Geometric Reformulation of Learning Models Can Help Prepare Better Teachers

Vladik Kreinovich[1] and Olga Kosheleva[2]

[1]Department of Computer Science
[2]Department of Teacher Education
University of Texas at El Paso
500 W. University
El Paso, Texas 79968, USA
olgak@utep.edu, vladik@utep.edu

1. Motivations

- Teaching is important but difficult, and preparing teachers is also an important and difficult task.
- Many researchers are analyzing how to improve teacher preparation.
- When preparing teachers, it is desirable to use the corresponding research results.
- Most of these research results are based on complex models and/or on complex data analysis.
- As a result, future teachers receive the advice as a black box, whose motivation they do not quite understand.
- To make sure that the future teacher follow the advice, we need to clarify its motivations.

2. Motivations (cont-d)

- A clear geometric picture is easier to understand that the corresponding formula.
- So, to make pedagogical recommendations clearer, we can reformulate them in geometric form.
- In this talk, we provide two cases of such reformulation.
- Both cases deal with the order in which we present different parts of the material.

3. Concrete-First vs. Abstract-First

- Each topic in math and sciences usually contains some new abstract notion(s) and related examples.
- We can start with examples, and delay the introduction of abstract ideas and notions as much as possible.
- Alternatively, we can start with the abstract ideas.
- Intuitively, it seems that students would learn better if they are first presented with numerous examples.
- However, empirically, the abstract-first approach often works better.
- We provide a geometric explanation for this seemingly counter-intuitive empirical phenomenon.

4. Learning: a Natural Geometric Representation

- The process of learning means that we change the state of a student:
  - from a state in which the student did not know the material (or does not have the required skill) to
  - a state in which the student has knowledge of the required material (or has the required skill).
- Let s_0 denote the original state of a student.
- Let S denote the set of all the states corresponding to the required knowledge or skill.
- We start with a state which is not in the set S ($s_0 \notin S$).
- We end up in a state s which is in the set S.
- Let $d(s, s')$ be the difficulty (time, effort, etc.) needed to go from state s to state s'.

5. Learning (cont-d)

- Our objective is to help the students learn in the easiest (fastest, etc.) way.
- In terms of d, this means that we want to go:
  - from the original state $s_0 \notin S$
  - to the state $s \in S$ for which the effort $d(s_0, s)$ is the smallest possible.
- In geometric terms, the smallest possible effort means the shortest possible distance.
- Thus, our objective is to find the state $s \in S$ which is the closest to s_0.
- Such closest state is called the *projection* of the original state s_0 on the set S.

6. Learning Complex Material: Geometric Interpretation

- Let us take into account that the material to be learned consists of several pieces.
- Let S_i, $1 \le i \le n$, denote the set of states in which a student has learned the i-th part of the material.
- Our ultimate objective is to make sure that the student learns all the parts of the material.
- In terms of states, learning the i-th part of the material means belonging to the set S_i.
- Thus, the student should end up in a state which belongs to all the sets $S_1, \ldots, S_n$.
- In other words, we want a state from the intersection $S \stackrel{\text{def}}{=} S_1 \cap \ldots \cap S_n$.

7. Learning Complex Material (cont-d)

- In these terms, if we present the material in the order S_1, S_2, …, S_n, this means that:
  - we first project s_0 onto the set S_1, resulting in a state $s_1 \in S_1$ which is the closest to s_0;
  - then, we project s_1 onto the set S_2, resulting in a state $s_2 \in S_2$ which is the closest to s_1;
  - …
  - at the last stage of the cycle, we project s_{n-1} onto S_n, resulting in $s_n \in S_n$ which is the closest to s_{n-1}.
- In some cases, we end up learning all the material – i.e., in a state $s_n \in S_1 \cap \ldots \cap S_n$.
- However, often, by the time the students have learned S_n, they have forgotten some earlier material.
- So, we need to repeat this process again (and again).

8. This Geometric Interpretation Makes Computational Sense

- The above "sequential projections" algorithm is actually actively used in many applications.
- Sometimes, all the sets S_i are convex.
- Then, the Projections on Convex Sets (POCS) method guarantees convergence to a point from $S_1 \cap \ldots \cap S_n$.
- In our terms, this means that the students will eventually learn all parts of the necessary material.
- In the more general non-convex case, the convergence is not always guaranteed.
- However, the method is still efficiently used, and often converges.

9. The Simplest Case: Two-Part Knowledge

- In this simplest case, there are only two options.
- We begin by studying S_1, then, we study S_2, then S_1 again, etc.
- We begin by studying S_2, then, we study S_1, then S_2 again, etc.
- We want to get from the original state s_0 to the state $\widetilde{s} \in S_1 \cap S_2$ which is the closest to s_0.
- The effectiveness of learning is determined by how close we get to the desired set $S = S_2 \cap S_2$.
- It is natural to conclude that the amount of this knowledge is reasonably small.
- Indeed, otherwise, we would have divided into a larger number of easier-to-learn pieces.

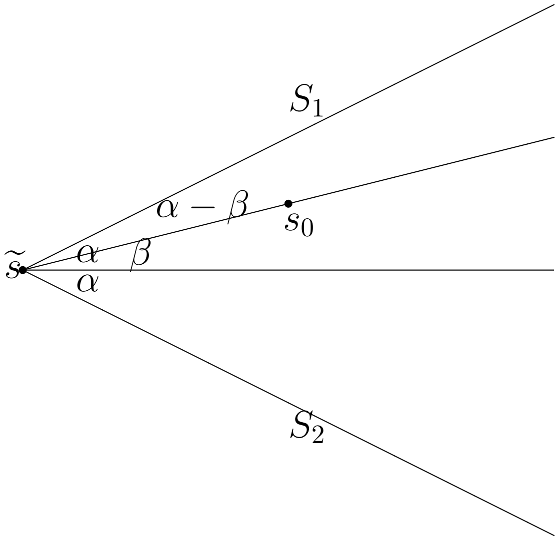
10. The Simplest Case (cont-d)

- In geometric terms, this means that the original state s_0 is close to the desired intersection set $S_1 \cap S_2$.
- So, the distance $d_0 \stackrel{\text{def}}{=} d(s_0, \widetilde{s})$ is reasonably small.
- So, all the states are close to each other.
- Thus, in the vicinity of the state $\widetilde{s}$, we can:
  - expand the formulas describing the borders of the sets S_i into Taylor series and
  - keep only terms which are linear in the (coordinates of the) difference $s - \widetilde{s}$.
- Thus, it is reasonable to assume that the border of each of the two sets S_i is described by a linear equation.
- This border is hence a (hyper-)plane: a line in 2-D space, a plane in 3-D space, etc.

11. The Simplest Case (cont-d)

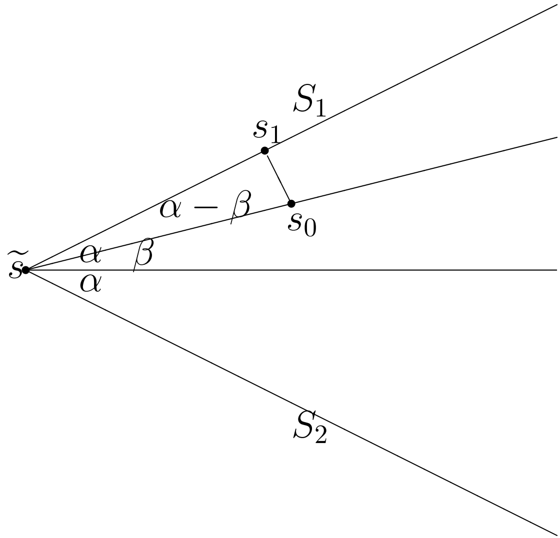
- As a result, we arrive at the following configuration.
- Let 2α denote the angle between the borders of the sets S_1 and S_2.
- So the angles between each of these borders and the midline is exactly α.
- Let β denote the angle between the direction from $\widetilde{s}$ to s_0 and the midline.
- In this case, the angle between the border of S_1 and the midline is equal to $\alpha - \beta$.
- So, we arrive at the following configuration:

12. The Simplest Case (cont-d)



- In the first option, we first project s_0 onto the set S_1.
- As a result, we get the following configuration:

13. The Simplest Case (cont-d)



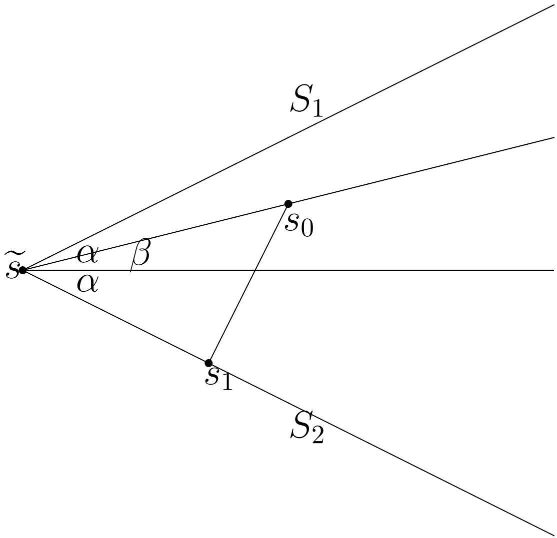
- Here, the projection line $s_0 s_1$ is orthogonal to the border of S_1.

14. The Simplest Case (cont-d)

- From the right triangle $\triangle \widetilde{s} s_0 s_1$, we conclude that
$$d_1 \stackrel{\text{def}}{=} d(\widetilde{s}, s_1) = d_0 \cdot \cos(\alpha - \beta).$$
- On the next step, we project the point s_1 from S_1 onto the line S_2 which is located at the angle 2α from S_1.
- Thus, for the projection result s_2, we will have
$$d_2 = d(s_2, \widetilde{s}) = d_1 \cdot \cos(2\alpha) = d_0 \cdot \cos(\alpha - \beta) \cdot \cos(2\alpha).$$
- After this, we may again project onto S_2, then again project onto S_1, etc.
- For each of these projections, the angle is equal to 2α.
- So, after each of them, the distance from the desired point $\widetilde{s}$ is multiplied the same factor $\cos(2\alpha)$.
- So, after k projection steps, we get a point s_k for which $d_k = d(s_k, \widetilde{s}) = d_0 \cdot \cos(\alpha - \beta) \cdot \cos^{k-1}(2\alpha)$.

15. The Simplest Case (cont-d)

- In the second option, we start with teaching S_2, i.e., if we first project the state s_0 into the set S_2.
- In this option, we get the following configuration:



16. The Simplest Case (cont-d)

- Here, we have

$$d_1 = d_0 \cdot \cos(\alpha + \beta),$$

$$d_2 = d(s_2, \widetilde{s}) = d_1 \cdot \cos(2\alpha) = d_0 \cdot \cos(\alpha + \beta) \cdot \cos(2\alpha), \ldots$$

$$d_k = d(s_k, \widetilde{s}) = d_0 \cdot \cos(\alpha + \beta) \cdot \cos^{k-1}(2\alpha).$$

- In general, $\cos(\alpha - \beta) \neq \cos(\alpha + \beta)$.
- So, a change in the presentation order can indeed drastically change the success of the learning procedure.

17. Analysis

- Starting with S_1 leads to a more effective learning than starting with S_2 if and only if

$$d_0 \cdot \cos(\alpha - \beta) \cdot \cos^{k-1}(2\alpha) < d_0 \cdot \cos(\alpha + \beta) \cdot \cos^{k-1}(2\alpha).$$

- This is equivalent to $\cos(\alpha - \beta) < \cos(\alpha + \beta)$.
- For the angles $x \in [0, \pi]$, the cosine $\cos(x)$ is a decreasing function.
- So, projection of S_1 is better if and only if $\alpha - \beta > \alpha + \beta$.
- Thus, we arrive at the following recommendation.

18. Recommendation

- We want to make learning more efficient.
- So, we should start with studying the material which is further away from the current state of knowledge.
- In other words, we should start with a material that we know the least.
- This recommendation explains why studying more difficult (abstract) ideas first enhances learning.
- It also ties in nicely with a natural commonsense recommendation that:
  - to perfect oneself,
  - one should concentrate on one's deficiencies.

19. Another Conclusion: an Explanation Why Early Start Can Inhibit Learning

- The age at which we teach different topics changes:
  - if it turns out that students do not learn, say, reading by the time they should,
  - a natural idea is to start teaching them earlier.
- Several decades ago, reading and writing started in the first grade.
- Now they start at kindergarten and even earlier.
- At first glance, the earlier we start, the better the students will learn.
- In practice, however, early start often inhibits learning.
- Our result provides a geometric explanation of this phenomenon.

20. Early Start (cont-d)

- Early start means a concentration on teaching skills that students can easily learn.
- According to our geometric model, this is detrimental in the long run.
- It is better to start with more challenging skills, skills that will not be acquired so easily.
- This will not lead to immediate spectacular results, but will eventually lead to better success.

21. Another Example: In What Order to Present the Material

- To make sure students learn, we need to repeat each of the learning items.
- 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?
- 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.
- The distance between the points can be viewed as a measure of similarity between the items.

22. In What Order (cont-d)

- In these terms, a presentation order is described by a curve $x(t)$ in the multi-D space.
- The experience of learning shows that often, presenting the items in random order is beneficial.
- To allow for this possibility, 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.
- Such 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.

23. In What Order (cont-d)

- 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
  - we start with some later point t_0.
- Hence, the characteristics of the process should be the same: $m(t) = m(t + t_0)$ and $A(t, s) = A(t + t_0, s + t_0)$.
- We conclude that $m(t) = \text{const}$.

24. In What Order (cont-d)

- 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 $A(t, s) = A(t - s, 0)$.
- So, $A(t, s)$ depends only on the difference between the times: $A(t, s) = a(t - s)$, where $a(t) \stackrel{\text{def}}{=} A(t, 0)$.
- In other words, the random process must be *stationary*.
- What autocorrelation function $a(t)$ should we use?

25. We Must Choose a Family of Functions

- The function $a(t)$ depends on how intensely we train.
- In more intensive training, we present the material faster.
- 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: $a(t) \to C \cdot a(t)$.
- The functions $a(t)$ and $C \cdot a(t)$ describe exactly the same learning strategy, but with different intensities.
- So, we should choose the best *family* $\{C \cdot a(t)\}_C$ of autocorrelation functions.

26. Which Family Is the Best?

- To select a family, we need to have a criterion.
- In other words, for every two families F and G, we decide whether:
  - F is better ($G < F$),
  - G is better ($F < G$),
  - or they are of the same quality ($F \sim G$).
- If the criterion selects several optimal families, we can use this uncertainty to optimize something else.
- For example, if teaching results are the same, we can optimize student satisfaction.
- In effect, this means that we use a new criterion with fewer optimal families.
- The final criterion should have only one optimal family.

27. Which Family Is the Best (cont-d)

- The relative quality of families should not depend on the choice of time unit:

$$\{C \cdot a(t)\} < \{C \cdot b(t)\} \Rightarrow \{C \cdot a(\lambda \cdot t)\} < \{C \cdot b(\lambda \cdot t)\}.$$

- It turns out that for each such scale-invariant final criterion, the optimal family has $a(t) = t^\alpha$ for some α.
- So, optimal configuration is a fractal process.
- When $\alpha = 2$, we have a straightforward trajectory, without any randomness.
- $\alpha = 0$ means that $x(t)$ and $x(s)$ for $t \neq s$ are completely uncorrelated, i.e., we have a white noise.
- Our experience showed that such *fractal* order indeed leads to improvement in learning.
- The optimal α depends on the students.

28. Acknowledgments

This work was partly supported:

- by Kazan University, and
- by the US National Science Foundation grants:
  - 1623190 (Model of Change for Preparing a New Generation for Professional Practice in Comp. Sci.),
  - HRD-1242122 (Cyber-ShARE Center of Excellence).