criterion is not final. We have to modify it until we come to a final criterion that will have that property.

F. The Optimality Criterion Must Be Scale-Invariant

The next natural condition that the criterion must satisfy is connected with the fact that the numerical value of the time t depends on the choice of the unit for measuring time.

If we replace the original unit of time by a new unit which is λ times larger (i.e., replace minutes by hours), then numerical values change from t to $\tilde{t} = t/\lambda$. The autocorrelation function that in the old units is described by a family $\{C \cdot a(t)\}$, in the new units, has a new form $\{C \cdot a(\lambda \cdot t)\}$,

Since this change is simply a change in a unit of time, it is reasonable to require that going from $a(t)$ from $a(\lambda \cdot t)$ should not change the *relative* quality of the autocorrelation functions, i.e., if a family $\{C \cdot a(t)\}_C$ is better than the family $\{C \cdot a'(t)\}_C$, then for every $\lambda > 0$, the family $\{C \cdot a(\lambda \cdot t)\}_C$ must be still better than the family $\{C \cdot a'(\lambda \cdot t)\}_C$.

G. Definitions and the Main Result

Definition 1:

- By an autocorrelation function we mean a monotonically non-strictly decreasing function from non-negative real numbers to non-negative real numbers.

- By a family of functions we mean the family $\{C \cdot a(t)\}_C$, where $a(t)$ is a given autocorrelation function and C runs over arbitrary positive real numbers.
- A pair of relations (\prec, \sim) is called consistent [8] if it satisfies the following conditions:
 - (1) if $a \prec b$ and $b \prec c$ then $a \prec c$;
 - (2) $a \sim a$;
 - (3) if $a \sim b$ then $b \sim a$;
 - (4) if $a \sim b$ and $b \sim c$ then $a \sim c$;
 - (5) if $a \prec b$ and $b \sim c$ then $a \prec c$;
 - (6) if $a \sim b$ and $b \prec c$ then $a \prec c$;
 - (7) if $a \prec b$, then $b \prec a$ or $a \sim b$ are impossible.

Definition 2:

- Assume a set A is given. Its elements will be called alternatives. By an optimality criterion we mean a consistent pair (\prec, \sim) of relations on the set A of all alternatives. If $b \prec a$, we say that a is better than b ; if $a \sim b$, we say that the alternatives a and b are equivalent with respect to this criterion.
- We say that an alternative a is optimal (or best) with respect to a criterion (\prec, \sim) if for every other alternative b either $b \prec a$ or $a \sim b$.
- We say that a criterion is final if there exists an optimal alternative, and this optimal alternative is unique.
- Let $\lambda > 0$ be a real number. By the λ -rescaling $R_\lambda(\rho)$ of a function $a(t)$ we mean a function $(R_\lambda a)(t) \stackrel{\text{def}}{=} a(\lambda \cdot t)$.
- By the λ -rescaling $R_\lambda(F)$ of a family F , we mean the set of the functions that are obtained from $f \in F$ by λ -rescaling.

In this paper, we consider optimality criteria on the set \mathcal{F} of all families.

Definition 3: We say that an optimality criterion on F is scale-invariant if for every two families F and G and for every number $\lambda > 0$, the following two conditions are true:

- if F is better than G in the sense of this criterion (i.e., $G \prec F$), then

$$R_\lambda(G) \prec R_\lambda(F);$$

- if F is equivalent to G in the sense of this criterion (i.e., $F \sim G$), then

$$R_\lambda(F) \sim R_\lambda(G).$$

As we have already remarked, the demands that the optimality criterion is final and scale-invariant are quite reasonable. The only problem with them is that at first glance they may seem rather weak. However, they are not, as the following theorem shows:

Theorem 1: [8] If a family F is optimal in the sense of some optimality criterion that is final and scale-invariant, then every density function $\rho(r)$ from this optimal family F which has the form $a(t) = A \cdot t^\alpha$ for some real numbers A and α .

In other words, the optimal configuration is a fractal random process. When $\alpha = 2$, we have a straightforward trajectory, without any randomness. The value $\alpha = 0$ means that values of

$x(t)$ and $x(s)$ for $t \neq s$ are completely uncorrelated, i.e., that we have a white noise. Intermediate values of α correspond to different levels of randomness.

Our experience showed that such fractal order indeed leads to improvement in learning. The exact value of the parameter α – corresponding to the fractal dimension of the corresponding trajectories – should be adjusted to the learning style of the students. A natural idea is to adjust it to the fractal dimension of their EEG characteristics.

IV. OTHER POSSIBLE APPLICATIONS OF OPTIMIZATION TO COMPUTERIZED EDUCATION

In the above text, we used an invariance-based approach to optimization. This approach is not limited to the above problem. In [8], this approach was used to provide theoretical justification for empirical heuristic formulas such as the choice of activation function $s(x) = 1/(1 + e^{-x})$ in artificial neural networks, the choice of “and”- and “or”-operations in fuzzy logic, etc.

This approach can not only be used to justify the existing formulas when we already know, from experiments, which formulas are empirically the best. It can also provide us with a guidance on which formulas could be best when no optimal formulas are known.

We hope that this approach will be useful in finding the best neural networks for training the education-related programs, for best fuzzy logic connectives in the analysis of expert rules, etc.

REFERENCES

- [1] R. Aló, K. Aló, and V. Kreinovich, “Towards Intelligent Virtual Environment for Training Medical Doctors in Surgical Pain Relief”, *Proceedings of The Eighth International Fuzzy Systems Association World Congress IFSA'99*, Taipei, Taiwan, August 17–20, 1999, pp. 260–264.
- [2] G. Barolat, “Current status of epidural spinal cord stimulation”, *Neurosurgery Quarterly*, 1995, Vol. 5, No. 2, pp. 98–124.
- [3] G. M. Barolat, J. He, S. Zeme, and B. Ketcik, “Mapping of sensory responses to epidural stimulation of the intraspinal neural structures in man”, *J. Neurosurgery*, 1993, Vol. 78, pp. 233–239.
- [4] I. I. Blekhman and I. I. Blekhman, *Vibrational Mechanics: Nonlinear Dynamic Effects, General Approach, Applications*, World Scientific, Singapore, 1999.
- [5] S. Horsch and L. Clayes, *Spinal cord simulation II*, Steinkopff Verlag, Darmstadt, 1995.
- [6] V. Kreinovich, E. Johnson-Holubec, L. K. Reznik, and M. Koshelev, “Cooperative learning is better: explanation using dynamical systems, fuzzy logic, and geometric symmetries”, In: Nguyen Hoang Phuong and A. Ohsato (eds.), *Proceedings of the Vietnam-Japan Bilateral Symposium on Fuzzy Systems and Applications VJFUZZY'98*, HaLong Bay, Vietnam, 30th September–2nd October, 1998, pp. 154–160.
- [7] B. B. Mandelbrot, *The fractal geometry of Nature*, Freeman, San Francisco, 1982.
- [8] H. T. Nguyen and V. Kreinovich, *Applications of continuous mathematics to computer science*, Kluwer, Dordrecht, 1997.
- [9] U. Rossi and J. Vernes, “Epidural spinal electrical stimulation for pain control: a ten-year experience”, *Proc. XII Annual Meeting of Australasian College of Rehabilitation Medicine*, 1992, pp. 17–21.
- [10] A. M. Zimichev, V. Kreinovich, et al. *Automatic control system for psychological experiments*, Final Report on Grant No. 0471-2213-20 from the Soviet Science Foundation, 1982 (in Russian).
- [11] A. M. Zimichev, V. Kreinovich, et al. *Automatic system for quick education of computer operators*, Final Report on Grant No. 0471-0212-60 from the Soviet Science Foundation, 1980–82 (in Russian).
- [12] G. K. Zipf, *Human behavior and the principle of least-effort*, Addison-Wesley, Cambridge, MA, 1949.