International Journal of Scientific Research and Engineering Development-– Volume 6 Issue 5,Sep-Oct 2023

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

MACHINE LEARNING AND ITS IMPLICATIONS

SiddihaRimzan, Sarah Behnaz ,Shamila, Fernando sis ,lora Helen, Carol Mathew and lebradahonsin.
Students of Atamie international school

Abstract:

In this article, we would deliver about Normal distribution in real and defines the how variances could be used for multiple events linear algebra, probability, calculus, and statistics—are the foundation of machine learning. Calculus aids in the learning and optimization of models, even if statistical ideas form the foundation of all models. When working with large datasets, linear algebra becomes quite useful, and probability aids in forecasting the course of future events. In your job in data science and machine learning, you will come across these mathematical concepts rather often.

In the chapter on probability, we saw that the binomial distribution could be used to solve problems such as "If a fair coin is flipped 100 times, what is the probability of getting 60 or more heads?" The probability of exactly x heads out of N flips is computed using many methods.

$$P(x) = \frac{N!}{x!(N-x)!}\pi^x(1-\pi)^{N-x}$$

where N is the number of flips (100), π is the chance of a head (0.5), and x is the number of heads (60). Consequently, you must calculate the likelihood of 60 heads, 61 heads, 62 heads, etc., and total up all of these probabilities in order to solve this problem. Consider how much time it must have taken to calculate binomial probabilities in the days before computers and calculators.

Abraham de Moivre, a statistician and gambling advisor in the eighteenth century, was frequently asked to perform these laborious calculations. De Moivre observed that the form of the binomial distribution approached a fairly smooth curve as the number of occurrences (coin flips) increased.

The distributions of many natural events are at least roughly normally distributed, which is the main reason for the significance of the normal curve. Analyzing measurement mistakes in astronomical observations—errors brought on by unreliable equipment and unreliable observers—was one of the early uses of the normal distribution.

In the 17th century, Galileo observed that minor errors were more common than large errors and that these faults were symmetric. This resulted in a number of proposed error distributions, but it wasn't until the early 1800s that it was determined that these errors had a normal distribution. The normal distribution formula was independently derived by mathematicians Adrain in 1808 and Gauss in 1809, who both demonstrated that errors.

2.Hypergeometric Distribution

Probabilities are computed using the hypergeometric distribution while sampling without replacement. Let's take an example where you take one card at random from a 52-card deck. Next, you sample a second card without changing it in the deck, and then a third (again without replacing cards). What is the likelihood that, out of the 52 cards in the deck, exactly two of the sampled cards will be aces given this sampling procedure? The following formula, which is based on the hypergeometric distribution, can be used to determine this probability:

$$p = \frac{({}_kC_x)({}_{(N-k)}C_{(n-x)})}{{}_NC_n} \quad \text{where}$$

k is the number of "successes" in the population
x is the number of "successes" in the sample
N is the size of the population
n is the number sampled
p is the probability of obtaining exactly x successes
${}_kC_x$ is the number of combinations of k things taken x at a time

In this example, k = 4 because there are four aces in the deck, x = 2 because the problem asks about the probability of getting two aces, N = 52 because there are 52 cards in a deck, and n = 3 because 3 cards were sampled. Therefore,

$$p = \frac{({}_4C_2)({}_{(52-4)}C_{(3-2)})}{{}_{52}C_3} =$$

3. Machine learning in Normal distribution

In Machine Learning, data satisfying Normal Distribution is beneficial for model building. It makes math easier. Models like LDA, Gaussian Naive Bayes, Logistic Regression, Linear Regression, etc., are explicitly calculated from the assumption that the distribution is a bivariate or multivariate normal. Also, *Sigmoid functions* work most naturally with normally distributed data.

Many natural phenomena in the world follow a log-normal distribution, such as *financial data* and *forecasting data*. By applying transformation techniques, we can convert the data into a normal distribution. Also, many processes follow normality, such as many *measurement errors in an experiment, the position of a particle that experiences diffusion, etc.*

So it's better to critically explore the data and check for the underlying distributions for each variable before going to fit the model.

Note: Normality is an assumption for the ML models. It is not mandatory that data should always follow normality. ML models work very well in the case of non-normally distributed data also. Models

like decision tree, XgBoost, *don't assume any normality and work on raw data as well. Also,* linear regression *is statistically effective if only the* model errors *are Gaussian, not exactly the entire dataset.*

Here I have analyzed the Boston Housing Price Dataset. I have explained the visualization techniques and the conversion techniques along with plots that can validate the normality of the distribution.

```
df = pd.read_csv("boston-dataset/boston_data.csv")
df.head(2)
```

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
0	0.15876	0.0	10.81	0.0	0.413	5.961	17.5	5.2873	4.0	305.0	19.2	376.94	9.88	21.7
1	0.10328	25.0	5.13	0.0	0.453	5.927	47.2	6.9320	8.0	284.0	19.7	396.90	9.22	19.6

A normal distribution has two primary parameters: the mean and the standard deviation. We may determine the distribution's shape and probability in relation to our problem statement with the use of these factors. The distribution's form varies as the parameter value does.

1 Mean

The average or mean value was employed by researchers to quantify central tendency. The distribution of variables that are measured as intervals or ratios can be described using it.

In a graph of a normal distribution, the majority of the data points are grouped around the mean, which establishes the location of the peak.

The normal distribution curve shifts to the left or right along the X-axis when the mean value is altered.

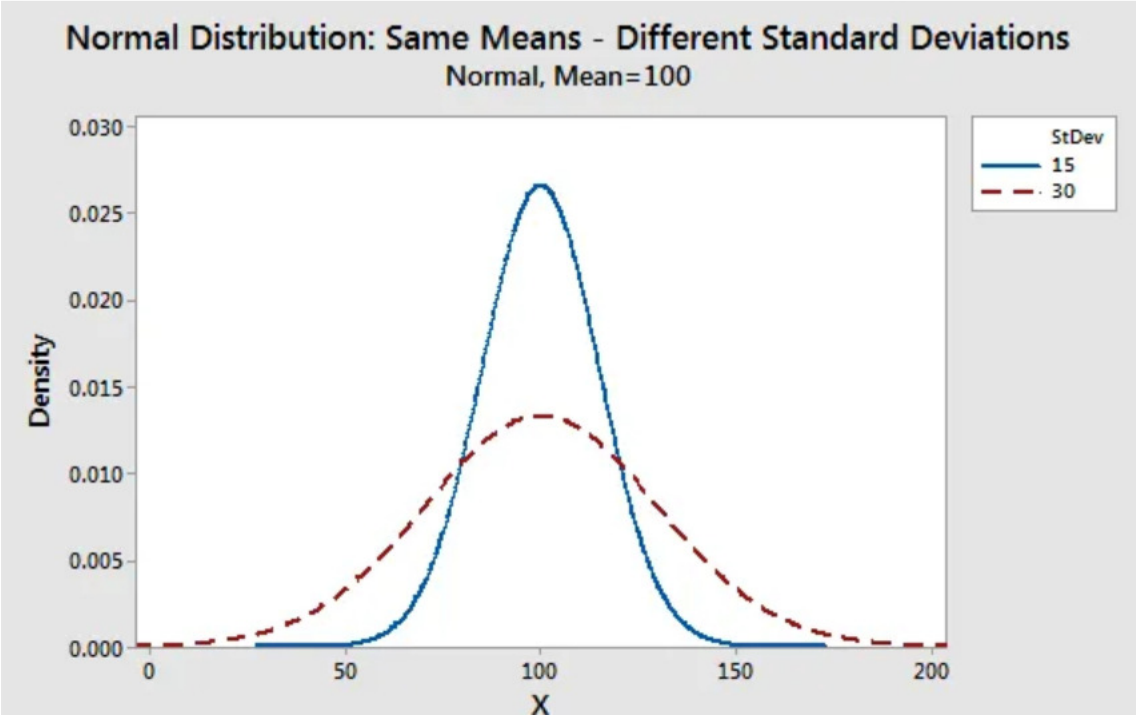
Mean Distortion

The standard deviation quantifies the degree of dispersion between the data points and the mean.

It reflects the distance between the mean and the data points and calculates the deviation of the data points from the mean.

The graph's width is determined by the standard deviation. Consequently, altering the standard deviation value causes the distribution's breadth along the x-axis to either contract or widen.

A steeper curve is often produced by a smaller standard deviation relative to the mean, and a flatter curve by a higher standard deviation.



The normal distribution has an exactly symmetrical shape.

This indicates that we can split the normal distribution curve in half and create two equal halves. Additionally, the symmetric shape is present when the number of observations on each side of the curve is equal.

2. The mode, median, and mean are all the same.

The point with maximum frequency, or the majority of the variable's observations, is referred to as the midpoint of a normal distribution.

All three of the central tendency measurements fall at the midway. Typically, a perfectly formed normal distribution has equal measurements.

4 Calculus

Deductive Reasoning Strategies 3 Although individuals seem to struggle with abstract logical tasks, it also appears that people are more successful when the tasks are posed in a familiar context. For example, in the authors found that education majors reasoned about contraposition better in a verbal, syllogistic environment than in a purely mathematical environment with abstract symbols and sentences. Wason and Shapiro also found that participants performed better on the Wason Selection Task when given a rule with familiar content, as opposed to an abstract rule.

For example, a rule such as “every time I go to Manchester, I travel by car” would be considered “familiar” whereas the rule “every card which has a D on one side has a 3 on the other side” would be considered “abstract.” It is important to note that although these studies do seem to suggest that familiar

contexts can help students to reason correctly, Stylianides et al. state that research, however, provides a weak basis on which to formulate hypotheses about the relation between students' performance in tasks with non-meaningful words and symbolic tasks that investigate the same logical principles and that research tends to favor logical reasoning in "meaningful verbal contexts."

For additional work regarding the teaching and learning of logical implication, see Yopp for a study related to eighth grade learning of the contrapositive and Attridge, et al. for undergraduate understanding of conditionals given previous logic experience. Also, see and recent work related to the Wason Selection Task. Findings from decades of research have provided insights into student thinking and the challenges that students encounter with calculus ideas such as limit, differentiation, and integration. For a history of this work, , and for more detailed reviews of the literature and findings specific to sub-topics in calculus,], and SIGMAA on RUME conference proceedings (http://sigmaa.maa.org/rume/).

The instruction students receive about key calculus ideas often includes theorem or theorem-like statements and students are expected to reason logically from them.

However, the vast majority of research into student understanding of this kind of logical reasoning has occurred in the context of introduction to proof or other proof-focused courses [11] and has not focused on students in introductory calculus. Although much work has been done separately on both the issues of logical implication and calculus learning, we know little about how students understand and work with ideas of logical implication that appear in theorems and theorem-like statements in a calculus context. Researchers have examined related ideas through studies of student thinking about sequences and series (nd some work has examined calculus students' meanings for quantifiers found in calculus theorems

. However, the focus in that work was specifically on quantifiers appearing in complex theorems. To date, beginning calculus student understanding of conditionals in the form of if-then statements that occur in introductory calculus has not 4 Case and Speer been closely examined. To explore this, we were interested in whether calculus students had the same kinds of difficulties with calculus-based conditional statement tasks as they did with the purely abstract tasks.

In other words, we wondered if calculus theorems provided enough of a "context" to support students' productive reasoning or whether those tasks were treated in the same way as the classic, abstract tasks.

To examine this, we focused on particular types of tasks set in abstract and calculus contexts that are characteristic of one kind of reasoning expected of students. In particular, in calculus, students are told to take for granted the truth of a particular theorem and then asked to draw conclusions given a true or false antecedent or consequent. It is important to note that we did not ask students to consider the truth or falseness of an entire conditional statement. Below are the tasks, in the abstract, that students were asked to consider:

• Inverse Task: Suppose that $p \Rightarrow q$ is true and you know that p is false. Is q true, false, or is it not possible to tell? Explain.

• Converse Task: Suppose that $p \Rightarrow q$ is true and you know that q is true. Is p true, false, or is it not possible to tell? Explain.

• Contrapositive Task: Suppose p ⇒ q is true and you know that q is false. Is p true, false, or is it not possible to tell? Explain.

• Modus-ponens Task: Suppose that p ⇒ q is true and you know that p is true. Is q true, false, or is it not possible to tell?

They are given that p ⇒ q for a particular p and q as well as a situation in which p (or q) is either true or false. An inference may then be made by reasoning with these two pieces of information. As an example of such a task, students may be told to assume that the following theorem is true: For all functions f, if f is differentiable at a point x = c, then f is also continuous at the point x = c. Then, given a particular function and point such as f (x) = x2 at x = 0, the student is then expected to investigate whether f (x) = x2 at the point x = 0 is differentiable. If the student determines that the antecedent is met, the student may then use the theorem to infer that f (x) = x2 is continuous at the point x = 0.

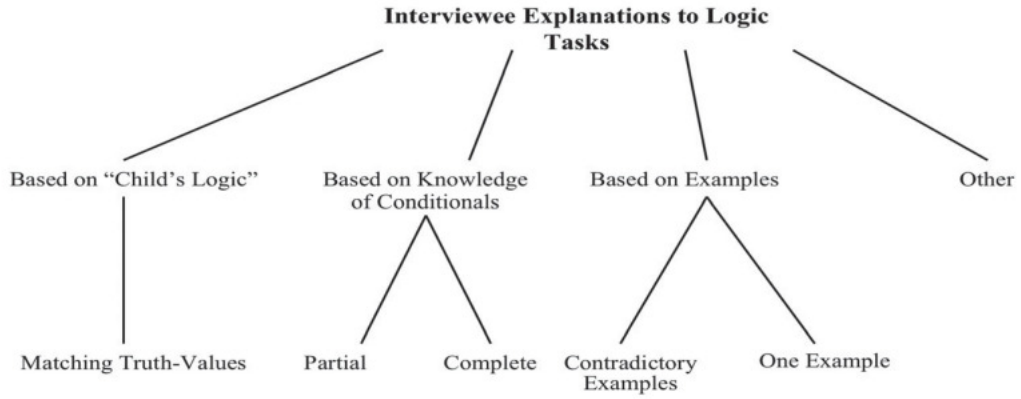
Given the importance of this kind of reasoning, we focus on student reasoning involving the inference of q (or p) given a true implication statement p ⇒ q and true or false p (or q).

Note that we are not concerned with students' abilities to validate the truth or falseness of an entire conditional statement given an antecedent and consequent. Rather, we are interested in how students infer the status of an antecedent or consequent given a true conditional statement and a true or false antecedent or consequent.

Although analyses of the survey data provided some insights (e.g., the calculus context seems to make some of the reasoning easier for students, the abstractly stated tasks are generally much more difficult for students, etc.),

we wanted to understand more about student thinking concerning the inferences, to gain further insight into the findings from the survey data analyses. From analysis of the interview data, we identified several different ways in which students approached the tasks. As displayed in Figure 3, there were three main ways of thinking (plus "other'), some of which had sub-categories that characterized the thinking at even finer levels of detail. The "other" category was used for

Deductive Reasoning Strategies



responses that were difficult to categorize and/or did not seem to fit the main categories.

Child's Logic and Knowledge of Conditonals

We first consider the strategies located on the two left-most branches in Figure 3. Interviewees who responded with "Child's Logic" (a common logical misunderstanding) tended to match truth-values (that is, they responded with "True" when given a true premise and responded with "False" when given a false premise). This strategy, when applied with complete consistency, generates correct answers to two of the four tasks. While interviewees sometimes exhibited "Child's Logic" on the calculus portion of the interview, they appeared more apt to consistently apply this kind of thinking on the abstract section.

For example, 5 of the 10 interviewees responded to all four abstract tasks with Child's Logic whereas only one responded to all four calculus tasks with Child's Logic. Interview responses based on some knowledge of conditionals were also given a category. Here, participants explained their work by following some rule or rules that they appear to have already internalized prior to their response. For example, at least one student explained that if you are given a conditional statement, only the modus ponens and contrapositive inferences could be made conclusively.

This approach is correct, however, it appeared to be based primarily on knowledge students had about conditionals and not on any reasoning actions that they performed during the interview.

Some believed that, given a conditional statement, it was only possible to make the modus ponens inference because the given conditional does not allow for any other possibilities.

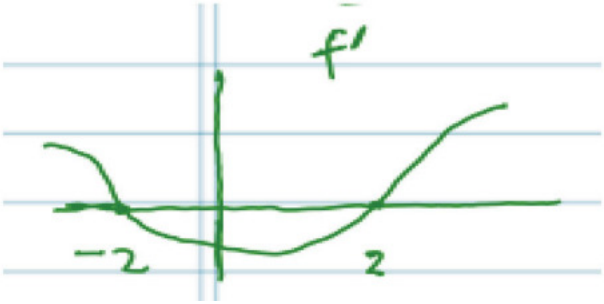
Thus, these students believe, incorrectly, that a conclusive deduction cannot be made regarding the task requiring contrapositive reasoning. This kind of partially correct thinking does provide correct responses for the inverse and converse tasks, since a conclusive inference cannot be made for 10 Case and Speer them.

However, this reasoning does not represent a complete understanding of how rules of logic apply to conditionals. In contrast with responses described below, these interviewees seemed to be recalling a rule to apply to the situation and were not engaged in extensive reasoning about the situations themselves. This strategy seemed to be the least prominent of the strategies used by the interviewees. 6.2.

Reasoning with Examples Many interviewees engaged with the tasks in a different way by generating an example or examples (via a graph or a verbalized scenario) to illustrate their thinking on at least one of the eight tasks. Although this occurred mostly with the calculus tasks, one student also used the example generating strategy when working on the abstract versions of the tasks. Two forms of this strategy were evident in the data: one utilized a single example to provide a justification and the other involved multiple examples. Interviewees used both forms of this strategy in their explanations of correct answers; however, they did not appear to be equally useful for reaching a correct conclusion. 6.2.1.

Reasoning with a Single Example In the single example approach, students generate an example graph or verbalize a relationship to explain the thinking behind their answer. Some students used this approach when providing explanations for why the contrapositive reasoning task is true. For example, Jordan was working with the theorem For all functions f, if f has a local maximum value at x = c, then c is a critical point of f and was asked to explain his answer to the associated contrapositive reasoning task

. Jordan drew a graph and tried to explain that if point c on the graph was not a critical point, then the function would not have a local maximum at that point. Although Jordan referred to various features of his graph as he tried to explain the thinking behind his correct response, he was unable to provide clear and convincing justification. After being asked to explain his answer, Jordan's response included quite a bit of hesitation and did not provide a clear chain of reasoning



5 History of Linear Algebra

In order to unfold the history of linear algebra, it is important that we first determine what Linear Algebra is. As such, this definition is not a complete and comprehensive answer, but rather a broad definition loosely wrapping itself around the subject. I will use several different answers so that we can see these perspectives

. First, linear algebra is the study of a certain algebraic structure called a vector space (BYU). Second, linear algebra is the study of linear sets of equations and their transformation properties. Finally, it is the branch of mathematics charged with investigating the properties of finite dimensional vector spaces and linear mappings between such spaces .

This project will discuss the history of linear algebra as it relates linear sets of equations and their transformations and vector spaces. The project seeks to give a brief overview of the history of linear algebra and its practical applications touching on the various topics used in concordance with it.

Around 4000 years ago, the people of Babylon knew how to solve a simple 2X2 system of linear equations with two unknowns. Around 200 BC, the Chinese published that "Nine Chapters of the Mathematical Art," they displayed the ability to solve a 3X3 system of equations (Perotti).

The simple equation of ax+b=0 is an ancient question worked on by people from all walks of life. The power and progress in linear algebra did not come to fruition until the late 17th century. The emergence of the subject came from determinants, values connected to a square matrix, studied by the founder of calculus, Leibnitz, in the late 17th century.

Lagrange came out with his work regarding Lagrange multipliers, a way to "characterize the maxima and minima multivariate functions." (Darkwing) More than fifty years later, Cramer presented his ideas of solving systems of linear equations based on determinants more than 50 years after Leibnitz (Darkwing)

. Interestingly enough, Cramer provided no proof for solving annxn system. As we see, linear algebra has become more relevant since the emergence of calculus even though it's foundational equation of ax+b=0 dates back centuries.

This prompted Jordan to generate the graph he used to illustrate his reasoning . During the discussion, Jordan was unable to provide a more compelling answer and eventually suggested the use of an equation to illustrate what was going on. A similar type of conversation also occurred when another student tried to use a single example as part of his explanation for his answer to the calculus contrapositive task.

When taking the single example approach to answering the calculus contrapositive task, students' struggles to explain their reasoning are not surprising. Visualizing one example in this situation is not going to provide the kind of solid evidence needed to obtain the appropriate conclusion for contrapositive reasoning. The two-example strategy described below was effective for students as they reasoned to obtain the correct answer for the converse and inverse reasoning tasks.

This appears to be effective because it involves generating two examples and then noting that they provide contradictory information about the truth status of the conclusion. This generates evidence that there is not enough information to decide whether the conclusion is true or false. We refer to this as the "contradictory examples" approach. Generating single examples to explore and explain the truth of the contrapositive variations to the theorems may have resulted in correct answers.

However, they were not productive approaches in the sense that students did not appear able to provide a complete explanation for their response. In large part, this is due to the nature of the contrapositive tasks. Here, contradictory examples cannot be obtained to properly infer the correct solution. In contrast with the example-generating approaches used by interviewees on the converse and inverse reasoning tasks, interviewees who gave correct answers to the abstract contrapositive reasoning tasks sometimes justified their answers by appealing to logic or by deriving the contrapositive rule itself via a contradiction argument. Findings from the survey data analysis support the notion that the pure, logical understanding exhibited by some of these students when answering the abstract contrapositive reasoning task may be useful even as they consider the calculus version of the same task

. More specifically, for the surveys, performance on the contrapositive calculus task was the only task variation that was correlated with performance on the 12 Case and Speer abstractly presented version (see Table 2). In other words, having some formal understanding of contrapositives makes it more likely that students would answer the calculus contrapositive task correctly. These findings suggest that, for the calculus contrapositive task, the example generating strategy may not be as effective compared to knowing and applying rules of logic.

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc.$$

Euler brought to light the idea that a system of equations doesn't necessarily have to have a solution (Perotti). He recognized the need for conditions to be placed upon unknown variables in order to find a

solution. The initial work up until this period mainly dealt with the concept of unique solutions and square matrices where the number of equations matched the number of unknowns.

With the turn into the 19th century Gauss introduced a procedure to be used for solving a system of linear equations. His work dealt mainly with the linear equations and had yet to bring in the idea of matrices or their notations. His efforts dealt with equations of differing numbers and variables as well as the traditional pre-19th century works of Euler, Leibnitz, and Cramer. Gauss' work is now summed up in the term Gaussian elimination. This method uses the concepts of combining, swapping, or multiplying rows with each other in order to eliminate variables from certain equations

After variables are determined, the student is then to use back substitution to help find the remaining unknown variables. As mentioned before, Gauss work dealt much with solving linear equations themselves initially, but did not have as much to do with matrices. In order for matrix algebra to develop, a proper notation or method of describing the process was necessary. Also vital to this process was a definition of matrix multiplication and the facets involving it. "The introduction of matrix notation and the invention of the word matrix were motivated by attempts to develop the right algebraic language for studying determinants. In 1848, J.J. Sylvester introduced the term "matrix," the Latin word for womb, as a name for an array of numbers. He used womb, because he viewed a matrix as a generator of determinants (Tucker, 1993).

The other part, matrix multiplication or matrix algebra came from the work of Arthur Cayley in 1855. Cayley's defined matrix multiplication as, "the matrix of coefficients for the composite transformation T2T1 is the product of the matrix for T2 times the matrix of T1" (Tucker, 1993). His work dealing with Matrix multiplication culminated in his theorem, the Cayley-Hamilton Theorem. Simply stated, a square matrix satisfies its characteristic equation. Cayley's efforts were published in two papers, one in 1850 and the other in 1858. His works introduced the idea of the identity matrix as well as the inverse of a square matrix. He also did much to further the ongoing transformation of the use of matrices and symbolic algebra. He used the letter "A" to represent a matrix, something that had been very little before his works. His efforts were little recognized outside of England until the 1880s.

Matrices at the end of the 19th century were heavily connected with Physics issues and for mathematicians, more attention was given to vectors as they proved to be basic mathematical elements. For a time, however, interest in a lot of linear algebra slowed until the end of World War II brought on the development of computers. Now instead of having to break down an enormous nxn matrix, computers could quickly and accurately solve these systems of linear algebra. With the advancement of technology using the methods of Cayley, Gauss, Leibnitz, Euler, and others determinants and linear algebra moved forward more quickly and more effective. Regardless of the technology though Gaussian elimination still proves to be the best way known to solve a system of linear equations (Tucker, 1993)

. The influence of Linear Algebra in the mathematical world is spread wide because it provides an important base to many of the principles and practices. Some of the things Linear Algebra is used for are to solve systems of linear format, to find least-square best fit lines to predict future outcomes or find trends, and the use of the Fourier series expansion as a means to solving partial differential equations. Other more broad topics that it is used for are to solve questions of energy in Quantum mechanics. It is also used to create simple every day household games like Sudoku. It is because of these practical

applications that Linear Algebra has spread so far and advanced. The key, however, is to understand that the history of linear algebra provides the basis for these applications. Although linear algebra is a fairly new subject when compared to other mathematical practices, it's uses are widespread.

With the efforts of calculus savvy Leibnitz the concept of using systems of linear equations to solve unknowns was formalized. Other efforts from scholars like Cayley, Euler, Sylvester, and others changed linear systems into the use of matrices to represent them. Gauss brought his theory to solve systems of equations proving to be the most effective basis for solving unknowns. Technology continues to push the use further and further, but the history of Linear Algebra continues to provide the foundation

. Even though every few years companies update their textbooks, the fundamentals stay the same.

6 Representation of Functions

We are mainly interested in representations that are potentially effective in high dimensions. Therefore we will focus on the ones that can be expressed as expectations. As an example, instead of the Fourier representation:

$$f(\boldsymbol{x}) = \int_{\mathbb{R}^d} a(\boldsymbol{\omega}) e^{i(\boldsymbol{\omega},\boldsymbol{x})} d\boldsymbol{\omega},$$

we will consider

$$f(\boldsymbol{x}) = \int_{\mathbb{R}^d} a(\boldsymbol{\omega}) e^{i(\boldsymbol{\omega}\boldsymbol{x})} \pi(d\boldsymbol{\omega}) = \mathbb{E}_{\boldsymbol{\omega}\sim\pi} a(\boldsymbol{\omega}) e^{i(\boldsymbol{\omega},\boldsymbol{x})} \tag{2}$$

where π is a probability measure on $\mathbb{R}^d$. The reason that we prefer (2) over (1) is as follows. The discrete analog of (1) is

$$f_m(\boldsymbol{x}) = \frac{1}{m} \sum_j a(\boldsymbol{\omega}_j) e^{i(\boldsymbol{\omega}_j,\boldsymbol{x})} \tag{3}$$

where the sum is performed on a regular grid $\{\boldsymbol{\omega}_j\}_{j=1}^m$ in the Fourier space. It is well-known that this kind of grid-based approximations satisfies

$$f - f_m \sim C(f) m^{-\alpha/d} \tag{4}$$

where $C(f)$ and α are fixed quantities depending on f. The appearance of $1/d$ in the exponent of m signals the curse of dimensionality. In contrast, for (2), by independently sample $\{\boldsymbol{\omega}_j\}_{j=1}^m$ from π, we obtain an approximation to f with a dimension-independent error rate:

$$\mathbb{E}|f(\boldsymbol{x}) - \frac{1}{m} \sum_{j=1}^m a(\boldsymbol{\omega}_j) e^{i(\boldsymbol{\omega}_j,\boldsymbol{x})}|^2 = \frac{\mathrm{var}(f)}{m}$$

From an algorithmic viewpoint, is typically associated with non-adaptive discretizations such as the spectral method or the ridglets and curvelets used in signal processing . We will see later that the forms and are closely associated with the random feature model and the two-layer neural network model

7 Gradient flows

The third component in machine learning is an algorithm for solving the optimization problem. In this section, we will discuss various gradient flow dynamics for the population or empirical risk. For simplicity we focus on the following loss functional

We first discuss gradient flows using a physics language .The loss functions or functionals defined above serve as the "free energy" of the problem. To begin with, we need to distinguish conserved and non-conserved "order parameters".

The coefficient a in is non-conserved. The probability distributions π or ρ are obviously conserved. First, let us examine the situation with the representation (6). Let I = I(a, π) be the loss functional. Denote by δIδa and δI δπ the formal variational derivative of I with respect to a and π respectively, under the standard L 2 metric. The gradient flow for a is simply given by

$$\frac{\sigma a}{\partial t} = -\frac{o\imath}{\delta a} \tag{29}$$

In the physics literature, this is known as the "model A" dynamics [39].

The gradient flow for π is given by a continuity equation:

$$\frac{\partial \pi}{\partial t} + \nabla \cdot \mathbf{J} = 0 \tag{30}$$

where the current $\mathbf{J}$ is given by:

$$\mathbf{J} = \pi \mathbf{v},\ \mathbf{v} = -\nabla V$$

$$V = \frac{\delta I}{\delta \pi}.$$

This is known as the "model B" dynamics [39] and V is known as the "chemical potential".

Remark 2. *It is well-known that the model B dynamics is also the gradient flow under the 2-Wasserstein metric [41, 65].*

For flow-based models, the parameters a and π are themselves one-parameter families of coefficients or probability distributions respectively: $a = (a_\tau)_{\tau\in[0,1]}, \pi = (\pi_\tau)_{\tau\in[0,1]}$. Given a functional $I, I = I((a_\tau), (\pi_\tau))$, a natural extension of the gradient flow to $(a_\tau), (\pi_\tau)$ is given by:

$$\frac{\partial a_\tau}{\partial t} = -\frac{\delta I}{\delta a_\tau}$$

$$\frac{\partial \pi_\tau}{\partial t} + \nabla \cdot \mathbf{J}_\tau = 0,$$

where

$$\mathbf{J}_\tau = -\pi_\tau \nabla \frac{\delta I}{\delta \pi_\tau}.$$

8 A smoothed particle methods

A popular modification of the particle method is the smoothed particle method. Here we illustrate how one can formulate the smoothed particle method for the integral transformbased model and the gradient flow .We will consider the special case when φ(x; w) = aσ(b Tx). Here w = (a, b) and σ(t) = max(0, t) is the ReLU activation function. Consider a smoothed particle approximation

$$\hat{\pi}_t(\boldsymbol{w}) = \frac{1}{m}\sum_{k=1}^{m} \phi_h(\boldsymbol{w} - \boldsymbol{w}_k(t)),$$

where ϕ_h is the probability density function of $\mathcal{N}(0, h^2 I)$. The smoothed particle discr tion of the flow-based model and the gradient flow is given by

$$\begin{aligned} f(\boldsymbol{x}; \hat{\pi}_t) &= \mathbb{E}_{\boldsymbol{w}\sim\hat{\pi}_t}[\phi(\boldsymbol{x}; \boldsymbol{w})] \\ &= \frac{1}{m}\sum_{k=1}^{m} \mathbb{E}_{\boldsymbol{\xi}}[\phi(\boldsymbol{x}; \boldsymbol{w}_k + h\boldsymbol{\xi})] \\ \frac{d\boldsymbol{w}_k}{dt} &= \mathbb{E}_{\boldsymbol{w}\sim\phi_h(\cdot-\boldsymbol{w}_k(t))}[\boldsymbol{v}(\hat{\pi}_t, \boldsymbol{w})] \\ &= \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{v}(\hat{\pi}_t, \boldsymbol{w}_k + h\boldsymbol{\xi})] \end{aligned}$$

where ξ ∼ N (0, Id+1). The right hand side of the last equality is the smoothed velocity. For this to be a practical numerical algorithm, we need a way to evaluate them and We will defer this to a future publication. To get some insight about the nature of this smooth particle method

we consider the special case when the data lies on the sphere, i.e.kxk = 1. Write $\boldsymbol{\xi} = (\xi_1, \boldsymbol{\xi}_2)$ with $\xi_1 \in \mathbb{R}$ and $\boldsymbol{\xi}_2 \in \mathbb{R}^d$, then the smoothed particle method becomes

$$\begin{aligned} f(\boldsymbol{x};\hat{\pi}) &= \mathbb{E}_{(a,\boldsymbol{b})\sim\hat{\pi}}[a\sigma(\boldsymbol{b}^T\boldsymbol{x})] \\ &= \frac{1}{m}\sum_{k=1}^{m}\mathbb{E}_{(\xi_1,\boldsymbol{\xi}_2)}[(a_k+h\xi_1)\sigma((\boldsymbol{b}_k+h\boldsymbol{\xi}_2)^T\boldsymbol{x})] \\ &= \frac{1}{m}\sum_{k=1}^{m}a_k\mathbb{E}_{\boldsymbol{\xi}_2}[\sigma(\boldsymbol{b}_k^T\boldsymbol{x}+h\boldsymbol{\xi}_2^T\boldsymbol{x})] \\ &= \frac{1}{m}\sum_{k=1}^{m}a_k\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}[\sigma(\boldsymbol{b}_k^T\boldsymbol{x}+h\xi)], \end{aligned} \tag{84}$$

where in the last equation we have used the assumption that $\|\boldsymbol{x}\| = 1$. Define a new activation function

$$\begin{aligned} \sigma_h(t) &= \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}[\sigma(t+h\xi)] = \int_{-t/h}^{\infty}(t+h\xi)\frac{1}{\sqrt{2\pi}}e^{-\xi^2/2}d\xi. \\ &= t\Phi(\frac{t}{h})+h\phi(\frac{t}{h}), \end{aligned} \tag{85}$$

where ϕ, Φ are the probability density and cumulative density functions of the standard normal distribution, respectively. Then the discretized model can be rewritten as

$$f(\boldsymbol{x};\hat{\pi}) = \frac{1}{m}\sum_{k=1}^{m}a_k\sigma_h(\boldsymbol{b}_k^T\boldsymbol{x}). \tag{86}$$

This is a new two-layer neural network with activation function σh, which can be viewed as a "smoothed" ReLU. It is easy to see supt∈R |σh(t)−σ(t)| = O(h). Figure 1 shows the difference between the two activation functions.

$$\begin{aligned} \frac{da_k}{dt} &= \mathbb{E}_{\boldsymbol{\xi}}[\mathbb{E}_{\boldsymbol{x}}[(f(\boldsymbol{x};\hat{\pi})-f^*(\boldsymbol{x}))\sigma(\boldsymbol{b}_k^T\boldsymbol{x}+h\boldsymbol{\xi}_2^T\boldsymbol{x})]] \\ &= \mathbb{E}_{\boldsymbol{x}}[(f(\boldsymbol{x};\hat{\pi})-f^*(\boldsymbol{x}))\sigma_h(\boldsymbol{b}_k^T\boldsymbol{x})] \\ \frac{d\boldsymbol{b}_k}{dt} &= \mathbb{E}_{(\boldsymbol{\xi_1},\boldsymbol{\xi_2})}[\mathbb{E}_{\boldsymbol{x}}[(f(\boldsymbol{x};\hat{\pi})-f^*(\boldsymbol{x}))(a+h\xi_1)\sigma'(\boldsymbol{b}_k^T\boldsymbol{x}+h\boldsymbol{\xi}_2^T\boldsymbol{x})\boldsymbol{x}]] \\ &= \mathbb{E}_{\boldsymbol{x}}[(f(\boldsymbol{x};\hat{\pi})-f^*(\boldsymbol{x}))a\sigma_h'(\boldsymbol{b}_k^T\boldsymbol{x})\boldsymbol{x}]. \end{aligned}$$

9 Analyzing the discretized model

We decompose the generalization error into two terms:

$$\mathcal{R}(\hat{f}_{m,n,t}) = \underbrace{\hat{\mathcal{R}}_n(\hat{f}_{m,n,t})}_{I_1} + \underbrace{\mathcal{R}(\hat{f}_{m,n,t}) - \hat{\mathcal{R}}_n(\hat{f}_{m,n,t})}_{I_2}. \tag{103}$$

Here I_1, I_2 are the optimization (training) error and generalization gap, respectively.

The general philosophy is that the generalization gap is bounded by a term of the form $\|f_{m,n,t}\|/\sqrt{n}$. Here $\|\cdot\|$ is some norm determined by the model. For example, for random feature models, this is the RKHS norm. For two-layer neural network models, this is the Barron norm [26]. Therefore to estimate the generalization gap, one needs to derive a priori bounds on these norms.

For any $\bar{\boldsymbol{a}} \in \mathbb{R}^m$, define

$$J(t) := t(\hat{\mathcal{R}}_n(\boldsymbol{a}_t) - \hat{\mathcal{R}}_n(\bar{\boldsymbol{a}})) + \frac{1}{2}\|\boldsymbol{a}_t - \bar{\boldsymbol{a}}\|_2^2. \tag{104}$$

Since $\hat{\mathcal{R}}_n(\boldsymbol{a})$ is convex, we have $dJ/dt \le 0$. So $J(t) \le J(0)$, i.e.

$$t(\hat{\mathcal{R}}_n(\boldsymbol{a}_t) - \hat{\mathcal{R}}_n(\bar{\boldsymbol{a}})) + \frac{1}{2}\|\boldsymbol{a}_t - \bar{\boldsymbol{a}}\|_2^2 \le \frac{1}{2}\|\boldsymbol{a}_0 - \bar{\boldsymbol{a}}\|_2^2.$$

This gives

$$\begin{aligned} \hat{\mathcal{R}}_n(\boldsymbol{a}_t) &\le \frac{\|\bar{\boldsymbol{a}}\|_2^2}{2t} \\ \|\boldsymbol{a}_t\|_2^2 &\le 2\|\bar{\boldsymbol{a}}\|_2^2 + 2t\hat{\mathcal{R}}_n(\bar{\boldsymbol{a}}). \end{aligned} \tag{105}$$

$$\mathcal{R}(\boldsymbol{a}) \lesssim \hat{\mathcal{R}}_n(\boldsymbol{a}) + \frac{\|\boldsymbol{a}\|^\iota/m + \|f^*\|_{\mathcal{H}}^\iota}{\sqrt{n}} \left(1 + \sqrt{\log((\|\boldsymbol{a}\|/\sqrt{m}+1)^2/\delta)}\right). \tag{106}$$

Proof. Let $\mathcal{F}_C := \{f_m(\cdot;\boldsymbol{a},\mathbf{B}^0) : \|\boldsymbol{a}\|/\sqrt{m} \le C\}$ and $\mathcal{H}_C := \{(f_m(\cdot;\boldsymbol{a},\mathbf{B}_0) - f^*)^2 : \|\boldsymbol{a}\|/\sqrt{m} \le C\}$. By Cauchy-Schwarz inequality, $|f_m(\boldsymbol{x};\boldsymbol{a},B_0)| \le C$ and $|f^*(\boldsymbol{x})| \le \sqrt{\int a(\boldsymbol{b})^2 d\pi(\boldsymbol{b})} \le \|f^*\|_{\mathcal{H}}$. Hence, $g(t) = (t-y_i)^2$ is $2(C+\|f^*\|_{\mathcal{H}})$-Lipschitz continuous. Then by the contraction property of Rademacher complexity, we have

$$\mathrm{Rad}_n(\mathcal{H}_C) \le 2(C + \|f^*\|_{\mathcal{H}})\mathrm{Rad}_n(\mathcal{F}_C) \le \frac{2C(C + \|f^*\|_{\mathcal{H}})}{\sqrt{n}},$$

where the last inequality follows from the fact that $\mathrm{Rad}(\mathcal{F}_C) \le \frac{C}{\sqrt{n}}$ [58]. Hence, with probability $1-\delta$ we have for any $\boldsymbol{a}$ satisfying $\|\boldsymbol{a}\|/\sqrt{m} \le C$,

$$\mathcal{R}(\boldsymbol{a}) \lesssim \hat{\mathcal{R}}_n(\boldsymbol{a}) + \frac{(C + \|f^*\|_{\mathcal{H}})C}{\sqrt{n}} + (\|f^*\|_{\mathcal{H}} + C)^2\sqrt{\frac{\log(2/\delta)}{n}} \tag{107}$$

$$\lesssim \hat{\mathcal{R}}_n(\boldsymbol{a}) + \frac{C^2 + \|f^*\|_{\mathcal{H}}^2}{\sqrt{n}} \left(1 + \sqrt{\log(1/\delta)}\right). \tag{108}$$

10 Analyzing the continuous model

The approach presented above is the standard approach in machine learning theory. It works since the loss functional is convex in this case. It is difficult to generalize this to more complicated situations due to the lack of convexity. Here we explore an alternative approach by studying the continuous problem. Our hope is that some of the PDE techniques can be leveraged to help our understanding. One such example is found in which proves a global convergence result for the gradient flow for two-layer neural networks by analyzing the PDE

$$\partial_t a(\boldsymbol{w}, t) = -\frac{\delta \hat{\mathcal{R}}_n}{\delta a}.$$

Similar to above, we define

$$J(t) := t(\hat{\mathcal{R}}_n(a_t) - \hat{\mathcal{R}}_n(a^*)) + \frac{1}{2}\|a_t - a^*\|^2_{L^2(\pi)}.$$

Then we have

$$\begin{aligned}\frac{dJ(t)}{dt} &= -t\|\frac{\delta \hat{\mathcal{R}}_n}{\delta a}\|^2_{L^2(\pi)} + \hat{\mathcal{R}}_n(a_t) - \hat{\mathcal{R}}_n(a^*) + \langle a_t - a^*, -\frac{\delta \hat{\mathcal{R}}_n}{\delta a}\rangle_{L^2(\pi)} \\ &\leq -t\|\frac{\delta \hat{\mathcal{R}}_n}{\delta a}\|^2_{L^2(\pi)} \leq 0,\end{aligned}$$

where the second inequality follows from the convexity of $\mathcal{R}_n$ with respect to a.
that

$$t(\hat{\mathcal{R}}_n(a_t) - \hat{\mathcal{R}}_n(a^*)) + \frac{1}{2}\|a_t - a^*\|^2_{L^2(\pi)} \leq \frac{1}{2}\|a_0 - a^*\|^2_{L^2(\pi)}.$$

Since $\boldsymbol{a}_0 = 0$ and $\hat{\mathcal{R}}_n(a^*) = 0$, we get

$$\begin{aligned}\hat{\mathcal{R}}_n(a_t) &\leq \frac{\|a^*\|^2_{L^2(\pi)}}{2t} \\ \|a_t\|_{L^2(\pi)} &\leq 2\|a^*\|_{L^2(\pi)}.\end{aligned}$$

References

[1] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savar´e. Gradient flows: in metric spaces and in the space of probability measures. Springer Science & Business Media, 2008.

[2] Dyego Ara´ujo, Roberto I Oliveira, and Daniel Yukimura. A mean-field limit for certain deep neural networks. arXiv preprint arXiv:1906.00193, 2019.

[3] Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. In Advances in Neural Information Processing Systems, pages 6481–6491, 2019.

[4] Benny Avelin and Kaj Nystr¨om. Neural ODEs as the deep limit of ResNets with constant weights. arXiv preprint arXiv:1906.12183, 2019.

5] Francis Bach. Breaking the curse of dimensionality with convex neural networks. Journal of Machine Learning Research, 18(19):1–53, 2017.

[6] Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. IEEE Transactions on Information theory, 39(3):930–945, 1993.

[7] Peter L Bartlett, Steven N Evans, and Philip M Long. Representing smooth functions as compositions of near-identity functions with implications for deep network optimization. arXiv preprint arXiv:1804.05012, 2018.

[8] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. Journal of Machine Learning Research, 3(Nov):463–482, 2002.

[9] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machinelearning practice and the classical bias–variance trade-off. Proceedings of the National Academy of Sciences, 116(32):15849–15854, 2019.

[10] Vladimir G Boltyanskii, Revaz V Gamkrelidze, and S Lev. Pontryagin. the theory of optimal processes. i. the maximum principle. Technical report, TRW Space Tochnology Labs, Los Angeles, California, 1960.

[11] Emmanuel J Cand`es. Harmonic analysis of neural networks. Applied and Computational Harmonic

Analysis, 6(2):197–218, 1999.

[12] Emmanuel J Cand`es and David L Donoho. Ridgelets: A key to higherdimensional intermittency? Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, 357(1760):2495–2509, 1999.

[13] Giuseppe Carleo and Matthias Troyer. Solving the quantum many-body problem with artificial neural networks. Science, 355(6325):602–606, 2017.

[14] Luigi Carratino, Alessandro Rudi, and Lorenzo Rosasco. Learning with SGD and random features. In Advances in Neural Information Processing Systems, pages 10213–10224, 2018.

[15] Ricky T Q Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In Advances in neural information processing systems, pages 6571–6583, 2018. 39
[16] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for overparameterized models using optimal transport. In Advances in neural information processing systems, pages 3036–3046, 2018

. [17] Philippe G Ciarlet. The finite element method for elliptic problems. Classics in applied mathematics, 40:1–511, 2002.

[18] George Cybenko. Approximation by superpositions of a sigmoidal function. Mathematics of control, signals and systems, 2(4):303–314, 1989.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018. [

20] Weinan E. A proposal on machine learning via dynamical systems. Communications in Mathematics and Statistics, 5(1):1–11, 2017.

[21] Weinan E. Machine learning: Mathematical theory and scientific applications. Notices of the American Mathematical Society, 66(11), 2019.

[22] Weinan E, Jiequn Han, and Arnulf Jentzen. Deep learning-based numerical methods for highdimensional parabolic partial differential equations and backward stochastic differential equations. Communications in Mathematics and Statistics, 5(4):349–380, 2017.

[23] Weinan E, Jiequn Han, and Qianxiao Li. A mean-field optimal control formulation of deep learning. Research in the Mathematical Sciences, 6(1):10, 2019.

[24] Weinan E, Chao Ma, and Lei Wu. Barron spaces and the compositional function spaces for neural network models. arXiv preprint arXiv:1906.08039, 2019.

[25] Weinan E, Chao Ma, and Lei Wu. A comparative analysis of the optimization and generalization property of two-layer neural network and random feature models under gradient descent dynamics. Science China Mathematics, pages 124, 2020; arXiv preprint arXiv:1904.04326, 2019.

[26] Weinan E, Chao Ma, and Lei Wu. A priori estimates of the population risk for two-layer neural networks. Communications in Mathematical Sciences, 17(5):1407–1425, 2019; arXiv preprint arXiv:1810.06397, 2018

. [27] Weinan E and Bing Yu. The deep Ritz method: a deep learning-based numerical algorithm for solving variational problems. Communications in Mathematics and Statistics, 6(1):1–12, 2018.

[28] George E Forsythe and Wolfgang R Wasow. Finite-difference Methods for Partial Differential Equations. Applied mathematics series. Wiley, 1967.

[29] David Gottlieb and Steven A Orszag. Numerical analysis of spectral methods: theory and applications, volume 26. SIAM, 1977.

[30] Bertil Gustafsson, Heinz-Otto Kreiss, and Joseph Oliger. Time dependent problems and difference methods, volume 24. John Wiley & Sons, 1995

[31] Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. Inverse Problems, 34(1):014004, 2017.

[32] Jiequn Han and Weinan E. Deep learning approximation for stochastic control problems. NIPS2016, Deep Reinforcement Learning Workshop, 2016.

[33] Jiequn Han, Arnulf Jentzen, and Weinan E. Solving high-dimensional partial differential equations using deep learning. Proceedings of the National Academy of Sciences, 115(34):8505–8510, 2018. 40. arXiv preprint arXiv:1901.06523, 2019.

