XOR as a Minimal Model of Topological Learning

S. K. Kwon[1*] and H. N. Kwon[2]
[1]Department of Physics, Pohang University of Science and Technology, Pohang 37673, Republic of Korea
[2]Department of Economics and Division of Computer Science, Hanyang University, Ansan 15588, Korea
[*]Correspondence: sekk@postech.ac.kr

Abstract

The XOR problem has long been cited as proof that neural networks require nonlinear activation functions to learn nonlinearly separable decision boundaries.
We argue that this interpretation is reversed.
XOR is not evidence that nonlinearity is required; it is evidence that learning is fundamentally topological.
We demonstrate that XOR can be solved by a purely piecewise-linear model, provided that the input space is partitioned into local patches and those patches are reconnected by a learned transition rule.
No nonlinear activation is applied to the value branches.
The softmax gate acts solely as a transition rule defining the topological partition of the input domain.

In this view, learning is the act of cutting and gluing regions of input space—a process that induces a nontrivial holonomy similar to a monopole-like singularity.
Because topological quantities are invariant under continuous deformation, all solutions of XOR—whether through ReLU, sigmoid, tanh, high-dimensional lifting, or discrete gating—are merely different coordinate representations of the same topological object.
Continuity, differentiability, and analytic activation functions are not the essence of learning; the topology of how input regions are divided and re-attached is.

1. Introduction

Since the 1980s, XOR has been regarded as the archetypal case that exposed the limitation of linear perceptrons[1].
The common explanation is that a single linear separator cannot classify XOR[2], so one must add a hidden layer and a nonlinear activation to "bend" decision boundaries[3].
This folklore assumes that *analytic nonlinearity* is the source of learning power.

We show instead that the failure of linear separability in XOR reveals something deeper: learning is not about curvature but about topology.
The model must carve the input space into linear regions, assign them to classes, and reconnect non-adjacent regions while keeping adjacent ones apart.
This process—cutting and gluing—is topological surgery on the input manifold.
Locally the system is linear; globally, the gluing rules create a nontrivial topology.

Continuity and differentiability, while convenient for optimization, are not ontological requirements.
Digital hardware—discrete, hysteretic, and non-differentiable—performs deep learning successfully at every moment in time and space.
Thus, the essential structure of learning cannot depend on smooth calculus.
The XOR problem, viewed correctly, demonstrates that what is learned is a *global topological relation*, not a local analytic nonlinearity.

2. Patch structure and holonomy

All XOR solutions, classical and modern, follow the same pattern: (*i*) divide the input plane into regions, (*ii*) assign binary labels to them, and (*iii*) define how the regions are reconnected.
In activation-based models (sigmoid, tanh, ELU), the network typically forms two negatively sloped boundaries that partition the square $[-1,1]^2$ into three macroscopic zones.
The middle strip acts as a transition region between classes.
The details differ—sigmoid boundaries are blurred, tanh and ELU boundaries are curved—but topologically, all yield the same configuration: two cuts, three regions, one transition band.

Our activation-free model makes this structure explicit.
We define the domain as $[-1,1]^2$ to expose the input symmetry under sign flips.
The model introduces two orthogonal boundaries, $x_1 = 0$ and $x_2 = 0$, producing four quadrants: $(+,+),(+,-),(-,+),(-,-)$.
XOR classification is realized by *gluing* opposite quadrants together: $(+,+)$ with $(-,-)$, and $(+,-)$ with $(-,+)$.
Adjacent quadrants remain separated; diagonal quadrants are identified.

This diagonal identification cannot be achieved by any single linear separator.
It is a global reattachment, equivalent to a surgery on the plane.
The transition between patches can be represented as a discrete internal variable—a *spin*.
Crossing a boundary flips the spin, and traversing a closed loop around the origin accumulates a mismatch.
This mismatch is a *holonomy*[4,5].
In gauge-theoretic language, it signals a monopole[6]: the XOR origin behaves as a topological singularity whose surrounding loop carries quantized phase.

The observed "nonlinearity" is therefore the manifestation of this holonomy, not analytic curvature.
What has been treated as functional nonlinearity is actually topological charge[7].

3. Universality across activation functions

Different activation functions merely dress the same topological core.
Sigmoid and ELU produce slanted boundaries; tanh and ReLU bend them; our gating model uses sharp, axis-aligned cuts. Yet all of them

1. partition input space into linear patches[8],
2. map patches into two logical classes, and
3. create a nontrivial loop around the origin where labels cannot be made globally single-valued.

Topology is blind to geometry. Angle, curvature, and smoothness are irrelevant.
The invariant is the connectivity pattern—the way space is cut and reattached.
That is why XOR can also be solved by "dimension lifting" or by increasing network width: these operations only deform the embedding, not the underlying topology[9].
The XOR solution remains topologically stable, confirming that learning captures invariants of connection, not of shape.

4. Gating models and experimental verification

We implemented two activation-free variants of XOR learning: *soft gating* and *hard gating*.

In **soft gating**, each local linear branch contributes smoothly; boundaries are already well defined, but slightly softened near the origin.
In **hard gating**, training is identical but inference uses a discrete winner-take-all rule, yielding perfectly sharp, axis-aligned quadrants and explicit diagonal identification.

These results were compared to classical activation-based networks.
The classical 2–4–1 MLP is trained on the standard XOR dataset consisting of the four corner points of $[-1,1]^2$: $\{(-1,-1),(-1,1),(1,-1),(1,1)\}$.
Panels (a)–(d) in Fig. 1 show sigmoid-, tanh-, ReLU-, ELU-based models; panels (e) and (f) show our soft- and hard-gated models on the same domain.
Despite superficial differences in curvature and softness, all the panels depict the same topological surgery on the plane.

Our proposed **2–2g–2h–1 model**—two inputs, two gated branches, two hidden linear units, and one output—achieves *exact XOR classification* with only **12 trainable parameters**, approximately **29 % fewer** than the conventional **2–4–1 multilayer perceptron** (17 parameters).
Despite this reduction, it reproduces the full decision structure of the nonlinear models, demonstrating that the topological mechanism alone suffices for perfect learning.
The hard-gated model achieves 100% accuracy on the four XOR corner points.

5. Consequences and outlook

The input plane of XOR begins with full linear symmetry under sign flips and quadrant permutations.
Learning breaks this symmetry by pairing opposite quadrants while keeping adjacent ones distinct.
What we perceive as "nonlinearity" is in fact *spontaneous symmetry breaking in the topology* of the input domain.

The resulting holonomy behaves as a conserved topological charge, directly analogous to magnetic flux quantization in gauge theory.
The singularity at the origin is mathematically analogous to the field of a monopole—it is the minimal topological defect required to make learning possible.

Smooth activations and high-dimensional embeddings merely parameterize this invariant in different coordinate systems.
The true object of learning is the topological connection among patches.
A trained network, even a deep one, can be viewed as a patch atlas on input space[10] equipped with discrete internal spin states that define how patches are identified.
Transporting this spin around singular points produces holonomy, and holonomy is the measurable witness that learning has occurred.

XOR is the minimal nontrivial manifold on which learning becomes possible.
It reveals that the essence of learning lies not in the algebra of activation functions but in the topology of connection.

The specific choice of gating function is not essential.
Gating is merely a device that partitions the input domain into patches; any rule—softmax, argmax, thresholding, or geometric partitioning—produces the same topological structure.
Thus, the success of the model stems not from the gating formula itself but from the topology it induces.
Learning without nonlinearity exposes what learning with nonlinearity hides behind it — the topological skeleton underpinning intelligence.

Figure 1. ***Comparative visualization of decision structures.***

Panels (a–d): conventional models using sigmoid, tanh, ReLU, ELU activations trained on the four corner points of $[-1,1]^2$: $\{(-1,-1),(-1,1),(1,-1),(1,1)\}$.
Two slanted or curved boundaries divide the space into three regions with a transition band.
Panels (e-f): the soft-gated model on $[-1,1]^2$ yields a smooth nonlinear decision surface due to the softmax transition, whereas the hard-gated model exhibits sharp linear boundaries corresponding to the learned partitions on $[-1,1]^2$.
(e) Soft gating: smooth transitions near the origin.
(f) Hard gating: perfectly sharp quadrant boundaries and diagonal pairing.
All the panels represent the same topological transformation—cutting and gluing of input space—revealing that apparent nonlinearity is a manifestation of underlying holonomy.

References

1. Rosenblatt, F. *The Perceptron: A probabilistic model for information storage and organization in the brain.* Psychological Review 65, 386–408 (1958).
2. Minsky, M. & Papert, S. *Perceptrons: An Introduction to Computational Geometry.* MIT Press (1969).
3. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. *Learning representations by back-propagating errors.* Nature 323, 533–536 (1986).
4. Berry, M. V. *Quantal phase factors accompanying adiabatic changes.* Proc. R. Soc. Lond. A 392, 45–57 (1984).
5. Simon, B. *Holonomy, the quantization condition, and quantum Hall effect.* Phys. Rev. Lett. 51, 2167–2170 (1983).
6. Dirac, P. A. M. *Quantised singularities in the electromagnetic field.* Proc. R. Soc. Lond. A 133, 60–72 (1931).
7. Nakahara, M. *Geometry, Topology and Physics.* 2nd ed., CRC Press (2003).
8. Montúfar, G. F., Pascanu, R., Cho, K. & Bengio, Y. *On the number of linear regions of deep neural networks.* NIPS 27, 2924–2932 (2014).
9. Poole, B. et al. *Exponential expressivity in deep neural networks through transient chaos.* NeurIPS 29, 3360–3368 (2016).
10. Gabrielsson, R. & Carlsson, G. *A topological view of deep learning.* Topological Data Analysis in Machine Learning, Springer (2020).

Declarations

Competing interests

The authors declare no competing interests.

Data availability

The code used to generate the results and figures in this paper is available on GitHub at: https://github.com/RubiksCube33/XOR_Topological_Learning