Explainable artificial intelligence and mathematics: What lies behind? Let us focus on this new research field

Massimiliano Ferrara

The growing complexity inherent in modern artificial intelligence (AI) models has necessitated an increased focus on the demand for explainability, commonly referred to as explainable artificial intelligence *(XAI). The primary objective of XAI is to render the decision-making processes of AI systems not only transparent, but also understandable to human users, thereby fostering greater trust and comprehension among stakeholders. As AI systems become more sophisticated and are deployed in critical areas such as healthcare, finance, and autonomous vehicles, the demand for clarity surrounding their operations intensifies.*

This paper delves deeply into the vital relationship between XAI and mathematics, asserting that mathematical principles are foundational to enhancing the interpretability, transparency, and overall trustworthiness of AI models. We will investigate the key mathematical constructs that underlie various XAI techniques, providing insights into how they function and contribute to explainability.

To illustrate the practical significance of these principles, we will examine specific case studies where mathematical frameworks have successfully improved the elucidation of AI model predictions. Furthermore, this paper will outline potential future avenues for research that aim to further integrate mathematical methodologies within XAI frameworks. By doing so, we hope to contribute to the development of more robust and interpretable AI systems that can be trusted and effectively utilized by humans in a multitude of applications.

1 Introduction

In recent years, artificial intelligence (AI) has undergone revolutionary advancements, leading to its integration into numerous critical applications spanning industries such as healthcare, finance, transportation, and beyond. With its capabilities to analyze vast datasets, recognize patterns, and make informed decisions, AI systems have become invaluable in promoting efficiency and innovation. However, as AI technologies evolve, so too has the complexity of the models driving their decision-making processes.

Simultaneously, this sophistication raises significant concerns regarding transparency and interpretability issues encapsulated in the term *black box* [7]. Many modern AI algorithms, particularly those based on deep learning, operate in ways that are not easily understandable by humans, rendering their decision-making processes opaque. This lack of clarity poses serious risks, especially in high-stakes environments. For instance, in healthcare, algorithmic decisions can influence clinical diagnoses and treatment plans, where a misinterpretation or erroneous model output could have drastic consequences on patient outcomes. In finance, automated systems determining creditworthiness must comply with regulations requiring transparency. When applicants are denied loans, they must be provided with understandable explanations.

To address these concerns, the field of explainable artificial intelligence (XAI) has emerged as a critical research area. XAI aims to develop methodologies and frameworks that enable humans to comprehend, trust, and exploit AI systems effectively. More than simply improving model interpretability, XAI encompasses a proactive approach to ensuring accountability and ethical standards in AI deployment. Increasingly, stakeholders are demanding that AI not only be high-performing, but also accessible and explainable to users and regulators alike.

This paper delves into the vital interplay between mathematics and XAI, asserting that a robust understanding of mathematical foundations is essential to fostering clearer interpretations of AI model behavior. Mathematics provides the frameworks necessary for developing advanced interpretability techniques, ranging from Shapley values to feature importance scores. The synthesis of these mathematical tools with AI narrows the gap between model complexity and user understanding, ultimately driving innovation in more reliable AI systems.

Through this exploration, the paper aims to outline the critical connections between mathematics and XAI methodologies while providing concrete case studies that illustrate these principles in action. By addressing the landscape of XAI in conjunction with its mathematical backbone, we aim to promote a comprehensive understanding of how these two domains can and should intersect.

2 Motivation

To appreciate the significance of explainability in artificial intelligence, it is essential to understand the historical context of AI development and the challenges that have accompanied the rise of complex algorithms. The roots of AI can be traced back to the mid-20th century, with early endeavors focused on rule-based systems that simulated basic reasoning capabilities. These systems relied primarily on human-crafted instructions and logic, making them relatively interpretable. However, as computational capacity burgeoned alongside data availability, the emergence of machine learning algorithms marked a paradigm shift, enabling systems to learn from data rather than rely solely on predefined rules [6].

Machine learning models, particularly those utilizing neural networks and deep learning architectures, have since demonstrated unparalleled performance in tasks such as image recognition, natural language processing, and game playing. However, this success has come at the cost of interpretability. As these models grow in complexity, comprising multiple hidden layers, millions of parameters, and intricate interactions, their internal workings become increasingly opaque. Users cannot easily ascertain how inputs are transformed into outputs, resulting in a sense of discomfort and mistrust, particularly in critical applications.

This "black box" nature of AI has prompted a renewed focus on explainability over the last decade. Researchers and practitioners recognize that building public trust in AI systems requires elucidating how and why decisions are made. Moreover, explainable AI is not merely a technical challenge, but also a societal imperative. Ethical implications abound when algorithms govern fundamental aspects of human lives, such as health and financial stability.

In this context, key concepts within XAI have emerged, defining a spectrum of approaches and frameworks designed to enhance interpretability. These include model-agnostic methods, which offer insights applicable across various algorithms, and instance-based explanations, which delve into the specifics of individual predictions. Researchers have utilized methods from diverse fields, including statistics, game theory, and information theory, to craft explanations that resonate with end-users. The importance of explainability is underscored by industry efforts and regulatory requirements, highlighting the critical need for interpretable models to ensure ethical practices [4, 9].

Simultaneously, this new domain raises discussions about the mathematical foundations of XAI techniques. Understanding the underlying mathematics is crucial for developing robust explanations that carry both technical accuracy and meaningful human insights. By integrating mathematical reasoning into XAI, we can promote the development of systems that not only function well, but also provide clear and actionable explanations for their behavior. As we transition into exploring the mathematical foundations of AI and XAI methodologies in the following sections, it becomes evident that the innovative approaches we seek will rely heavily on our understanding of the mathematical concepts that underpin this technology. A deeper engagement with these connections will pave the way for enhanced trust, usability, and societal acceptance of artificial intelligence.

3 Mathematical foundations of XAI

Mathematics provides the bedrock upon which many XAI methods are built. From linear algebra and calculus to more complex fields like information theory and topology, mathematical concepts facilitate the extraction of meaningful information from AI models.

3.1 Linear algebra and matrix decompositions

Linear algebra is fundamental in model interpretation, particularly in techniques like principal component analysis (PCA) and singular value decomposition (SVD). These methods reduce data dimensionality while preserving variance, making it easier to visualize and interpret high-dimensional data.

Principal component analysis (PCA)

PCA transforms data by projecting it onto orthogonal vectors that maximize variance. The transformation of a dataset X using PCA involves computing its covariance matrix Σ, and then deriving its eigenvalues and eigenvectors. The principal components are the eigenvectors corresponding to the largest eigenvalues:

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^{\mathsf{T}},$$

$$\Sigma v = \lambda v.$$

Here, v represents the eigenvectors (principal components), and λ the eigenvalues.

Singular value decomposition (SVD)

SVD generalizes PCA and decomposes a matrix into singular vectors and singular values. For a given matrix A, SVD can be represented as

$$A = U\Sigma V^{\mathsf{T}},$$

where U and V are orthogonal matrices, and Σ is a diagonal matrix of singular values.

3.2 Calculus and optimization

Gradient-based optimization techniques, derived from calculus, are essential for training AI models. Understanding gradients and Hessian matrices helps in explaining how models learn from data, and in identifying critical features and decision boundaries.

Gradient descent

Gradient descent minimizes a function $f(\theta)$ by iteratively moving in the direction of the steepest descent, defined by the negative gradient. The update rule is given by

$$\theta_{t+1} = \theta_t - \eta \nabla f(\theta_t),$$

where η is the learning rate, and $\nabla f(\theta_t)$ is the gradient of the function at θ_t.

Hessian matrices and curvature

The Hessian matrix H of a function $f(\theta)$ at point θ is a square matrix of second-order partial derivatives, representing the local curvature:

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial \theta_1^2} & \cdots & \frac{\partial^2 f}{\partial \theta_1 \partial \theta_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial \theta_n \partial \theta_1} & \cdots & \frac{\partial^2 f}{\partial \theta_n^2} \end{bmatrix}.$$

3.3 Information theory

Information theory quantifies uncertainty and information gain, aiding in the development of metrics such as entropy and mutual information. These metrics are vital for feature selection and model interpretability.

Entropy and Information Gain

Entropy $H(X)$ measures the uncertainty in a random variable X:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i).$$

Information gain measures the reduction in entropy when a dataset is split based on an attribute:

$$\mathrm{IG}(Y \mid X) = H(Y) - H(Y \mid X).$$

3.4 Mutual information

Mutual information $I(X;Y)$ quantifies the amount of information obtained about one random variable through another:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}.$$

4 The contribution of game theory: New perspectives

Game theory is a branch of mathematics that investigates the strategic interactions among rational agents and explores their wide-ranging applications across diverse fields, including artificial intelligence (AI). Within the domain of explainable artificial intelligence (XAI), game theory offers a foundational methodology for enhancing our understanding and improving the transparency of AI models.

A pivotal concept in game theory is the representation of strategic interactions as "games," where participants engage in rational decision-making to optimize their objectives. By applying these principles to AI explainability, we can regard the decision-making processes of AI models as a game involving the artificial system and human users attempting to comprehend its actions.

Game theory furnishes a conceptual framework for examining the strategies deployed by AI models to convey their decisions in a clear and comprehensible manner. For instance, utilizing concepts such as Nash equilibrium allows us to analyze how AI models and human users can collaborate effectively to facilitate meaningful explanations of the system's decisions.

Furthermore, game theory can assist in modeling situations where the explainability of AI may conflict with other objectives, such as computational efficiency or predictive accuracy. By evaluating multi-agent games and identifying strategic trade-offs, we can devise strategies that reconcile these competing considerations and create explainable AI frameworks that satisfy a variety of requirements.

In conclusion, integrating game theory into the XAI realm can offer novel insights and methodologies for addressing challenges related to the transparency and interpretability of artificial systems. By leveraging fundamental concepts from game theory to analyze and optimize the interactions between AI models and human users, we can foster the development of intelligent systems that are not only powerful and accurate, but also comprehensible and acceptable to society.

Shapley values and their role

Shapley values, which originate from cooperative game theory, guarantee a fair distribution of payoffs among participants [10]. In the context of XAI, Shapley values quantify the contribution of each feature to the overall prediction. The Shapley value for a feature i is defined as

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)],$$

where N represents the complete set of features, and $v(S)$ is the value function that denotes the prediction when the subset S of features is utilized.

Application in SHAP

Shapley additive explanations (SHAP) apply Shapley values to provide consistent and verifiable feature attributions. Delve into the mathematical formulation of SHAP using the Shapley value equation above and demonstrate with an example.

5 Case studies

To illustrate the synergy between mathematics and XAI, we consider several case studies where mathematical techniques have enhanced explainability.

LIME and SHAP
Local interpretable model-agnostic explanations (LIME) and Shapley additive explanations (SHAP) are popular XAI methods that rely on mathematical principles. LIME uses locally weighted linear regression to approximate a model's behavior around a specific prediction, while SHAP leverages cooperative game theory to distribute contributions of features fairly.

LIME
Detail the mathematical methodology behind LIME, including the optimization of local surrogates and interpretability of linear approximations. Provide a detailed example showcasing a step-by-step application of LIME to a specific prediction instance.

SHAP
Discuss SHAP's foundation in Shapley values from cooperative game theory. Highlight the mathematical derivation of Shapley values and their contribution to fair attribution of feature importance. Include a case study that rigorously applies SHAP to a real-world dataset, illustrating how feature contributions are computed and interpreted.

Decision trees and rule extraction
Decision trees, inherently interpretable models, use recursive partitioning based on feature values to generate easily understandable rules. Techniques like decision tree surrogate models create interpretable approximations of complex models.

Recursive partitioning
Explain the mathematical basis of recursive partitioning, including impurity measures like Gini impurity and entropy in the context of decision trees. Provide a case study that demonstrates the construction of a decision tree and the derivation of decision rules from the model [8]:

$$\text{Gini}(S) = 1 - \sum_{i=1}^{n} (p_i)^2.$$

Rule extraction methods
Detail methods for extracting rules from black-box models, such as model distillation and surrogate decision trees, with mathematical explanations of each approach. Include examples of rule extraction processes, illustrating the transformation of complex model outputs into human-understandable rules.

Bayesian networks
Bayesian networks utilize probability theory to represent and reason about the dependencies among variables. These networks simplify the visualization and understanding of probabilistic relationships, aiding in the interpretability of predictions.

Probabilistic graphical models
Discuss the mathematical foundation of Bayesian networks, including concepts of conditional independence and factorization of joint distributions. Provide an example application of Bayesian networks in a specific domain, highlighting how probabilistic dependencies are modeled and interpreted:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \text{Parents}(X_i)).$$

6 The role of mathematics in future XAI developments

As the field of artificial intelligence continues to evolve, the integration of advanced mathematical techniques into explainable artificial intelligence (XAI) is becoming increasingly critical. This intersection not only enhances model interpretability, but also opens new avenues for research and application, contributing to the overall trustworthiness of AI systems. The role of mathematics in future XAI developments can be categorized into several key areas: the exploration of advanced modeling techniques, the establishment of quantitative metrics for explainability, the application of optimization methods, and the potential contributions from emerging fields such as topological data analysis and information theory [3].

6.1 Advanced modeling techniques

Traditional machine learning algorithms have relied on wellestablished mathematical frameworks, such as linear regression and decision trees. However, with the rise of deep learning and other complex models, researchers are exploring innovative mathematical representations that enhance explainability. For instance, neural networks can be enhanced by integrating concepts from calculus, specifically through techniques such as those mentioned below.

Gradient-based explanation methods
The gradients of the loss function concerning input features are paramount for understanding model behavior. The backpropagation algorithm, expressed mathematically as

$$\delta^l = \nabla_a C \odot \sigma'(z^l),$$

is vital for calculating error derivatives across hidden layers, where δ^l represents the error term, C is the cost function, a is the activation output, σ is the activation function, and z is the weighted

input. Through gradient calculations, we can gain insights into which features influence the most the model's predictions.

Interpretability via attention mechanisms
Attention mechanisms in neural architectures, particularly transformer models, allow the model to focus on specific parts of the input sequence. Mathematically, the attention score can be defined as

$$\text{Attention}(Q, K, V) = \text{softmax}\Big(\frac{QK^{\mathsf{T}}}{\sqrt{d_k}}\Big)V,$$

where Q (queries), K (keys), and V (values) are derived from the input representations. Understanding the attention weights can help determine the importance of various input components in the model's predictions, making it easier to devise explanations that correspond to the input features responsible for specific outputs.

6.2 Quantitative metrics for explainability

To evaluate and compare the effectiveness of different explanation methods, it is imperative to establish rigorous quantitative metrics. Mathematics plays a crucial role in developing these metrics, which can quantify various aspects such as the following.

Fidelity and consistency
The fidelity of an explanation refers to how accurately it reflects the behavior of the underlying model. One way to mathematically validate this is through measures based on approximating the original model f with an interpretable model g.

Simplicity and completeness
Explainability metrics often emphasize the trade-off between complexity and comprehensiveness. For instance, a metric S to evaluate the simplicity of an explanation might be defined as

$$S(g) = \frac{1}{|g|}\sum_{i=1}^{|g|}\text{Length}(g_i),$$

where g_i are the components of the explanation. Here, a lower score indicates that an explanation is simpler, which is typically desirable.

6.3 Optimization methods

Mathematics is fundamental in optimizing models for both performance and explainability. Model interpretability often requires trade-offs that can be addressed through optimization techniques, such as those mentioned below.

Multi-objective optimization
This paradigm allows the simultaneous optimization of multiple conflicting objectives, for instance, maximizing model accuracy while minimizing complexity. An example objective function is provided by

$$\text{minimize} \quad \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda \cdot \text{Complexity}(f),$$

where L is the loss function measuring the error, y_i is the true output, x_i is the input data, and λ is a trade-off parameter for model complexity.

Regularization techniques
Regularization techniques, such as L1 (lasso) and L2 (ridge) regularization, help prevent overfitting while enhancing interpretability by encouraging sparsity in the model weights. The L1 regularization term can be mathematically expressed as

$$R(\theta) = \lambda \sum_{j=1}^{p} |\theta_j|,$$

where θ_j are the model parameters and p is the number of features. Sparse solutions lead to simpler models that are easier to interpret.

6.4 Contributions from emerging fields

As AI progresses, emerging mathematical fields are beginning to influence how we understand and develop explainable models, as in the examples below.

Topological data analysis (TDA)
TDA focuses on the shape of data and has been proposed as an avenue to reveal insights into high-dimensional datasets that may inform model behavior. Techniques such as persistent homology can provide a geometric understanding of the data manifold, potentially revealing relationships that enhance interpretability.

Information theory
Applying concepts from information theory allows researchers to quantify the information gains achieved through XAI methods. Measures such as mutual information can be leveraged to determine how much information an explanation conveys about the prediction

$$I(X; Y) = H(X) + H(Y) - H(X, Y),$$

where H represents entropy. Understanding the mutual information between input features X and predictions Y can guide the development of more informative explanations.

7 Future directions

The integration of advanced mathematical techniques into XAI is an ongoing field of research [1, 2, 5]. Future work may involve the following.

7.1 Topological data analysis (TDA)

TDA applies concepts from algebraic topology to uncover the shape and structure of data. Persistent homology, a key tool in TDA, can reveal robust features that contribute to model explanations.

Persistent homology

Explain persistent homology's mathematical foundation and its utility in identifying significant data features that persist across multiple scales. Include examples of how TDA has been applied to complex datasets and the insights it has provided.

Causal inference

Mathematical techniques from causal inference can help distinguish causation from correlation in AI models, providing deeper insights into the underlying mechanisms driving predictions.

Causal models

Introduce causal models and the mathematical formulation of causal relationships (e.g., do-calculus). Discuss applications in interpreting model decisions, providing examples of causal inference techniques applied to real-world AI predictions.

Information geometry

Information geometry examines the differential-geometric structure of statistical models. This perspective can enhance our understanding of model parameter spaces and improve interpretability.

Geometric understanding of models

Explain the mathematical principles of information geometry, including divergence measures and their role in interpreting statistical models. Provide examples of how information geometry can be applied to examine and understand deep learning models.

8 Conclusions

The integration of advanced mathematical techniques into future XAI developments is not merely a theoretical exercise, but a practical necessity. The mathematical tools underpinning AI systems will continue to shape the evolution of methods designed to promote transparency and interpretability. As we navigate a landscape of increasingly intricate models, the role of mathematics will persist as a cornerstone in the quest to unravel the complexities of AI, ensuring these systems serve humanity while adhering to standards of trust and accountability.

References

[1] C. M. Bishop, *Pattern recognition and machine learning*. Inf. Sci. Stat., Springer, New York (2006)

[2] H. Edelsbrunner and J. L. Harer, *Computational topology*. An Introduction. American Mathematical Society, Providence, RI (2010)

[3] I. Goodfellow, Y. Bengio and A. Courville, *Deep learning*. Adapt. Comput. Mach. Learn., MIT Press, Cambridge, MA (2016)

[4] T. Hastie, R. Tibshirani and J. Friedman, *The elements of statistical learning* (2nd edition) Springer Ser. Statist., Springer, New York (2009)

[5] J. A. Lee and M. Verleysen, *Nonlinear dimensionality reduction*. Inf. Sci. Stat., Springer, New York (2007)

[6] S. M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems,* pp. 4768–4777, Association for Computing Machinery, New York (2017)

[7] C. Molnar, *Interpretable machine learning: A guide for making black box models explainable.* Lulu.com (2019)

[8] J. Pearl, *Causality*. Models, Reasoning, and Inference. (2nd edition) Cambridge University Press, Cambridge (2009)

[9] M. T. Ribeiro, S. Singh and C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier. In *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 1135–1144, Association for Computing Machinery, New York (2016)

[10] L. S. Shapley, A value for *n*-person games. In *Contributions to the theory of games,* vol. 2, pp. 307–317, Ann. of Math. Stud 28, Princeton University Press, Princeton, NJ (1953)

Massimiliano Ferrara is a full professor of mathematical economics, business analytics, decision support systems, and economic intelligence at the Mediterranean University of Reggio Calabria, and a research affiliate at ICRIOS – the Invernizzi Center for Research on Innovation, Organization, Strategy and Entrepreneurship, Bocconi University, Milan. He served as a member of the AMASES Scientific Committee (2017–2022), and is a council full member delegate at the European Mathematical Society (EMS) since 2020. He is an editor, co-editor, and associate editor of various international scientific journals, notably: *Nature – Scientific Reports* (member of the editorial committee), *Soft Computing, Dynamic Games and Applications,* and *Journal of Dynamics and Games* (associate editor). He has served as a referee for over 250 international scientific journals on economics, pure and applied mathematics, indexed by SCOPUS, WoS, and MathSciNet. His research interests include mathematical economics, game theory, explainable artificial intelligence (XAI), machine and deep learning modeling, business analytics, optimization theory, nonlinear analysis, dynamical systems, and epidemic modeling.

massimiliano.ferrara@unirc.it