

# Mathematics for Machine Learning: A Bridge Course

Samuel Deng

Columbia University

New York, New York, United States of America

samdeng@cs.columbia.edu

## Abstract

We present *Mathematics for Machine Learning*, a one-semester mathematics course designed to strengthen students' foundations before further study and research in machine learning (ML) and data science. Oftentimes, the mathematical prerequisites needed for serious study of ML are taught in a disjointed manner. Our course is designed to bridge this gap and provide emphasis on concepts heavily employed in modern ML, such as spectral analysis in linear algebra or convex optimization in calculus. We structured our course around the three "pillars" of math that underlie much of modern ML: (i) linear algebra, (ii) calculus and optimization, and (iii) probability and statistics. Weaving each of these together is a central story — all concepts, ideas, and proofs are introduced relative to two ubiquitous concepts in machine learning: least squares regression and gradient descent, providing a consistent anchoring narrative and constant motivation for mathematical ideas.

## ACM Reference Format:

Samuel Deng. 2025. Mathematics for Machine Learning: A Bridge Course. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 2 (SIGCSE TS 2025)*, February 26 – March 1, 2025, Pittsburgh, PA, USA. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/3641555.3705218>

## 1 Introduction

In this poster, we outline *Mathematics for Machine Learning (M4ML)*, a one-semester course with the goal of strengthening students' mathematical foundations for rigorous study of machine learning and data science at the advanced undergraduate or graduate level. This course was piloted in an initial six week iteration in the summer of 2024 for undergraduates and masters students at Columbia University, with very positive feedback (see Section 4).

The goal of this course is to give students the preparation and confidence to pursue machine learning at our institution by addressing a root issue: lack of sufficient mathematical preparation. A prevalent issue in upper-division machine learning courses in our department, felt by both instructors and students, is that, even if they have formally completed the prerequisites, students coming into such courses lack sufficient foundations in the three mathematical "pillars" of machine learning: (i) linear algebra, (ii) multivariable calculus, and (iii) probability and statistics. This is particularly felt in our core machine learning course, where there is a steep jump

in mathematical maturity. This course remedies this issue by providing a second exposure to these prerequisites, focusing on each of these mathematical "pillars" with a view towards developing fluency in the topics that matter specifically for machine learning.

The course is aimed at advancing undergraduates who have already completed (but not necessarily mastered) undergraduate level linear algebra, probability and statistics, and multivariable calculus courses. Optional (but recommended) prerequisites include an introductory programming course (for light programming problems in Python) and a first course in reading and writing proofs (at the level of, say, discrete mathematics).

## 2 Main Issues

There are several main issues we see our course addressing. Although we found these as issues at our institution, we anticipate the same issues cropping up in other computer science departments.

**Variance in prerequisite courses.** Our department allows students to take multivariable calculus, linear algebra, and probability and statistics in different departments, with different professors and syllabi. This introduces much variance in what students might learn in these courses, and knowledge gaps in any of these three areas can lead to downstream confusion. This course aims to have a consistent syllabus that patches up such potential holes.

**Insufficient focus on core ML concepts.** Due to being "of-flooded" to different departments, the prerequisite courses may not focus on specific techniques and concepts that have particular importance in machine learning. For example, a multivariable calculus course may focus on line and path integrals while a student interested in machine learning may need more time learning about optimization and the implications of convexity. This course aims to focus on the core topics from each prerequisite in greater detail than their coverage in prerequisite undergraduate courses.

**Lack of motivation for theory.** Prerequisite undergraduate math courses often present theory without sufficient motivation in why a student might need that theory, especially with an eye towards more applied domains. Our course revolves around two main concepts in machine learning (least squares regression and gradient descent) that are motivated early on as realistic and ubiquitous throughout the field of machine learning. With an eye towards understanding these two concepts with as much rigor and depth as possible, we develop the theory *in order to* deepen our understanding of regression and gradient descent, which gives the course a coherent narrative and scaffolding, all while motivating each new piece of theory we introduce. And, because regression and gradient descent are mathematically deep and interesting in their own right, we are able to flesh out many different perspectives of least squares and gradient descent as students acquire new mathematical tools.

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SIGCSE TS 2025, February 26-March 1, 2025, Pittsburgh, PA, USA

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0532-8/25/02

<https://doi.org/10.1145/3641555.3705218>

### 3 Course Overview

We now provide a brief overview of our course, which will be fleshed out more fully in the poster when we may share the full course webpage, graphics, and materials. In Sections 3.2, 3.3, and 3.4, we present the *modules* of the course, where each module corresponds, roughly, to three hours of lecture.

#### 3.1 Course “Story”

We designed our course to have an overarching story. The course is structured around two main ideas that underlie modern machine learning: least squares regression and gradient descent. Very informally, least squares regression is a classic way of modeling problems in machine learning (the “what”), and gradient descent is the workhorse algorithm that drives much of modern machine learning (the “how”). Every week, we develop and motivate these two ideas in lecture with the tools and concepts students learn from each part of the course. As the class goes on, students develop different perspectives on these two ideas from each “pillar” of mathematics. The hope is that, by the end of the course, students develop a deep understanding of both these ideas in ML while also having two concrete “applications” to motivate all the abstract mathematical tools and concepts they learn in the course.

To emphasize the development of these parallel “stories,” we centered each module around accessible 3D renderings for the students to interact with that guide the development of each lecture. Due to space restrictions, examples will be shown in the full poster, along with an extended description of each module.

#### 3.2 Part 1: Linear Algebra

The course begins with a tour of linear algebra with specific focus on building geometric intuition with least squares regression as analogous to projection onto a subspace spanned by data. In **Module 1 (vectors, matrices, and least squares)** and **Module 2 (bases subspaces, and orthogonality)**, students learn how data is represented in matrices and vectors in machine learning, and they culminate in a geometric proof from purely linear algebraic principles of ordinary least squares. **Module 3 (singular value decomposition)** gives students the tool of the *pseudoinverse* to get a more general solution to ordinary least squares when the number of equations is less than or greater than the number of unknowns. **Module 4 (Eigenvalues and eigenvectors)** introduce diagonalization and the all-important positive semidefinite matrix, motivating quadratic forms and demonstrating that the least squares objective, all along, was a quadratic form. We hint at optimizing such objects in rough analogy to optimizing quadratics in single-variable calculus, which leads smoothly to Part 2: Calculus and Optimization.

#### 3.3 Part 2: Calculus and Optimization

Picking up from the understanding that the least squares objective can be seen as a positive semidefinite quadratic form, **Module 5 (differentiation and vector calculus)** reviews multivariable differentiation and arrives at the same ordinary least squares solution as in Part 1 but from an entirely different perspective and proof: optimization of a quadratic function. **Module 6 (Taylor series and linearization)** re-introduces the Taylor series and multivariable calculus as the art of taking linear or quadratic approximations of

possibly unwieldy functions, and this motivates and allows us to introduce gradient descent, the second parallel story of the course. After using Taylor series to prove a basic local convergence theorem for gradient descent, **Module 7 (Optimization and the Lagrangian)** tours local optimization and constrained optimization, bolstering the regression story with constrained optimization, which naturally leads to the discussion of regularization and ridge regression. **Module 8 (Convex Optimization)** brings the two stories together — students now have the machinery to prove that least squares regression admits a *convex function*, so applying gradient descent to solve the problem will provably converge to a global optimum. This gives a third, iterative solution to least squares regression, bringing our course full circle.

#### 3.4 Part 3: Probability and Statistics

Finally, we motivate the final third of the course on probability and statistics in **Module 9 (Probability Theory, Models, and Data)** by grounding the epistemological assumptions of machine learning in probability theory, with an eye towards analyzing the statistical properties of least squares. The language of probability allows us to posit a statistical error model, and we prove that, under this model, least squares’ conditional expectation is the true parameter vector, and its conditional expectation is the a function of the noise variance. **Module 10 (Law of large numbers and statistical estimation)** reintroduces the law of large numbers and the notion of a *statistical estimator*, and students see the *Gauss-Markov Theorem* — that least squares is also, in a formal sense, an *optimal* unbiased estimator. The notion of statistical estimation also gives us language to talk about *stochastic gradient descent*, the main workhorse algorithm that powers most of modern machine learning. **Module 11 (Central Limit Theorem, Distributions, and MLE)** and **Module 12 (Multivariate Gaussian Distribution)** finally connect least squares back to perhaps the most familiar statistical distribution, the Gaussian, by showing that it arises both from the statistical paradigm of maximum likelihood estimation with Gaussianity assumptions and as a sampling distribution.

### 4 Initial Student Experiences

We piloted this course in the summer of 2024 in a six-week format, and we received overwhelmingly positive evaluations and feedback, albeit from a small sample of 14 students. At the end of the course, students received an anonymous evaluation survey, where questions had categorical answer choices from (1) Poor, (2) Fair, (3) Good, (4) Very Good, and (5) Excellent. The two relevant questions were *Amount Learned* and *Overall Course Quality*, which both received 8 votes of (5) Excellent and 1 vote of (4) Very Good from 9 total responders.

Additional student evaluations and feedback will be present in the poster, when more space is permitted.

### 5 Acknowledgments

The author acknowledges support from the Columbia CS department’s SEAS Teaching Fellowship for financial support during the pilot version of this course. The author also appreciates valuable discussions with Daniel Hsu, Tony Dear, Jae Woo Lee, Nakul Verma, and Adam Cannon in assisting the creation of this course.