# Does Representation Matter? Comparing Algebraic and Geometric Approaches to Teaching L1/L2 Regularization

## Effects on Conceptual Understanding, Problem-Solving, and Knowledge Transfer

**Ivan Nikolov**[1]

**Supervisor(s): Gosia Migut**[1]**, Ilinca Rențea**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Ivan Nikolov
Final project course: CSE3000 Research Project
Thesis committee: Gosia Migut , Ilinca Rențea , Jorge Martinez Castaneda

## Abstract

As machine learning becomes a standard part of computing and engineering curricula, teaching its core concepts effectively has become an important educational challenge. Regularization is one such concept: L1 and L2 penalties are widely used to prevent overfitting, with L1 producing sparse solutions and L2 shrinking weights more smoothly. It is commonly taught through two representational formats - an algebraic one, presenting the loss function and its penalty terms, and a geometric one, depicting constraint regions and their intersection with the loss contours - yet instructors choose between them without clear evidence on which better supports learning. This study asks whether teaching regularization algebraically or geometrically leads to different student performance, and in which kinds of understanding. Two interactive notebooks, matched on learning objectives, length, and reading difficulty, were developed to teach the concept in each format and are released openly; they were compared in a between-subjects experiment with students who had completed an introductory machine learning course. Learning was measured with a post-test spanning conceptual understanding, problem-solving, and knowledge transfer, alongside a thematic analysis of students' written explanations. Both formats supported practical reasoning about regularization, but not identically: the algebraic group performed better overall, with its clearest advantage in conceptual understanding, no reliable difference on problem-solving, and an inconclusive result on transfer. The two groups largely shared the same core understanding but expressed it through different vocabularies - a penalty on the loss function versus a shrinking constraint region in weight space. Representational choice therefore appears to shape how students explain regularization more than whether they grasp it, suggesting that combining the two formats may best support learning.


## 1 Introduction

Machine learning is increasingly treated as a core competency in computing and engineering education, as demand for AI/ML skills has grown in both professional job markets and academic contexts [1; 2]. This has created a pressing need for instructional strategies which reliably support understanding of ML's core concepts.

Regularization is one such concept. L1 and L2 regularization are among the most widely applied techniques in machine learning, used to prevent overfitting by penalizing large model weights during training [3; 4].When teaching L1 and L2 regularization, an important instructional goal is to help students relate the penalty terms in the loss function to their practical effects on model behaviour: reducing overfitting, controlling model complexity, and producing different coefficient patterns. In particular, L1 regularization is associated with sparse solutions, whereas L2 regularization tends to shrink weights more smoothly [5; 6].

These techniques are commonly taught through two distinct representational formats. The algebraic format presents regularization through the loss function and its penalty terms, emphasizing the mathematical formulation and how the penalty enters the optimization. The geometric format presents regularization visually, depicting the constraint regions (the diamond L1 boundary and circular L2 boundary) and their intersection with the loss contours, making the origin of sparsity in L1 and smooth shrinkage in L2 visually apparent. Each format appears to emphasize different aspects of the concept, making it useful to examine how these representations may support different kinds of learning. This matters because the choice of representation may shape not only whether students can recall a definition, but whether they can solve problems and transfer the concept to new situations - the kinds of understanding that distinguish durable learning from surface familiarity [7; 8].

Despite the prevalence of both formats in textbooks and courses, there is little empirical work directly comparing their instructional effectiveness for this topic. This study addresses that gap.

> **Research question:** Does teaching L1 and L2 regularization algebraically or geometrically lead to any differences in student performance across conceptual understanding, problem-solving, and knowledge transfer?

The main contribution of this study is an experimental evaluation of the two representational formats for teaching L1 and L2 regularization, producing evidence about which format more effectively supports which type of understanding. The two interactive notebooks developed for this study are made publicly available to support replication, further research, and reuse by other educators in their own teaching.

The remainder of the paper is structured as follows. Section 2 reviews related work. Section 3 describes the methodology. Section 4 presents the quantitative and qualitative results. Section 5 discusses the findings. Section 6 reflects on the ethical aspects of the research and discusses the reproducibility of the methods. Section 7 concludes with implications for ML teaching practice and directions for future research.

## 2 Related Work

### 2.1 Machine Learning Education

Machine learning education is still a relatively new research area. Although machine learning is now widely used in university programmes and professional fields, there is still limited research on how specific machine learning concepts should be taught. Shapiro, Fiebrink, and Norvig argue that machine learning has become important enough that it should receive more attention in undergraduate computing curricula

[9]. Shapiro and Fiebrink later proposed a research agenda for machine learning education, identifying students' understanding of algorithms and their parameters as an important open problem [10]. This issue is relevant to regularization, where students must understand not only the presence of a penalty term, but also how changing this term affects the learned model.

Other work has examined what students should learn in introductory machine learning courses. Burnik argues that machine learning education should be organized around ordered competences instead of separate algorithms [2]. Sulmont, Patitsas, and Cooper argue that the field needs its own pedagogical content knowledge, because teachers need to know which parts of machine learning are difficult for students and how to address them [11]. These studies suggest that machine learning education should focus not only on implementation, but also on conceptual understanding of the mechanisms behind algorithms.

There is also related work on tools and learning environments for teaching machine learning concepts. Fleischer, Biehler, and Schulte studied educationally designed Jupyter Notebooks for teaching data-driven machine learning, highlighting how notebooks can integrate code, explanatory text, and visualisations to support interactive learning [12]. Gresse von Wangenheim et al. reviewed visual tools and found that many are designed to help learners build intuition about models through visual interaction [13]. Rentea, Migut, and Krijthe studied interactive visualizations for topics such as gradient descent and PCA [14]. This work shows that visual and interactive materials are common in machine learning education, but it provides less evidence about when such materials are more effective than symbolic or algebraic explanations.

Several studies also show that students can struggle with the underlying mechanisms of machine learning concepts. Valdenegro-Toro and Sabatelli describe misconceptions about overfitting in bachelor- and master-level courses [15]. They found that students often know that overfitting is a problem, but cannot always explain what causes it or how to address it correctly. Angelescu similarly found misunderstandings of central topics such as cross-validation, principal component analysis, and gradient descent in a bachelor-level course [16]. Jin reported related misconceptions about gradient descent [17]. Together, these findings motivate research that looks beyond whether students recognize terminology and instead examines whether they understand the mechanisms behind machine learning techniques. Regularization is a suitable case for this kind of investigation because it requires students to connect a mathematical penalty term with its practical effect on model behaviour, including model complexity, overfitting, and the different coefficient behaviour associated with L1 and L2 penalties. The present study therefore examines whether students learn this concept differently depending on how it is represented instructionally.

## 2.2 Learning with Multiple Representations

The comparison between algebraic and geometric explanations is grounded in research on learning with multiple representations. In this context, a representation is not only a way of presenting information, but also a tool that can shape how students reason about a concept. Paivio's Dual Coding Theory suggests that verbal and visual information are processed through different but connected systems [18]. Mayer's Cognitive Theory of Multimedia Learning similarly argues that students can learn more deeply when they actively connect verbal explanations with visual information [19]. Ainsworth's DeFT framework further explains that different representations can support learning by complementing each other, constraining possible misunderstandings, and helping learners construct a more complete understanding [7].

However, visual representations are not automatically better than symbolic ones. Their usefulness depends on the concept being taught, the task students need to complete, and the learners' prior knowledge. Research from mathematics education illustrates this point. Koedinger and Nathan found that students sometimes performed better on word-based algebra problems than on the same problems written as symbolic equations, because the word problems were more familiar [20]. In contrast, Koedinger, Alibali, and Nathan showed that symbolic representations can become more useful for more complex problems, where students need to reason abstractly [21]. This tension between geometric and symbolic representations has been studied most directly in undergraduate linear algebra, where the same objects can be expressed geometrically, algebraically, or in abstract structural terms. Sierpinska distinguishes between synthetic-geometric, analytic-arithmetic, and analytic-structural modes of thinking and attributes many student difficulties not to any single representation but to students' inability to move flexibly between them [22]. Studying matrix algebra students, Dogan-Dunlap found that geometric representations complemented rather than replaced symbolic and computational reasoning [23], and reviews of this literature treat the coordination of geometric and algebraic descriptions as a central learning challenge in its own right [24].Together, these findings suggest that geometric and symbolic formats support different kinds of reasoning, and that a geometric representation only helps when learners can connect it to the underlying formal structure.

This perspective is directly relevant to L1 and L2 regularization. Textbooks commonly present regularization as a method for controlling model complexity and reducing overfitting [3; 4]. L1 regularization is associated with sparse solutions, while L2 regularization tends to shrink coefficients more smoothly [5; 6]. These differences can be explained algebraically, through the loss function and penalty term, or geometrically, through constraint regions, loss contours, and the position of the optimal solution. From a multiple-representations perspective, the algebraic format may better support formal reasoning about the objective function, whereas the geometric format may better support intuition about the solution space. However, there is limited direct empirical evidence comparing these two formats for teaching regularization. The present study addresses this gap by experimentally comparing their effects on conceptual understanding, problem-solving performance, and knowledge transfer.

## 3 Methodology

This section describes the development and evaluation of the instructional materials, the design of the experiment, and the methods used to analyse the collected data. An overview of the full methodological procedure is shown in Figure 1.

### 3.1 Interactive Materials Development

Before developing any material, a shared set of learning objectives was defined to specify exactly which concepts the notebooks should teach (see Appendix A). Two Jupyter Books were then created to cover these objectives, presenting regularization through two complementary representations: an algebraic representation, focused on formulas and numerical coefficient shrinkage, and a geometric representation, focused on constraint regions and visual intuition in weight space.

The algebraic notebook follows a worked example adapted from Hull's *Machine Learning in Business* textbook [25], where an overly complex polynomial model is fitted to predict salary from age. Specifically, a 5th-degree polynomial is fitted to 10 training observations, producing a model that fits the training points almost perfectly but generalizes poorly, the unregularized coefficients reach values in the tens of thousands. The notebook walks students step by step through this example: first by showing the overfitting problem directly, then by introducing the Ridge and Lasso objectives, and finally demonstrating the effect of varying lambda on the resulting coefficients for both regularization techniques. Three interactive visualizations accompany the algebraic notebook. The first two show how the learning curve changes for different values of $\lambda$, for Ridge and Lasso respectively. The third visualization shows the difference in shrinking paths between the two techniques: under L1, weights are driven exactly to zero one by one as $\lambda$ grows, performing feature selection, while under L2 all weights shrink smoothly toward zero, but none are eliminated, producing dense rather than sparse weights.

The geometric notebook introduces regularization through the lens of constraint regions and weight space geometry [4]. The notebook follows a simplified two-weight model, which allows the regularization problem to be drawn as a two-dimensional picture. Students are first introduced to the weight plane, then to loss contours, ellipses that represent the levels of training error. After that, regularization is reframed geometrically, rather than adding an additional penalty to the loss function, weights are restricted to lie inside a bounded region around the origin. Ridge regression constrains weights to a circle, where the loss ellipse can touch almost anywhere producing dense weights. On the contrary, Lasso constrains weights to a diamond, where the loss ellipse frequently makes first contact at the sharp corners, leading to sparse weights. An interactive visualization supports this explanation: sliders control the size of $\lambda$ and the position from which the loss contours approach the constraint region, allowing students to directly observe how the shape of the boundary determines where the solution lands.

Since the study compares two representational formats, the notebooks were developed in parallel to cover the same learning objectives at a comparable level of complexity. The two books shared the same introduction and conclusion, differing only in the main chapter, and were closely matched in reading difficulty, measured with the Flesch–Kincaid grade level[26; 27] (7.2 vs 7.9, with identical syllables per word). The geometric chapter was somewhat longer (2,354 vs 1,982 words), while the algebraic notebook contained more interactive visualizations (three versus one); the two were designed to impose broadly comparable workload, with the additional reading in one offset by the additional interactivity in the other.

### 3.2 Expert Evaluation

To assess the correctness and clarity of the materials, an expert evaluation was conducted prior to the experiment with a teaching assistant (TA) from a machine learning course. A TA was chosen deliberately: besides being familiar with both the subject matter and the post-test, a TA could devote more time to a careful, detailed review than a more senior reviewer such as a professor, whose availability is typically more limited. The evaluation was intended as a content-validity and clarity check to screen the materials before the experiment, a purpose for which a single, well-matched expert is appropriate. The TA was given both interactive notebooks and an evaluation rubric (see Appendix B) assessing the materials across five areas: correctness, clarity, quality of visualizations, alignment with the assessment, and overall readiness for use. Each area was rated on a five-point quality scale and accompanied by open-ended comments.

All five criteria were rated 4 or 5 out of 5: correctness, visualization quality, and overall readiness each received the maximum of 5, while clarity and alignment with the assessment each received 4. Both notebooks were judged to be correct and to use clear language, with concise explanations and consistent formatting. The feedback consisted mainly of minor clarity issues, all of which were addressed before the experiment. One comment had implications beyond clarity: the TA observed that the post-test might slightly favour the algebraic notebook, since several questions concern overfitting, a topic that notebook emphasised more heavily. In response, the two notebooks were rebalanced so that overfitting received comparable coverage in both, ensuring that neither representation was advantaged by the assessment's emphasis.

### 3.3 Experiment Design

To address this research question, a between-subjects experiment was conducted to measure whether the representational format through which regularization is taught, algebraic versus geometric, produces different patterns of conceptual understanding, problem-solving, and knowledge transfer. This section describes the methodology of the experiment.

**Participants.** The experiment recruited 20 Computer Science and Aerospace Engineering Bachelor students who have successfully completed an introductory Machine Learning course but had no prior familiarity with regularization as a topic.

**Procedure.** The procedure began with participants completing a pre-test (see Appendix C). Based on the pre-test results, participants were divided into two groups: the algebraic notebook group (N = 10) and the geometric group (N = 10).

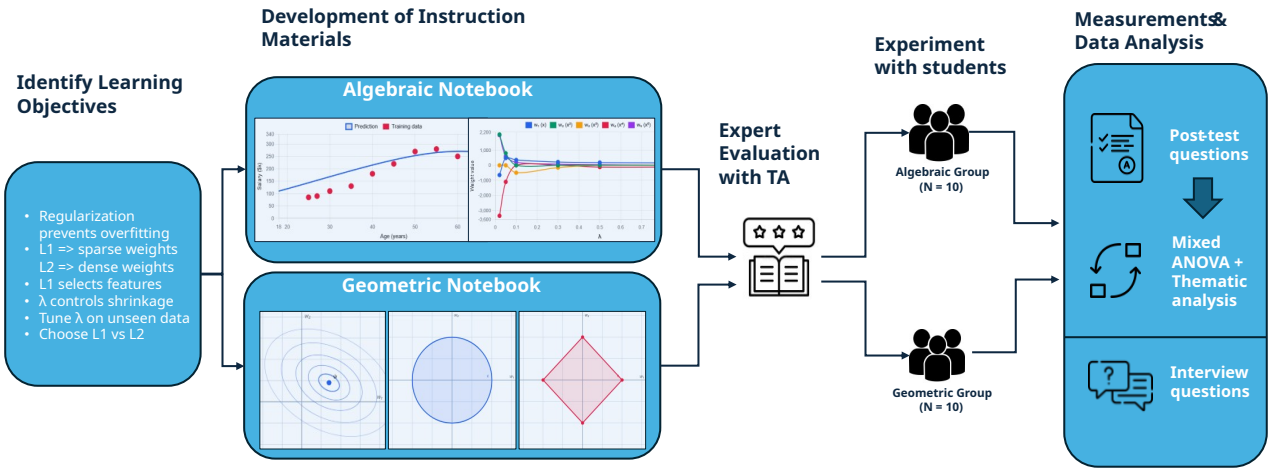


Figure 1: Overview of the study methodology

The group assignment was done using stratified randomization [28] across two variables: field of study, ensuring equal distribution of Computer Science and Aerospace students in each group, and pre-test score, ensuring equivalent prerequisite knowledge. Following the group assignment, the participants worked through the interactive materials and filled a post-test upon completion (see Appendix D). Subsequently, a structured interview was conducted to gather qualitative data on their learning experience and reasoning strategies (see Appendix E).

**Interview.** After the post-test, some of the participants (N = 8) took part in a short structured interview. A post-test score shows how much a student learned and at what level, but not how they reasoned, which parts of the materials helped them, or what still confused them afterwards. The interview was added to capture these aspects, which a graded test cannot. Pairing a test with interviews is a common mixed-methods approach, in which the interview clarifies and adds to what the test measures [29].

The interview had four questions, each aimed at something the post-test does not capture. The first asked which parts of the materials helped the student understand the topic, giving direct feedback on the materials themselves [30]. The second used a retrospective think-aloud approach by asking the student to talk through their reasoning on one of the post-test questions after completing it, so that their thinking, and not only the final answer, could be examined [31]. The third asked whether they felt they were reasoning or guessing, which shows how aware they were of their own understanding [32]. The fourth asked what still felt unclear after the materials, pointing to any gaps that remained [30].

**Measurements.** To measure the knowledge acquired by the students, they were asked to complete a pre-test which measured their prerequisite knowledge of foundational concepts, overfitting, loss functions, regression models, required to engage with the instruction materials and a post-test upon completion. The pre-test served as a baseline to ensure equivalence between the two groups rather than for computing knowledge gain. The post-test was created to match the learning objectives of the instruction materials, and the questions were classified into 3 separate categories, conceptual understanding, problem solving and knowledge transfer using Bloom's taxonomy [33]. Conceptual understanding tested whether students could explain or interpret the core ideas of regularization without solving anything or apply knowledge in a new situation, corresponding to the remember and understand levels of Bloom's taxonomy. Problem-solving questions required students to apply conceptual knowledge to a defined situation with a correct answer, corresponding to the apply and analyse levels. Knowledge transfer questions test whether students can transfer their understanding of the underling principles to a new context, corresponding to the evaluate level. This three-category structure was chosen to assess not only whether students learned from the materials, but whether the two representational formats produce different profiles of understanding, an outcome that a single aggregate score would not be able to detect. The questions were graded according to a rubric with equal weighting

**Data Analysis.** The collected data was analysed using a mixed ANOVA [34], with the representational format, algebraic versus geometric, as the between-subjects factor and question category as the within-subjects factor. A mixed ANOVA was appropriate because the study compared two independent groups while also comparing multiple related post-test category scores from the same participants. It was preferred over separate t-tests per category because it accounts for the within-subject correlation between category scores and reduces the risk of inflated Type I error from multiple separate comparisons. Prior to the main analysis, Shapiro-Wilk tests were applied to verify the assumption of normality. Where the interaction was significant, follow-up independent t-tests were conducted for each category separately, with Bonferroni correction [35] applied. To complement the quantitative findings, the open-ended post-test answers were analysed using inductive thematic analysis [36]. Initial open coding was used to identify recurring ideas in the responses,

which were then grouped into broader themes. This approach was suitable because it allowed reasoning patterns and types of explanations to be identified from the data rather than from predefined categories. The interview responses were analysed using descriptive statistics, which was suitable because the interview sample was small and the goal was to summarize students' reported experiences clearly rather than make statistical generalizations.

# 4 Results

## 4.1 Pre-test Equivalence

Because participants were assigned to groups using stratified randomisation on pre-test score and field of study, group equivalence on prerequisite knowledge was confirmed prior to the main analysis. Both groups achieved identical pre-test performance (algebraic: $M = 4.70$, $SD = 0.48$; geometric: $M = 4.70$, $SD = 0.48$; out of a maximum of 5), with no significant difference between them (Mann–Whitney $U = 50.0$, $p = 1.00$; independent t-test $t(18) = 0.00$, $p = 1.00$). This confirms that the stratification succeeded and that the two groups were equivalent in prerequisite knowledge, supporting the attribution of any post-test differences to the instructional format rather than to prior knowledge. Pre-test scores were near-ceiling across the entire sample (all 20 participants scored 4 or 5 of 5, with 14 achieving full marks), consistent with the pre-test's intended role as a check that participants possessed the required foundational knowledge rather than as a baseline for computing knowledge gain. Self-reported prior exposure to regularization was minimal across the sample: all participants reported no prior experience or only passing familiarity with the terms.

## 4.2 Descriptive statistics

Figure 2 reports mean post-test scores by category and group. The algebraic group outperformed the geometric group overall (algebraic: $M = 26.9$, $SD = 4.1$; geometric: $M = 21.5$, $SD = 5.8$, out of 36). The two groups were closest on problem-solving (algebraic: $M = 10.1$, $SD = 1.5$; geometric: $M = 9.0$, $SD = 2.4$), with a larger separation on knowledge transfer (algebraic: $M = 7.3$, $SD = 1.9$; geometric: $M = 5.5$, $SD = 1.8$). The widest divergence was on conceptual understanding, where the algebraic group scored substantially higher ($M = 9.5$, $SD = 1.7$) than the geometric group ($M = 7.0$, $SD = 2.5$).

## 4.3 Inferential analysis

To determine whether the two formats led to different performance within each question category, scores were analysed with a 2 (representational format: algebraic vs. geometric, between subjects) $\times$ 3 (question category: conceptual, problem-solving, transfer, within subjects) mixed ANOVA, followed by planned independent-samples $t$-tests comparing the groups within each category. Shapiro-Wilk tests supported normality in five of the six group $\times$ category cells; the algebraic transfer cell departed mildly from normality ($W = 0.81$, $p = .02$), so all between-group comparisons were additionally verified with Welch's $t$-test and the Mann-Whitney $U$ test, which yielded the same conclusions throughout. Mauchly's test indicated that sphericity was met ($W = 0.90$, $p = .40$), so uncorrected within-subjects tests are reported.

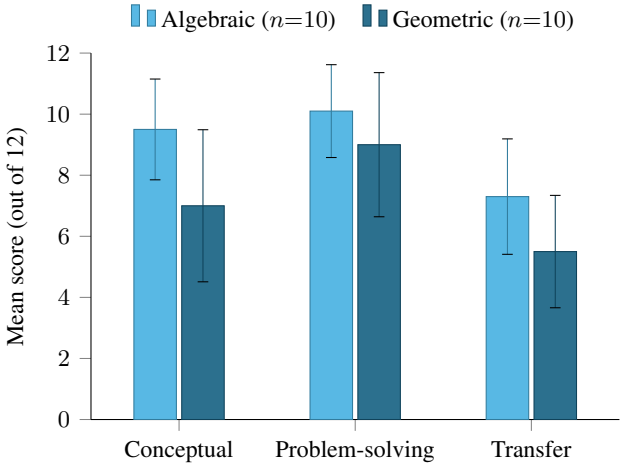


Figure 2: Post-test scores by category and group ($N = 20$). Bars show group means out of 12; error bars indicate $\pm 1$ SD.

The ANOVA showed a significant main effect of representational format, with the algebraic group scoring higher overall ($F(1, 18) = 5.84$, $p = .027$, $\eta_p^2 = .25$), and a significant main effect of question category ($F(2, 36) = 28.10$, $p < .001$, $\eta_p^2 = .61$): collapsing across groups, problem-solving was answered most fully and transfer least. Consistent with this, the overall (36-point) difference between groups was significant with a large effect size ($t(18) = 2.42$, $p = .027$, $d = 1.08$).

The central question-whether one format outperformed the other within each category-was addressed with planned independent-samples $t$-tests, using Bonferroni correction for the three comparisons. The algebraic group scored significantly higher on conceptual understanding, and this difference survived correction with a large effect size ($t(18) = 2.64$, $p_{\text{adj}} = .050$, $d = 1.18$). The two groups did not differ reliably on problem-solving ($t(18) = 1.24$, $p = .23$, $d = 0.55$), and although the algebraic group scored higher on knowledge transfer, this difference was significant only before correction and not after ($t(18) = 2.16$, $p = .045$, $p_{\text{adj}} = .13$, $d = 0.97$). The two formats therefore differed in their effect on conceptual understanding, while the evidence for a difference on transfer was weaker and inconclusive, and no difference was detected on problem-solving.

For completeness, the format $\times$ category interaction was not statistically significant ($F(2, 36) = 1.37$, $p = .27$, $\eta_p^2 = .07$). This means that, although the conceptual difference was the only one to reach significance, the analysis cannot establish that the algebraic advantage was *larger* for conceptual understanding than for the other categories. The conceptual result is therefore best interpreted as the category in which a group difference was most clearly detected, rather than as a statistically isolated effect.

### 4.4 Thematic analysis of open-ended responses

The open-ended post-test responses were analysed using an inductive thematic analysis based on Braun and Clarke's six-phase approach [36]. Codes were created from the responses themselves and then grouped into families covering conceptual reasoning, representational vocabulary, methodological awareness, depth of explanation, and transfer. Because one student could produce several coded segments, both the number of coded segments and the number of distinct students behind them are reported, and all counts are treated as descriptive rather than as statistical tests. Five patterns emerged.

The clearest difference between the two groups was not what students concluded, but the language they used to reason. Algebraic-group students almost always explained regularization as a penalty added to the loss function (penalty framing: 16 segments from 8 of 10 algebraic students, versus 2 segments in the geometric group). Geometric-group students instead explained it through a shrinking constraint region (constraint-region framing: 21 segments from 8 of 10 geometric students, versus none in the algebraic group). This split was almost complete and reflects the different emphases of the two textbooks.

Despite this difference, both groups shared the same core understanding. Reading a large training-validation gap as a sign of overfitting, knowing that too much regularization causes underfitting, and choosing $\lambda$ based on the validation error all appeared at similar rates in both groups (for example, the underfitting-at-large-$\lambda$ code appeared in 20 algebraic versus 19 geometric segments, and the gap-as-overfitting code in 19 versus 18). This agreement on how to apply the ideas fits the quantitative result that the two representations did not differ reliably on problem-solving.

Two other patterns were stronger in one group than the other. First, algebraic students were much more likely to suggest checking a result before trusting it, for example by looking at the validation error or trying several values of $\lambda$ before deciding. This appeared in 20 segments from 8 algebraic students, but only once in the geometric group. Algebraic students therefore tended to test their claims rather than judge them by inspection alone. Second, geometric students more often gave short answers that stated an outcome without explaining why it happened. This minimal-mechanism code appeared in 14 segments from 5 geometric students and in none of the algebraic responses, and a few geometric answers also confused overfitting with underfitting (6 segments from 5 geometric students). These were minority patterns rather than features of the whole group, and most geometric students still gave full and correct constraint-based explanations.

The fifth pattern came from the housing-feature transfer item from the post-test (Q9), which asked students to evaluate a colleague's claim about a Lasso model dropping one of two correlated features. Here the quality of reasoning varied within both groups rather than between them. Many students correctly noticed that the two features, the number of bedrooms and the total number of rooms, were probably correlated, and several suggested Ridge regression as a compromise that shrinks both coefficients instead of dropping one (correlation identified: 7 algebraic and 5 geometric students; alternative regularizer suggested: 6 and 5 students respectively). Other students assumed that a coefficient set to zero by Lasso meant the feature was truly unimportant, when it was more likely just redundant given the other feature (3 algebraic and 5 geometric students). Because both correct and mistaken reasoning appeared in each group, this is better seen as a general feature of how students handled an unfamiliar transfer task than as a difference between the two representations.

### 4.5 Structured interview insights

Eight of the twenty participants took part in a short structured interview after the post-test (geometric: N = 5; algebraic: N = 3). Because the interview sample was small and unbalanced, the responses are summarised descriptively and no statistical comparison between groups is made.

The interviews showed that students found the interactive widgets, coefficient table, and summary table most helpful. These parts appeared to make regularization more concrete by showing how changes in $\lambda$ affected the model and its coefficients. Students also used the tables to compare L1 and L2 regularization more directly, especially when explaining differences in coefficient shrinkage.

The think-aloud responses showed similar reasoning patterns across the two groups. Students usually identified whether the model was overfitting or underfitting, connected this to the value of $\lambda$, and then suggested adjusting $\lambda$ or checking validation performance. In the L2 error question, most students interpreted the increase in both training and test error as a sign that the regularization strength was too high, rather than as evidence that L2 regularization itself was unsuitable. This supports the post-test result that the two groups did not differ clearly on problem-solving. Although the two groups often used different representational language, their practical reasoning steps were largely similar.

The final interview responses showed that most students did not experience the post-test as pure guessing, and the remaining unclear points were generally limited. Several comments referred to issues that went beyond the direct scope of the assessment, such as needing more time to internalize the topic, wanting to know more about how $\lambda$ is chosen in practice, or wondering whether validation-based tuning can itself lead to overfitting. At the same time, a few students still found the exact reasoning behind sparse versus dense coefficient behaviour difficult to explain. This was most explicit in the algebraic interviews, where students could state the difference between absolute-value and squared penalties but were less certain why this produced different coefficient behaviour. Overall, the interview responses suggest that most core ideas were understood, while the remaining uncertainties were either minor, more technical, or related to extensions beyond the assessed tasks.

## 5 Discussion and Limitations

### 5.1 Discussion

This study compared algebraic and geometric representations for teaching L1 and L2 regularization. The algebraic group performed better overall, with the clearest statistically supported difference in conceptual understanding; the groups did

not differ reliably on problem-solving, and the algebraic advantage on transfer did not survive correction. Two features of the design temper how strongly this can be read. The format $\times$ category interaction was not significant, so the conceptual advantage cannot be claimed as specific to that category rather than a generally higher level of performance. More importantly, the qualitative data show that the two groups reached largely the same understanding and differed mainly in how they expressed it, which suggests that the measured gap reflects the language of the assessment as much as the depth of learning.

The most likely source of this gap is the alignment between the algebraic notebook, the post-test, and standard machine learning terminology. The conceptual questions asked students to reason about overfitting, generalization, model weights, penalty strength, training and test error, and the effect of increasing $\lambda$ - ideas expressed directly in the algebraic notebook, where regularization was introduced as a penalty added to the loss function. Algebraic-group students could therefore answer in much the same vocabulary they had just been taught. This mirrors Koedinger and Nathan's finding that a representation's apparent advantage often tracks how familiar it is to the learner rather than any inherent superiority [20].

The geometric notebook taught the same concepts through a different language: weight space, loss contours, constraint regions, and the shapes of the L1 and L2 boundaries. Although the rubric accepted constraint-based explanations, geometric students still had to translate this visual understanding into the terms the questions used - connecting a smaller constraint region to stronger regularization, movement toward the origin to weight shrinkage, and a restricted solution space to lower model flexibility. The difficulty appears to lie in this translation rather than in the representation itself, echoing Sierpinska's argument that many student difficulties stem not from any single representation but from moving flexibly between them [22].

The qualitative results support this reading. The main difference between groups was not what they concluded but the vocabulary they reasoned in: algebraic students framed regularization as a penalty on the loss function, geometric students as a constraint region in weight space, while their practical judgements - reading a large train-validation gap as overfitting, treating too much regularization as a cause of underfitting, choosing $\lambda$ from validation error - appeared at similar rates in both. One asymmetry is worth noting: algebraic students were more likely to justify a claim by checking validation error or trying several values of $\lambda$, suggesting the algebraic framing encouraged slightly more explicit reasoning about model evaluation.

The geometric format communicated its representation successfully: students readily used the intended visual concepts - circles, diamonds, boundaries, constraint regions. But this vocabulary did not translate into higher scores; some geometric answers stated the correct outcome without explaining the mechanism in terms of overfitting, generalization, or validation performance. The representation shaped how students reasoned without being the form this assessment most rewarded.

The absence of a reliable difference on problem-solving is the clearest positive finding. Both groups generally understood that $\lambda = 0$ leaves a model unregularized and prone to overfitting while a large $\lambda$ over-regularizes and underfits, and both could act on common patterns such as a large train-test gap. On the reasoning that matters most in practice, the two representations were interchangeable; they differed in how students explained that reasoning, not in whether they could carry it out.

The transfer results are the least conclusive. The algebraic group scored higher on average, but not reliably so after correction, and the Lasso house-price item was hard for both groups. Many students correctly saw that "number of bedrooms" and "total number of rooms" are likely correlated and that Lasso may drop one because the other already carries its information; others read a zero coefficient as evidence the feature was unimportant. Students seem to learn that L1 performs feature selection without fully grasping that the selection depends on the other features present - a gap that neither representation closed.

Taken together, the findings suggest that representational choice shapes how students explain regularization more than whether they understand it. The interviews also point to a difference the post-test did not capture: algebraic students could state the difference between the absolute-value and squared penalties but were often unsure why this leads to sparse rather than smoothly shrunk coefficients, whereas the geometric format may have given students a clearer sense of why L1 produces sparse weights and L2 smooth shrinkage. Because this aspect was not assessed directly, combining the two formats is worth investigating as a way to build stronger understanding, consistent with multimedia-learning accounts in which integrated verbal and visual representations support deeper understanding than either alone [19].

## 5.2 Limitations

This study has several limitations. First, the sample was small, with 20 participants divided into two groups of 10. This limits statistical power and makes the results sensitive to individual differences. Some effects, especially the transfer difference, may not have reached significance because the study was underpowered.

Second, participants were Computer Science and Aerospace Engineering students who had already completed an introductory machine learning course. Their near-ceiling pre-test scores show that they had the required background knowledge, but the findings may not generalize to students with less prior experience.

Third, the assessment may have aligned more naturally with algebraic and general ML terminology. The rubric was designed to be representation-neutral, but the questions still required students to explain overfitting, generalization, weight shrinkage, model complexity, and $\lambda$. These concepts are expressed more directly in the algebraic notebook, whereas geometric students may have needed to translate from constraint-region reasoning into this terminology.

Fourth, the instructional materials were reviewed by only one teaching assistant. This helped check correctness and clarity, but stronger validation would require multiple experts.

In addition, both the scoring rubric and the thematic analysis would benefit from reliability checks. For example, using multiple raters and reporting inter-rater reliability would make the scoring more robust, while intra-rater reliability checks would help confirm that the same researcher applied the rubric and qualitative codes consistently over time.

Finally, the study measured only immediate post-test performance. It does not show whether either representation leads to better long-term retention or delayed transfer. Future work should use larger and more diverse samples, include a delayed post-test, and test whether combining algebraic and geometric explanations produces stronger learning than either format alone.

# 6 Responsible Research

## 6.1 Ethical Considerations

This study involved human participants and was therefore conducted under the oversight of the TU Delft Human Research Ethics Committee (HREC), whose approval, together with the accompanying data management plan, was obtained before any data was collected. Participation was voluntary throughout. Before taking part, each participant received and was taken through an informed consent form describing the purpose of the study, the procedure (pre-test, instructional notebook, post-test, and optional interview), the data that would be collected, and how it would be stored and used. Participants were told that they could withdraw at any point without giving a reason and without consequence, and only those who gave explicit consent took part.

The study posed minimal risk. It collected no special-category personal data; the only background information recorded was field of study, needed for the stratified group assignment, and a self-report of prior exposure to regularization. Responses to the pre-test, post-test, and interviews were stored under participant codes rather than names, so that no answer could be traced back to an individual, and the interview data was kept in anonymised form. All data was held on secured TU Delft infrastructure(OneDrive) during the project. Because the topic was purely educational and the participants were adults completing tasks similar to ordinary coursework, no foreseeable physical, psychological, or reputational harm was associated with participation.

## 6.2 Reproducibility

The methodology was designed so that the study can be reproduced independently. Section 3 describes the full procedure, participant selection, stratified randomization, the instructional phase, and the post-test and interview, in enough detail to be followed step by step. All materials required to repeat the study are provided: the two interactive Jupyter books, the pre-test, the post-test, the expert-evaluation (reviewer) questions, the grading rubric, the thematic-analysis codebook, and the interview questions. The notebooks are released openly, and the remaining instruments are included in the appendices or in the research project's repository, so that another researcher could deliver the same instruction and apply the same assessment.

To support reanalysis as well as replication, the anonymised data - pre-test and post-test scores, category-level results, and the coded qualitative segments - has been deposited in the 4TU Research Data repository. The statistical procedures (the mixed ANOVA, the normality and sphericity checks, and the Bonferroni-corrected follow-up tests) and the inductive thematic-analysis approach are standard and fully specified, so the quantitative pipeline can be re-run on the shared data and the qualitative analysis retraced through the codebook. One limit to exact reproducibility should be acknowledged: the qualitative coding and rubric scoring were carried out by a single researcher, so some interpretive judgement is unavoidable. The published codebook and rubric make these judgements transparent and repeatable, but full replication of the coding would benefit from the inter- and intra-rater reliability checks noted in the limitations.

## 6.3 Use of Generative AI

Generative AI tools (Claude Opus 4.8) were used in a supporting role in this project, with the author retaining full responsibility for the final content. They were used mainly to improve grammar and help structure the writing, to generate the code for the interactive widgets in the two notebooks, and to brainstorm candidate ideas for some of the pre- and post-test questions. All such output was reviewed, verified, and edited by the author, and no personal participant data was entered into these tools.

# 7 Conclusion and Future Work

This study asked whether teaching L1 and L2 regularization algebraically or geometrically changes what students learn, and in which kinds of understanding. To address this, 20 students were randomly assigned to one of two interactive notebooks, identical in content and structure except for whether regularization was framed algebraically or geometrically. After a pre-test confirmed the two groups were equivalent, they completed a nine-question post-test spanning conceptual understanding, problem-solving, and knowledge transfer, with responses analysed quantitatively through a mixed ANOVA and qualitatively through thematic analysis of free-text answers and structured interviews. The answer is that both formats work: students in both groups reasoned soundly about the core ideas - why $\lambda = 0$ overfits, why a large $\lambda$ underfits, and why increasing $\lambda$ shrinks weights - and the two representations were effectively interchangeable on the practical, problem-solving reasoning that matters most in use.

Where they differed was narrower than a single overall score suggests. The algebraic group scored higher, most clearly on conceptual explanation, but the qualitative evidence indicates the two groups understood the concept similarly and diverged mainly in vocabulary - a penalty on the loss function versus a constraint region in weight space. The algebraic advantage therefore appears to owe much to the assessment sharing its language, rather than to a deeper grasp of regularization.

The contribution of this work is an experimental comparison of two standard ways of teaching regularization, with the finding that representation shapes how students express

understanding more than whether they reach it. For teaching practice this argues against choosing one format over the other: the algebraic view anchors the standard terminology of penalties, weights, and $\lambda$, while the geometric view makes the origin of L1 sparsity and L2 shrinkage visible, and the two are most useful together.

These conclusions rest on a small sample, an immediate post-test, and an assessment whose wording favoured the algebraic format, so they should be read as indicative rather than definitive. The clearest next steps are to replicate with more, and more varied, students; to add a delayed post-test for retention; to strengthen scoring and coding with multiple raters and reliability checks; and, most directly, to test whether teaching both representations together outperforms either alone.

## References


[1] Amit Verma, Kamal Lamsal, and Payal Verma. An investigation of skill requirements in artificial intelligence and machine learning job advertisements. 36(1):63–73, 2022.

[2] Elmedin Burnik. Competences in machine learning: The order of competences that students need to learn in ml, 2022.

[3] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[4] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition, 2009.

[5] Robert Tibshirani. Regression shrinkage and selection via the lasso. 58(1):267–288, 1996.

[6] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC, 2015.

[7] Shaaron Ainsworth. Deft: A conceptual framework for considering learning with multiple representations. 16(3):183–198, 2006.

[8] Richard E. Mayer. Multimedia learning. In *Psychology of Learning and Motivation*, volume 41, pages 85–139. Academic Press, 2002.

[9] R. Benjamin Shapiro, Rebecca Fiebrink, and Peter Norvig. How machine learning impacts the undergraduate computing curriculum. *Communications of the ACM*, 61(11):27–29, 2018.

[10] R. Benjamin Shapiro and Rebecca Fiebrink. Introduction to the special section: Launching an agenda for research on learning machine learning. *ACM Transactions on Computing Education*, 19(4):1–6, 2019.

[11] Elisabeth Sulmont, Elizabeth Patitsas, and Steve Cooper. Can you teach me to machine learn? In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, pages 948–954. ACM, 2019.

[12] Franz Yannik Fleischer, Rolf Biehler, and Carsten Schulte. Teaching and learning data-driven machine learning with educationally designed jupyter notebooks. 21(2):7, 2022.

[13] Christiane Gresse von Wangenheim, Jean C. R. Hauck, Fernando S. Pacheco, and Matheus F. Bertonceli Bueno. Visual tools for teaching machine learning in k-12: A ten-year systematic mapping. 26:5733–5778, 2021.

[14] I. Renea, Gosia Migut, and Jesse Krijthe. Are interactive visualizations in machine learning education helping students? In *Proceedings of the 30th ACM Conference on Innovation and Technology in Computer Science Education*. ACM, 2025.

[15] Matias Valdenegro-Toro and Matthia Sabatelli. Machine learning students overfit to overfitting. In *Proceedings of the Third Teaching Machine Learning and Artificial Intelligence Workshop*, volume 207 of *Proceedings of Machine Learning Research*, pages 25–33. PMLR, 2023.

[16] I. Angelescu. Discovering the misconceptions that influence learning of machine learning, 2022.

[17] Karen Jin. Students' misconceptions of gradient descent algorithm in an machine learning course. 34(6):150–151, 2019.

[18] Allan Paivio. *Mental Representations: A Dual Coding Approach*. Oxford University Press, New York, 1986.

[19] Richard E. Mayer. Cognitive theory of multimedia learning. In Richard E. Mayer, editor, *The Cambridge Handbook of Multimedia Learning*, pages 31–48. Cambridge University Press, Cambridge, 2005.

[20] Kenneth R. Koedinger and Mitchell J. Nathan. The real story behind story problems: Effects of representations on quantitative reasoning. *Journal of the Learning Sciences*, 13(2):129–164, 2004.

[21] Kenneth R. Koedinger, Martha W. Alibali, and Mitchell J. Nathan. Trade-offs between grounded and abstract representations: Evidence from algebra problem solving. *Cognitive Science*, 32(2):366–397, 2008.

[22] Anna Sierpinska. On some aspects of students' thinking in linear algebra. In Jean-Luc Dorier, editor, *On the Teaching of Linear Algebra*, pages 209–246. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000.

[23] Hamide Dogan-Dunlap. Linear algebra students' modes of reasoning: Geometric representations. *Linear Algebra and its Applications*, 432(8):2141–2159, 2010.

[24] Jean-Luc Dorier and Anna Sierpinska. Research into the teaching and learning of linear algebra. In Derek Holton, editor, *The Teaching and Learning of Mathematics at University Level: An ICMI Study*, pages 255–274. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001.

[25] John C. Hull. *Machine Learning in Business: An Introduction to the World of Data Science*. Independently published, 3 edition, 2021.

[26] Rudolf Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, 1948.

[27] J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Research Branch Report 8-75, Naval Technical Training Command, Millington, TN, 1975.

[28] Walter N. Kernan, Concato Viscoli, Robert W. Makuch, Lawrence M. Brass, and Ralph I. Horwitz. Stratified randomization for clinical trials. *Journal of Clinical Epidemiology*, 52(1):19–26, 1999.

[29] Jennifer C. Greene, Valerie J. Caracelli, and Wendy F. Graham. Toward a conceptual framework for mixed-method evaluation designs. 11(3):255–274, 1989.

[30] Paul Black and Dylan Wiliam. Assessment and classroom learning. 5(1):7–74, 1998.

[31] K. Anders Ericsson and Herbert A. Simon. Verbal reports as data. 87(3):215–251, 1980. Revised edition published as *Protocol Analysis: Verbal Reports as Data*, MIT Press, 1993.

[32] John H. Flavell. Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. 34(10):906–911, 1979.

[33] Benjamin S. Bloom, editor. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain*. David McKay, New York, 1956.

[34] Sohad Murrar and Markus Brauer. Mixed model analysis of variance. *Journal of Consumer Psychology*, 28(2):312–330, 2018.

[35] Carlo Emilio Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.

[36] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

# A Learning Objectives

The following learning objectives were identified as the shared set of concepts that both notebooks were designed to teach:

- **Why we use regularization** - it prevents overfitting by discouraging the model from using large weights.
- **How to recognize overfitting** - a model that scores much higher on training data than on test data is overfitting.
- **The difference between L1 and L2** - L1 (Lasso) sets some weights to exactly zero; L2 (Ridge) makes all weights smaller but keeps them.
- **Why L1 is useful** - by zeroing out weights, it automatically selects the features that matter and drops the ones that don't.
- **What $\lambda$ controls** - larger $\lambda$ means smaller weights; too small and the model overfits, too large and it underfits.
- **How to choose $\lambda$** - test several values and pick the one that performs best on data the model hasn't seen.
- **Choosing the right method** - know when to use L1 vs. L2, and interpret the results carefully when features are correlated.

# B Reviewer Questions

The reviewer was asked to assess the instructional notebook against the following five criteria.

1. **Correctness.** Is the content of this notebook accurate throughout? Please note any errors, imprecise statements, or misleading explanations you found.
2. **Clarity and level.** Are the explanations clear and pitched appropriately for a student who has completed one introductory ML course-neither assuming too much nor over-explaining? Note anything you expect would confuse a student.
3. **Visualizations.** Do the interactive visualizations work as intended, and do they add to understanding beyond the text? Note any that are unclear, distracting, or not helpful.
4. **Alignment with assessment.** Does the notebook give a student enough to answer the post-test questions correctly? Flag any question that the materials do not adequately prepare a student for.
5. **Overall readiness.** Are these materials good enough to use in a study measuring student learning? What are the most important changes, if any, you would recommend?

# C Pretest Questions

The following questions were administered before the tutorial to assess participants' prior knowledge. The **correct answer is shown in bold**. The final item is a self-report survey question with no correct answer.

1. A model predicts $\hat{y} = a + w_1 x_1 + w_2 x_2$. If $a = 5$, $w_1 = 2$, $w_2 = -3$, and the input is $x_1 = 4$, $x_2 = 1$, what is $\hat{y}$?
   (a) 8
   (b) **10**
   (c) 16
   (d) 3
2. The true values are $y = [10, 20]$ and the predicted values are $\hat{y} = [12, 16]$. What is the mean squared error?
   (a) 1
   (b) 3
   (c) **10**
   (d) 20
3. A model gets near-perfect accuracy on its training data but performs badly on new, unseen data. This is best described as:
   (a) Underfitting
   (b) **Overfitting**
   (c) Good generalization
   (d) Low variance
4. Given weights $w = [3, -4]$, what are the sum of *squared* weights and the sum of *absolute* weights, respectively?
   (a) **25 and 7**
   (b) 7 and 25
   (c) 49 and 1
   (d) 25 and 1
5. Why keep weights small? Which statement is most accurate?
   (a) Large weights make a model simpler and less likely to overfit
   (b) **Large weights tend to make a model more complex and more likely to overfit**
   (c) The size of the weights has no effect on overfitting
   (d) Small weights always reduce training accuracy to zero
6. Before this tutorial, how much experience have you had with regularization (L1/Lasso, L2/Ridge)? *(Survey item - no correct answer.)*
   (a) None - I've never encountered it
   (b) I've heard the terms but haven't used them
   (c) I understand the basic idea but haven't applied it
   (d) I've used it in practice (e.g. in scikit-learn or similar)

# D Post-test questions

## D.1 Conceptual understanding

1. What is the purpose of regularization in machine learning?
2. How do L1 and L2 regularization differ in their effect on model weights?
3. As the regularization strength $\lambda$ increases, what generally happens to the model weights?

### D.2 Problem-solving

4. A student increases $\lambda$ in an interactive regularization widget. The training error increases slightly, but the test error decreases. What does this suggest about the model, and why might this pattern indicate improved generalization?
5. A model trained without regularization achieves 98 percent training accuracy and 61percent test accuracy. A colleague suggests increasing $\lambda$ from zero. Would you agree? Explain what the large gap between training and test accuracy suggests, and what you would expect to happen to both training accuracy and test accuracy as $\lambda$ increases.
6. You train four models with different regularization strengths. Which $\lambda$ would you choose, and why? Explain what is happening when $\lambda = 0$ and when $\lambda = 10$.

| λ | Training error | Validation error |
|---|---|---|
| 0 | 0.02 | 0.30 |
| 0.01 | 0.05 | 0.20 |
| 0.1 | 0.10 | 0.12 |
| 10 | 0.35 | 0.38 |

### D.3 Knowledge transfer

7. A colleague applies L1 regularization and is pleased that it has set 80percent of the model weights to exactly zero, telling you the model is now "much simpler and therefore more accurate." Do you agree with their reasoning? Explain.
8. After adding L2 regularization to a model, a colleague finds that both the training error and the test error have increased compared to the unregularized model. They conclude that regularization "doesn't work for this dataset." Do you agree? Explain.
9. You train a Lasso model to predict house prices. Of two features-'number of bedrooms" and 'total number of rooms"-Lasso keeps one and sets the other to exactly zero. A colleague worries the model is now ignoring important information about the house. How would you respond?

## E Structured interview questions

The structured interview consisted of four open-ended questions asked after the post-test.

1. Was there a specific part of the materials that helped you understand regularization better, and what did it make clearer for you?
2. Think about the question where a colleague added L2 regularization and found that both the training and test error increased. Walk me through how you thought about that one: what did you consider first?
3. Were there any questions where you felt you were guessing rather than reasoning? What made those harder?
4. Is there anything about regularization that you still feel unclear on after going through the materials?