My experiences in taking the course 6.437

Zhengqi Gao
MIT EECS
May 23, 2023

This semester, I enrolled in the course 6.437 (new number: 6.7800), Inference and Information, at MIT. Undoubtedly, this course has been the most challenging course I have ever taken in my life; it definitely warrants a dedicated article from multiple perspectives. For those unfamiliar, MIT EECS Ph.D. program imposes a Technical Qualifying Examination (TQE) requirement. As part of this requirement, Ph.D. students must complete four courses from a predetermined list, earning a minimum of three A's and one B. While I have already secured three A's, I found myself pondering at the beginning of this semester: Why not take the 'notoriously' hard 6.437? Well, it turns out this thought put me onto an arduous, challenging, yet highly-awarding journey. I spent a tremendous amount of time on this course and gained really a lot from the course. **This article is designed for those hesitating to take 6.437 or not, and those who are interested in what a hardcore course inside MIT looks like.**

First, let me provide some background on the course and myself. In my own words, this course primarily focuses on statistical inference from an abstract and theoretical standpoint. I took the course in spring 2023, but our syllabus is similar to the one used in spring 2022, which can be found at this link. The first half of the semester covers topics such as hypothesis testing, inference, modeling, and information geometry. The second half delves into the asymptotic properties of the concepts taught in the first half, examining how inference, modeling, and decision errors behave as the number of samples (N) approaches infinity. The course has a broad scope, providing a comprehensive analysis of these topics. Personally, I have built my statistics background through self-learning from the book 'Pattern Recognition and Machine Learning' and by writing its solution manual. I have some prior experience in real analysis. I do love math and love exploring the theoretical perspectives of an algorithm, but I won't say I am good at it.

The course instructors, Professors Wornell and Golland, are excellent. They provide thorough explanations and conduct whiteboard derivations, which greatly enhance understanding. The lectures take place on Tuesdays and Thursdays, and I typically review the corresponding lecture notes immediately after each session, usually taking me around an hour to complete. Every week, we are assigned a problem set (Pset) that covers the material from both lectures. Each Pset consists of approximately four problems, with each problem having around five sub-questions. I dedicate my evenings to working on the Pset, aiming to solve one problem per night. While I manage to achieve this goal sometimes, there are occasions when I struggle. The Pset problems are challenging, and here's the twist: even if you grasp all the concepts covered in the lecture notes, it is highly likely that you may still find the Pset difficult to solve. When doing the Pset problem, you need to read a lengthy question with tons of symbols and first spend some time to understand what the symbols really tell you and what the question is actually asking, and then try to derive something on scratch paper. Essentially, the Pset presents a challenge in identifying the connection between the posed question and the tools we have learned in class. Sometimes, it even directly uses an extension of what is taught in class. But understanding (strictly proving) why the extension is correct will take you some time. As an example, which is used most frequently, we learn the I-projection of an exponential family $\mathcal{E} = \{p \in \mathcal{P}^{\mathcal{Y}} : p(y; x) = \exp[xt(y) - \alpha x + \beta(y)]\}$ onto a linear family $\mathcal{L} = \{p \in \mathcal{P}^{\mathcal{Y}} : \mathbb{E}[t(y)] = \bar{t}\}$. In the Pset, after we do some simplifications, the problem could be converted to the I-projection of $\mathcal{E}$ onto $\mathcal{S} = \{p \in \mathcal{P}^{\mathcal{Y}} : \mathbb{E}[t(y)] \leq \bar{t}\}$. But we didn't learn '$\leq$', we only learn '$=$' as in the linear family! Well... you need to prove by yourself that the KL divergence is convex, so actually I-projection onto $\mathcal{S}$ is equivalent to its boundary, which is actually $\mathcal{L}$.

The midterm and final exams follow a similar format to the Pset. We are given three hours to complete three problems, each consisting of around four sub-questions. The total score for each exam is 30 points, with the average typically falling below 20 points. It may seem surprising, but even MIT students often struggle to achieve a score of 66%. It's worth noting that the grading rubric is quite lenient. For instance, in my midterm, I was only able to solve the first sub-question in Problem 3 and had no idea how to approach the other two sub-questions. However, I made an effort to write down my thoughts and attempts. The grader acknowledged my "significant" progress and awarded me full credit for the entire Problem 3... You see, even under such loose grading rubis, we could only just pass the exam, if we use the usual standard: 60% as pass. The grading system for the course is quite generous, and it is stated that solving the equivalent of one question already reflects a B-level performance. Personally, I have never encountered exams as challenging as these before. This is what happen in both the midterm and the final: After two hours, I found myself only able to provide solutions for one problem, which led to a sense of panic that I had to consciously control. But luckily, I did well in the midterm, ranking 15 out of about 80 students, and I am now finger-crossing on my final, which I think I did badly yesterday...

I have faced significant challenges with the Pset and exams, but I would say that it is because of these tasks, especially while preparing for the exams, that I have gained a much better understanding of the course material. Previously, I usually forget who is in the denominator of the KL divergence equation. However, now that information is firmly embedded in my memory. I have also acquired knowledge about various properties, alternative forms (such as the variation form), and lower and upper bounds of KL divergences. These topics were not covered explicitly in the lectures but were instead explored and proven in the Pset. Personally, I have developed a strong appreciation for information geometry, which has provided me with a clear visualization of concepts like I-projection and has allowed me to view the expectation-maximization algorithm as a series of alternating I-projection and reverse I-projection steps. Additionally, I now have a deeper understanding of why the exponential family holds such importance. Initially, I knew that certain distributions, like the Gaussian and Bernoulli, belonged to the exponential family. However, I now understand that efficient estimators, I-projection, and numerous other concepts are built upon the foundation of the exponential family. Furthermore, I have come to recognize the remarkable power of linear families, such as their ability to represent marginalization constraints. The latter half on asymptotics is hard to me; sometimes it is even confusing.

To anyone who may be hesitant, I wholeheartedly recommend embracing the opportunity to enroll in 6.437. However, it is important to be prepared to dedicate 10-20 hours to the course, possess a resilient mindset, and have a genuine interest in statistical inference. Although it is undoubtedly a challenging endeavor, reflecting on my own experience, I am immensely grateful that I made the decision to embark on this academic journey. Best of luck if you choose to undertake 6.437!