Advanced Microeconomics II

Introduction
Outline

1. The course will begin with a review/overview of some preliminary technical topics that are fundamental to the rest of the course, including philosophy of social science, Bayesian probability theory and elementary information theory.

2. The theme of the course is the theoretical and empirical explanation of the relation of aggregate and individual behavior in social situations.

3. Two building blocks of this effort are the next
topics of the course: the entropy-constrained behavior model, which applies information theory to individual behavior; and the social interaction model, which represents the situation of a typical individual in a social context.

4. The course will then turn to applying these ideas to fundamental problems in political economy.

5. The first problem is the theory of decentralized spontaneous organization of an economy in which production is specialized through a division of labor and regulated by the exchange of products as commodities, the focus of Smithian and Marxian political economy. We will use the basic tools to examine the theory of
money and monetary neutrality in this type of economy.

6. The course will then extend this analysis to capitalist economies in which productive labor takes the form of wage labor and the organization of the economy is dominated by capitalist pursuit of surplus value.

7. We will then take up the study of advanced industrial capitalist economies, in which the financial sector is highly developed and the problems of social coordination include aggregate demand fluctuations and unemployment of resources.

8. The course will then apply these basic ideas to a
revision of marginalist theory with particular attention to the impact of information theory on distribution and inequality.

9. The later part of the course will take up a series of specific models and examples related to these themes.
Trigger warning:

- Learning the material in this course may change the way you understand political economy and the interpretation of economic data irreversibly, and make it impossible for you to maintain the views you now hold.

- Learning can a painful and stressful process that is traumatic to those who cling uncritically to received ideas.
Lecture 1
Preliminaries
Philosophy of social science

- Social sciences differ essentially from natural sciences because their object of study (human society) includes the investigator (the human mind).

- Social sciences have a *self-referential* character not present in the natural sciences.

- Social sciences incorporate the methods of natural science but have to transcend them.
Example: thermodynamics and social science

- For example, statistical physics has developed powerful methods for analyzing the aggregate behavior of complex systems of particles.

- Economics shares with physics the problem of studying aggregation in complex systems. Statistical physical methods can illuminate the problem of aggregation in social systems with many degrees of freedom.

- Because human beings are not passive physical
particles, however, it would be a reductive mistake to analyze social aggregation problems without taking into account human intelligence and planning.
Philosophy of science

- Scientific investigation aims at explaining phenomena we observe and experience in the real world.

- Science explains by organizing phenomenal observations and experience in terms of theories that have explanatory power.

- Good theories provide unified and consistent concrete explanations of a broad range of phenomena.

- Good theories are *parsimonious*, avoiding
unnecessary complications and multiplication of assumptions and parameters.
Probability

- The word “probability” is used in many different senses in ordinary language.

- In this course we will use it to mean "degree of belief" in a statement or hypothesis in the broadest sense, in accord with the Bayesian approach to probability. (See Jaynes for a complete discussion.)

- With this concept of probability we can coherently refer to the probability of pretty much anything we can talk about: the
probability of single unrepeateable events, of arbitrary propositions, of hypotheses, of mathematical theorems being true.

- This sense of “probability” is very closely connected to information, since one’s degree of belief in anything depends on the information one has about it.
Propositions and the laws of probability

- A “proposition” is any kind of statement that might be true or false. Propositions include measurements of quantitative variables, theories, hypotheses, statements of fact about the world. We can refer to propositions symbolically as $A$, $B$, …. 

- The marginal probability of a statement, $P[A]$, according to the Bayesian approach, is a
measure of degree of belief that $A$ is true, which can vary from 0 to 1.

- The \textit{conditional probability of $B$ given $A$}, $P[B \mid A]$, is the degree of belief that $B$ is true given that $A$ is known to be true.

- The \textit{joint probability} of $A$ and $B$, $P[AB]$, is the degree of belief that both $A$ and $B$ are true.
The product rule

- In order to believe that both $A$ and $B$ are true, it is necessary to believe first of all that $A$ is true, and then that $B$ is true. Applying this reasoning to degrees of belief we derive the *product rule* for probabilities:

$$P[A \cap B] = P[B \mid A] \cdot P[A] = P[A \mid B] \cdot P[B]$$  \hspace{1cm} (1)

- Alternatively, we could define conditional probability as:

$$P[B \mid A] = \frac{P[A \cap B]}{P[A]} \text{ if } P[A] > 0$$  \hspace{1cm} (2)

- Many errors in probabilistic reasoning arise
from assuming incorrectly that the “product rule” is $P[A B] = P[B] P[A]$. This expression is valid *only* in the case where $P[B \mid A] = P[B]$, that is, when the propositions are *independent* in the sense that the probability of one does not depend on whether the other holds or not.
The complement rule

- If someone believes \( A \) is false, they must believe \( \overline{A} \) is true. (The problems of the “law of the excluded middle” and Hegelian “negation of the negation” have to do with the many different ways in which any proposition can be false.) Applying this point to degrees of belief:

\[
P[A] + P[\overline{A}] = 1
\]  

(3)

- All of the implications of probability theory can be derived from (1) and (3).
Probabilities and frequencies

- A system of Bayesian probabilities assigns non-negative numbers summing to one to the universe of propositions under consideration.

- In many important cases, propositions (or theories) assert the frequency with which we observe some particular outcome.

- Frequencies also are systems of non-negative numbers that sum to one.

- It is easy to confuse probabilities in the Bayesian sense with frequencies.
Example: urn problem 1

- A common statistical problem is to infer the composition of an urn from the observation of a limited sample drawn from it.
  - If we know that the urn contains balls of only three colors, red, green, and blue, for example, the composition of the urn is the three numbers \( \{R, G, B\} \) representing the number of balls of each color.
  - The size of the urn is \( N = R + G + B \).
The frequencies of the colors in the urn are
\( \{ r = \frac{R}{N}, \ g = \frac{G}{N}, \ b = \frac{B}{N} \}. \)
Example: urn problem 2

- How much degree of belief, $P[A]$, we put in the statement $A$, “the urn has frequencies $\{r, g, b\}$” will depend on what information we have about the urn, including, for example, the composition of some sample drawn from it.

- The probability $P[A]$ is a degree of belief in a particular frequency of balls in the urn.
Bayes’ Theorem

- Bayes’ Theorem is an interpretation of the definition of conditional probability (2). Suppose that we have some hypothesis, or theory, $H$, and we want to determine whether some particular empirical data or observations, $D$, support or contradict the hypothesis. Using (2):

\[
P[H \mid D] = P[H] \frac{P[D \mid H]}{P[D]} \tag{4}
\]

- Translating this key formula into words, $P[H \mid D]$ is the *posterior probability*, the probability of the
hypothesis given the data, \( P[H] \) is the prior probability, the probability of the hypothesis before the data is known, \( P[D \mid H] \) is the likelihood, the conditional probability of the data given the hypothesis, and \( P[D] \), the marginal probability of the data, is a normalizing factor.

- The likelihood considered as a function of \( D \) is a normalized probability, since the hypothesis specifies a probability for every possible value of the data. But the likelihood considered as a function of \( H \) is not normalized, since the hypotheses considered might predict a very low or high probability for the data observed.
Bayes’ Theorem and updating probabilities with new information

- Bayes’ Theorem is a procedure for updating probabilities in the sense of degrees of belief in the light of new information contained in the data.

- The new information restricts the relevant set of possibilities in the original joint prior probability assignment $P[H, D]$. Mathematical theorems show that Bayesian updating is the only consistent way to modify degrees of belief to take account of new evidence (see Jaynes).

- Either probabilistic reasoning is based on some
consistent prior (whether the prior is made explicit or not) or it is vulnerable to errors of inconsistency.
Example--diagnosis

- A patient is being tested for a rare condition that appears in .01% of the population. The test has 1% false negatives and 2% false positives. The test comes back positive. What is the probability of $H$, “the patient has the condition”?

- The prior probability is $P[H] = .0001$. The data is the positive outcome of the test. The likelihood is $P[D \mid H] = .99$. From Bayes’ Theorem we have

$$P[H \mid D] \propto P[H] P[D \mid H] = .0001 \times .99 = .00099$$

$$P[\overline{H} \mid D] \propto P[\overline{H}] P[D \mid \overline{H}] = .9999 \times .02 = .019998$$
■ Normalizing the probabilities:

\[ P[H \mid D] = \frac{0.00099}{0.00099 + 0.019998} = 0.0471698 \]

\[ P[H^c \mid D] = \frac{0.019998}{0.00099 + 0.019998} = 0.95283 \]

■ This example underlines the significance of prior information. Even though the test looks quite good, and a positive test result lowers the odds against the patient having the condition from 10000:1 to 20:1, the absolute probability of the patient having the condition remains low.
Example--fairness of coin tosses

- We observe $n$ trials of an experiment each of which can result in a *hit* or a *miss*, of which $n_1$ are hits and $n_2 = n - n_1$ are misses. What odds do we put on the occurrence of a hit in the next trial, assuming no change in the conditions of the experiment?

- Suppose we think that if we ran the experiment many, many times, the proportion of hits would be approach some limit $p$, where $0 \leq p \leq 1$. The value of $p$ is the hypothesis.
The data is the observed number of hits and misses, \( \{n_1, n_2\} \).

Given \( p \), the likelihood of observing \( n_1 \) hits in \( n \) trials is the binomial probability

\[
\binom{n}{n_1} p^{n_1} (1 - p)^{n-n_1} = \frac{n!}{n_1!(n-n_1)!} p^{n_1} (1 - p)^{n-n_1}
\]  

(5)

If we have no information about \( p \) in this experiment we might adopt the \textit{uniform prior}

\[
P[p \in [p' + dp]] = \left\{ \begin{array}{ll}
dp & 0 \leq p' \leq 1 \\
0 & p' < 0, \ p' > 1\end{array} \right.
\]

(6)

From Bayes’ Theorem we have

\[
P[p \mid \{n_1, n_2\}] \propto P[p] P[\{n_1, n_2\} \mid p]
\]
\[
\begin{cases}
\binom{n}{n_1} p^{n_1} (1 - p)^{n-n_1} d\, p & 0 \leq p \leq 1 \\
0 & p' < 0, p' > 1
\end{cases}
\]

- The posterior probability of observing a hit in the next trial would then be proportional to

\[
\int_0 (n + 1) \binom{n}{n_1} p^{n_1} (1 - p)^{n-n_1} d\, p = \frac{n_1 + 1}{n + 2}
\]

- This example underlines the important point that the Bayesian method does not “estimate” one value of \( p \), say the highest posterior probability \( p \), and assume that it is the true value by “rejecting” others; instead the Bayesian method imputes a positive (though perhaps small) posterior probability to every value of \( p \).
consistent with the data and averages over these posterior probabilities to evaluate the posterior probability on the next outcome of the experiment.
Visualizing posterior probabilities

\[ P[p|\{n_1, n-n_1\}] \]

- \text{Prob[Hit]}
- \text{P[p|\{n_1, n-n_1\}]}
Information theory

Shannon’s theorem 1

- The central result in information theory is Shannon’s Theorem, which defines informational entropy as a quantitative measure of information.

- Shannon was seeking a mathematical theory of the maximum capacity of an information transmission system (a telegraph wire, or a telephone circuit, or a fiber-optic cable).

- Shannon’s theorem concerns a situation where
there is a finite set of $n$ messages known to both the transmitter and receiver. (These messages might be phonemes in the case of speech, or letters, or digits, or the two possible messages “the British troops are traveling by water” and “the British troops are traveling by land”.)

One scheme to transmit the messages is to agree on a numbering of them and to transmit the corresponding number over the transmission channel. This method requires $\log_2[n]$ bits of information, since $n = 2^{\log_2[n]}$. The time it will take to send the message depends on how many bits of information the channel can transmit in a given time, such as a second, its *bandwidth*. 
Shannon’s theorem 2

- Shannon noted that if some messages were sent more often than others, it would be possible to improve the *average* transmission rate by assigning short code words to the more frequently transmitted messages.

- (Samuel Morse's telegraphic code assigns the shortest codes, dot and dash, to the most frequently encountered letters, "e" and "t".)

- If there is a frequency $f_i > 0$ assumed for each message $i$, with $\sum_i f_i = 1$, Shannon proved that it
is possible to devise a coding in which the $i$th message is assigned a code word of $\lceil -\log_2[f_i] \rceil$ bits. Furthermore Shannon showed that this is the best one can do to minimize the average message length as the set of messages becomes large, that is, as $n \to \infty$.

- A set of messages with specified frequencies is called an *ensemble*.

- If we take logarithms to the base $e$, the measure of information is the *nat*. A nat is equal to $\log_2[e] = 1.4427$ bits.

- We will take logarithms to base 2 unless another base is specified.
The notation \( \lceil x \rceil \) means the Ceiling\([x]\), the smallest integer greater than or equal to \( x \).
Entropy

- Using Shannon-optimal coding, the average message length for an ensemble will be the informational entropy of the frequencies:

\[
H[f_1, \ldots, f_n] = - \sum_i f_i \log[f_i]
\]  

(7)

- If the \( n \) messages will be sent with equal frequency, \( f_i = \frac{1}{n} \), the informational entropy is:

\[
H\left[\left\{\frac{1}{n}, \ldots, \frac{1}{n}\right\}\right] = - \sum_i \frac{1}{n} \log \left[\frac{1}{n}\right] = \log[n]
\]

- If one of the \( n \) messages will be sent every
time, \( f_1 = 1, f_{i \neq 1} = 0 \), the informational entropy is (setting \( 0 \log[0] = 0 \)):

\[
H[\{1, 0, \ldots, 0\}] = 1 \log[1] = 0
\]

■ (In this case there is no need to send any message at all!)

■ Thus to every frequency distribution there corresponds a code length assignment, and, equally important, to every coding of messages there corresponds an implicit frequency distribution over the messages.