Massachusetts Institute of Technology
6.042J/18.062J: Mathematics for Computer Science
Professor David Karger

Lecture 20
April 29, 1996

Notes 20

1 Probability and Set Theory

Before discussing some more complicated probability problems, let's develop some useful connections between probabiliy and set theory. Recall from the last lecture that

Definition 1 A *sample space* for a problem (or experiment) is a set consisting of all possible outcomes.

Definition 2 A *probability space* consists of a sample space S and a probability function $Pr : S \to \mathbb{R}$, such that

1. For all $s \in S$, $0 \leq Pr(S) \leq 1$.
2. $\sum_{s \in S} Pr(s) = 1$.

Definition 3 An *event* is a subset of the sample space.

Definition 4 The probability of an event S' is the sum of the probabilities of all outcomes in the event.

Using these definitions, we can establish some basic facts.

Theorem 1.1 Let A and B be events such that $A \subseteq B$. Then $Pr(A) \leq Pr(B)$.

Proof.

$$Pr(A) = \sum_{s \in A} Pr(s) \leq \sum_{s \in B} Pr(s) = Pr(B).$$

■

For an event A, let $\bar{A}$ denote the *complement* of A.

Theorem 1.2 For any event A,

$$Pr(\bar{A}) = 1 - Pr(A).$$

Proof.

$$\begin{aligned} Pr(\bar{A}) + Pr(A) &= \sum_{s \in A} Pr(s) + \sum_{s \notin A} Pr(s) \\ &= \sum_{s \in S} Pr(s) \\ &= 1 \end{aligned}$$

where the first equality follows from the definition of complementation and the second follows from the definition of a probability space. ■

Also this has a simple proof, often it's much easier to calculate a probability by first calculating the probability of the complement and then subtracting from 1.

We can use the inclusion-exclusion principle to get probabilities of unions.

Theorem 1.3 For any events A and B, $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$.

Proof.

$$\begin{aligned} Pr(A \cup B) &= \sum_{s \in A \cup B} Pr(s) \\ &= \sum_{s \in A} Pr(s) + \sum_{s \in B} Pr(s) - \sum_{s \in A \cap B} Pr(s) \\ &= Pr(A) + Pr(B) - Pr(A \cap B). \end{aligned}$$

■

Sometimes it's hard to calculate the exact probabilities of unions and intersections. But the advantage of a result like Theorem 1.3 is that it allows us to give bounds on these probabilties.

Example 1 Toss a fair coin twice. The sample space is $S = \{HH, HT, TH, TT\}$. Let A be the event of getting at least one head, i.e. $A = \{HH, HT, TH\}$, and let B be the event of getting at least one tail, i.e. $B = \{HT, TH, TT\}$. Since the coin is fair all outcomes are equally likely, so $P(A) = 3/4$, and $P(B) = 3/4$. By the above theorem,

$$\begin{array}{ccccccc} P(A \cup B) & = & P(A) & + & P(B) & - & P(A \cap B) \\ 1 & = & 3/4 & + & 3/4 & - & 1/2 \end{array}$$

Example 2 A study determines that out of 1000 men, 2% are "too heavy" and 3% are "too tall" to be an astronaut. What can we say about the percentage of men that are unfit to be astronauts (because they are either too tall or too heavy)?

Let S be the sample space of all men. Let the event A be the set of men that are "too heavy", and let B be the set of men that are "too tall". Thus we are given $P(A) = 0.02$ and $P(B) = 0.03$. By Boole's inequality, we can conclude that

$$P(A \cup B) \leq P(A) + P(B) = 0.05$$

On the other hand, we can see that the lower bound is

$$P(A \cup B) \geq \max(P(A), P(B)) = 0.3$$

Since we don't know $P(A \cap B)$, this is all we can conclude about $P(A \cup B)$.

Exercise 1 Prove that $P(A_1 \cup A_2 \cup \cdots \cup A_k) \leq P(A_1) + P(A_2) + \cdots + P(A_k)$.

Exercise 2 Prove that $P(A_1 \cup A_2 \cup \cdots \cup A_k) \geq \sum_{i=1}^{k} P(A_i) - \sum_{1 \leq i < j \leq k} P(A_i \cap A_j)$.

Exercise 3 Generalize the previous two exercises and prove the result.

Independence

It is often useful to be able to compute $P(A \cap B)$ from $P(A)$ and $P(B)$. In a special case, but a common one, we can do so.

Definition 5 Events A and B are said to be *independent* iff

$$P(A \cap B) = P(A)P(B).$$

Example 3 Suppose we a drawing a card from the randomly shuffled deck (i.e. each card has a probability of $1/52$ to be picked). Let A be the event of drawing a spade, and let B be the event of drawing a 5. Then (since there are 13 spades in the deck) $P(A) = 13/52 = 1/4$ and (since there are four 5's) $P(B) = 4/52 = 1/13$. The probability of drawing the 5 of spades is $P(A \cap B) = 1/52 = 1/4 \cdot 1/13 = P(A) \cdot P(B)$. Therefore the events A and B are independent.

Example 4 Let the sample space S be the cells of an 3×3 array, with each cell having probability of $1/9$ of being selected. Let A be the event that a cell in the first row is selected, and let B be the event that a cell in the first column is selected. $A \cap B$ is the event that the upper-left cell of the array is selected. $P(A) = 1/3$ and $P(B) = 1/3$. On the other hand, $P(A \cap B) = 1/9$, so the events A and B are independent.

Example 5 Consider the sample space from the previous example, and let A be the event that a cell in the first two rows is selected, and let B be the event that a cell in the second two rows is selected. Then $A \cap B$ is the event that a cell in the second row is selected. Now, $P(A) = P(B) = 2/3$, but $P(A \cap B) = 1/3$. The events A and B are *not* independent.

If events A and B are independent, then knowing whether or not A occurred doesn't influence one's estimate as to whether B occurred. We can talk more precisely in this vein using the notion of *conditional probability.*

Independence of more than two events

We say that events A and B are independent if $P(AB) = P(A)P(B)$. But what about more than two events?

Consider an experiment consisting of n coin flips and let S_i be the sample space for the ith flip, i.e. $S_i = \{H_i, T_i\}$. Let S be the sample space for the entire experiment. Clearly, $S = S_1 \times S_2 \times \cdots \times S_n$. We wish to say that the coin flips are independent if $P(x_1 x_2 \cdots x_n) = P(x_1)P(x_2)\cdots P(x_n)$ if $x_1 x_2 \cdots x_n$ is a sequence of coin flips.

Suppose $n = 3$ and the coin is biased so that the number of tails is always odd. In our sample space we have:

$$\begin{array}{lll} P(H_1H_2H_3) = 0 & P(H_1H_2T_3) = 1/4 & P(H_1T_2H_3) = 1/4 \\ P(T_1H_2T_3) = 0 & P(H_1T_2H_3) = 1/4 & P(T_1T_2H_3) = 0 \\ P(H_1T_2T_3) = 0 & P(T_1T_2T_3) = 0 & \end{array}$$

In this example $P(H_i) = P(T_i) = 1/2$ for $i = 1, 2, 3$. The pairwise combinations also work. For example $P(H_1H_2) = 1/4$ and so on. However, $P(H_1H_2H_3) = 0 \neq P(H_1)P(H_2)P(H_3)$.

We now make the following definition:

A set $\{A_1, A_2, \cdots A_n\}$ of events is k-wise independent if for any subset $\{A_{i_1}, A_{i_2}, \cdots A_{i_j}\}$ of j events, for $j \leq k$ we have

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_j}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_j})$$

2 Conditional Probability

In a lot of problems (including the one we looked at in the last lecture), we want to consider the probability that an event A happens, *given* that we know a certain other event B has happened. We write

$$Pr(A \mid B)$$

to denote the probability of event A, given B.

There is a very simple formula for this, which can be viewed as follows. Once we know that event B has happened, we are in the subset of the sample space equal to B. We now want to know what fraction of this set consists of outcomes in A. Thus,

$$Pr(A \mid B) = \frac{Pr(A \cap B)}{Pr(B)}.$$

We will take this formula as the definition of the *conditional probability of A given B*.

Example 6 Suppose a baseball team is playing a best-2-out-of-3 series. The probability it wins the first game is $\frac{1}{2}$. The probability of winning a game following a win in the previous game is $\frac{2}{3}$; the probability of winning following a loss is only $\frac{1}{3}$.

The question is: what is the probability the team wins the series if it wins the first game?

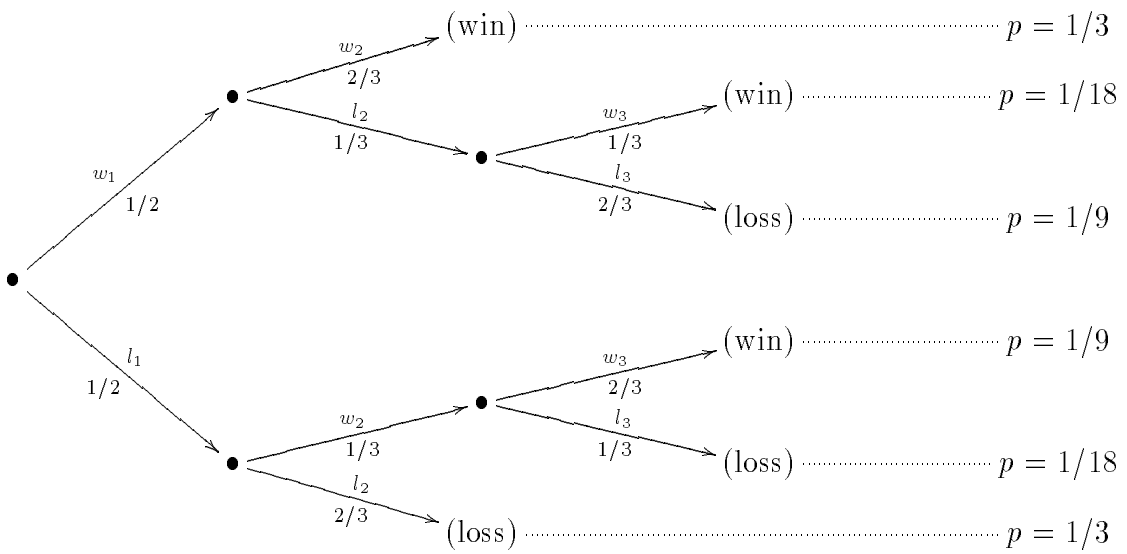
There are actually two types of conditioning going on here:

1. The probability of winning a game depends on the probability of having won the previous game.
2. We're asked for the probability of winning the series, *given* that the team wins the first game.

We solve this problem in the following step-by-step way.

1. Draw the sample space. (See the figure below.)
2. Assign probabilities to the outcomes. These are all calculated using the same principle; let's show we calculate the probability of the outcome (tree branch) WLL.

$$\begin{aligned} Pr(WLL) &= Pr(\text{Win 1st}) \cdot Pr(\text{Lose 2nd} \mid \text{Win 1st}) \cdot Pr(\text{Lose 3rd} \mid \text{Win 1st}, \text{Lose 2nd}) \\ &= \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{2}{3} \\ &= \frac{1}{9} \end{aligned}$$



3. Now we identify the events. A is the event that the team wins the series, and B is the event that the team wins the first game.
4. Finally, the answer we want is $Pr(A \mid B)$. We compute this using the formula, and adding up the probabilities from the figure:

$$Pr(A \mid B) = \frac{Pr(A \cap B)}{Pr(B)} = \frac{\frac{1}{3} + \frac{1}{18}}{\frac{1}{3} + \frac{1}{9} + \frac{1}{18}} = \frac{\frac{7}{18}}{\frac{9}{18}} = \frac{7}{9}.$$

5. We can also answer the question "What is the probability that the series lasts three games?". From the picture, we add the probabilities of all the series of length three. $1/18 + 1/9 + 1/9 + 1/18 = 1/3$.

3 Another Example

Conditional probability can lead to some counter-intuitive answers; but things really do work out if you just follow the definition. Here's a problem that always confuses people. It's based on calculating the probability of an event A, given an event B that comes *after* it in time. In such cases, the conditional probability is often called the *a posteriori probability* — though it's really the same thing mathematically.

Suppose I have two coins — one is a fair coin, which comes up heads or tails with probability $\frac{1}{2}$ each, and the other is an unfair coin, which always comes up heads. Now I choose one of the two coins with probability $\frac{1}{2}$ each, and I flip this chosen coin once. If it comes up heads, what is the probability that I picked the fair coin?

The trickiness here is that our intuitive notion of conditional probability does not always match the definition as $Pr(A \cap B)/Pr(B)$ — in this case, the notion of "causality" seems to go backwards. But in such cases, we always have to follow the definition.

As in the previous example, we solve this problem in the following steps.

1. Draw the sample space. (See the figure below.)
2. Assign probabilities to the events, again as in the previous example.
3. Define the relevant events. Here, A is the event that we choose the fair coin, and B is the event that the result is heads.
4. $$Pr(A \mid B) = \frac{Pr(A \cap B)}{Pr(B)} = \frac{\frac{1}{4}}{\frac{1}{4} + \frac{1}{2}} = \frac{1}{3}.$$

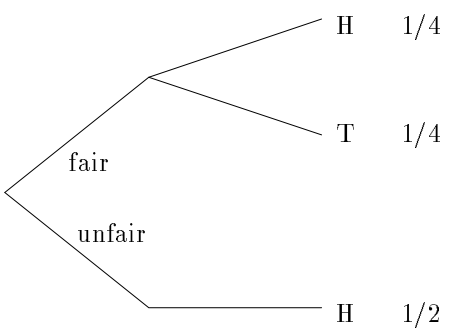


Figure 1: Flipping Coins

The key point here is that we knew the probability of picking the fair coin was exactly $\frac{1}{2}$. Without this, we could not have computed $Pr(A \mid B)$.

For example, say that we didn't know that we picked the fair coin with probability 1/2, we performed the experiment 100 times and it came up heads every time. What can we say about the probability of having picked the fair coin? The answer is "absolutely NOTHING!". We can see this by drawing the sample space of the new experiment (see figure 3).

$Pr(A \mid B) = \frac{p2^{-100}}{1-p+2^{-100}} = \frac{p}{(1-p)2^{-100}+1}$

The probability of this event appears to be very small, but if $p = 1$ then $Pr(A \mid B) = 1$.

4 Yet Another Example

Assume that 10% of the American polulation have disease X. You don't know if you have the disease but there is a test. If you have the disease, there is a 10% chance that the test

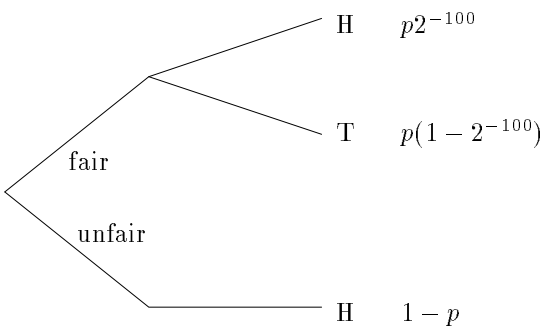


Figure 2: Flipping Coins

will come out negative. If you don't have the disease there is a 30% chance that the test will come out positive. Now suppose that a random American is selected and tested. What is the probability that this person has the disease given that the test is positive?

Let A be the event that the person has the disease and B be the event that the test is positive. The probability that both events occur is $0.1 \cdot 0.9 = 0.09$, so $Pr(A \cap B) = 0.09$. The probability that the test comes out positive is $Pr(B) = 0.1 \cdot 0.9 + 0.9 \cdot 0.3 = 0.36$ (draw the sample space to convince yourself), so we have

$Pr(A \mid B) = \frac{.09}{.36} = \frac{1}{4}$

Even if you test positive, you are not likely to have the disease, even though the test has low false rates. Indeed, $Pr($ test is correct $) = .09 + .63 = .72$. The test is accurate more than 70% of the time, but that does not mean that it is likely to be accurate when it is positive. This is because the vast majority of people don't have the disease and thus there are more people that don't have the disease but test positive (27%) than people who do have the disease and test positive (9%).
It's worth looking at the alternative test which always says no. This test is accurate 90% of the time, which is better than 72%!

Example:One-Time Pads

In this example from Cryptography, we wish to ascertain what information an eavesdropper can learn by viewing an encrypted message.

Let us assume that Alice wants to send an n-bit message (m $= m_1 m_2 \cdots m_n$ where $m_i = 0$ or 1) to Bob. She picks an n-bit key (k $= k_1 k_2 \cdots k_n$) at random and the ciphertext (encypted message) is constructed according to the rule ($c_i = m_i \oplus k_i$ where $\oplus$ denotes "exclusive or" or

"addition $mod(2)$ ")). Note that if Bob knows the key then he can reconstruct the message since $(m_i \oplus k_i) \oplus k_i = m_i$.

To formulate this in a probabilistic framework, we let $P(M = x)$ denote what the eavesdropper knows initially about the message. He may have some initial guess about how likely each n-bit message is. After seeing the ciphertext his knowledge will be denoted by $P(M = x \mid C = y)$. If our scheme is secure, then the two quantities should be the same. The eavesdropper should learn nothing about the message after seeing the ciphertext.

Now:

$$P(M = x \mid C = y) = \frac{P((M = x) \ \& \ (C = y))}{P(C = y)}$$

But

$$\begin{aligned} P((M = x) \ \& \ (C = y)) &= P(M = x)P(C = y \mid M = x) \\ &= P(M = x)2^{-n} \end{aligned}$$

since $P(C = y \mid M = x) = P(k = x \oplus y)$ and we assumed that the key is chosen at random. Similarly

$$\begin{aligned} P(C = y) &= \sum_x P(M = x)P(C = y \mid M = x) \\ &= \sum_x P(M = x)2^{-n} \\ &= 2^{-n} \end{aligned}$$

Thus $P(M = x \mid C) = y = \frac{P(M=x)2^{-n}}{2^{-n}} = P(M = x)$.

Example: Berkeley Admissions

Recently, a lawsuit was filed against UC Berkeley. It charged that women had a smaller chance of getting admitted than men. The plaintiff claimed that in every department at Berkeley, the probability of admission given that one was female was smaller than the probability of admissions given that the applicant was male. The defense countered with the fact that overall the situation was reversed – the probability of admission to Berkeley overall was higher given that the applicant was female. How could this happen?

Let us consider a smaller example where there are only two departments, D_1 and D_2. Let our sample space be the set of all applicants. We are interested in the following events:

- F = "the applicant is female"
- M = "the applicant is male"
- A = "the applicant was admitted"

- D_i = "the application made was to department i"

Consider the following table of probabilities:

	F	M	Overall
D_1	4/5	1/1	5/6
D_2	0/5	1/9	1/14
Overall	4/10	2/10	6/20

These numbers correspond to the following quantities:

	F	M	Overall
D_1	$P(A \mid FD_1)$	$P(A \mid MD_1)$	$P(A \mid D_1)$
D_2	$P(A \mid FD_2)$	$P(A \mid MD_2)$	$P(A \mid D_2)$
Overall	$P(A \mid F)$	$P(A \mid M)$	$P(A)$

From the table we see that $P(A \mid FD_i) < P(A \mid MD_i)$ for all i, but $P(A \mid F) > P(A \mid M)$. This is explained by looking at the formula for $P(A \mid F)$:

$$\begin{aligned} P(A \mid F) &= P(AD_1 \mid F) + P(AD_2 \mid F) \\ &= P(A \mid FD_1)P(D_1 \mid F) + P(A \mid FD_2)P(D_2 \mid F) \end{aligned}$$

To see this note that $P(A) = P(A \cap D_1) + P(A \cap D_2)$ since D_1 and D_2 are disjoint and so $P(A \mid F) = P(AD_1 \mid F) + P(AD_2 \mid F)$. Thus $4/10 = 4/5 \cdot 5/10 + 0/5 \cdot 5/10$.

It turns out that the defense prevailed in the case.

Bayes Rule

Are $P(A|B)$ and $P(B|A)$ related? Yes, both relate to $P(A \cap B)$:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad \text{and} \qquad P(B|A) = \frac{P(B \cap A)}{P(A)},$$

so

$$P(A|B)P(B) = P(A \cap B) = P(B|A)P(A).$$

This yields the famous *Bayes Rule:*

$$P(A|B) = \frac{P(B|A)}{P(B)}P(A).$$

In Bayes Rule, A is the event we are interested in, B is the event that represents outcome of an experiment that is to reinforce our belief that A has/has not happened. The probability $P(A)$ is called the *prior* or *a priori* probability of A. The probability $P(A|B)$ is called the *posterior* or *a posteriori* probability of A (after it is known that B has occurred).

Example 7 At a school 40% of students are male (call that A), and 60% of males and 80% of all students play sports (call that B). What is the probability that a randomly chosen student is a male, given that the student plays sports? We are given $P(A) = 40\%$, $P(B) = 80\%$ and $P(B|A) = 60\%$. We want to compute $P(A|B)$. By Bayes Rule

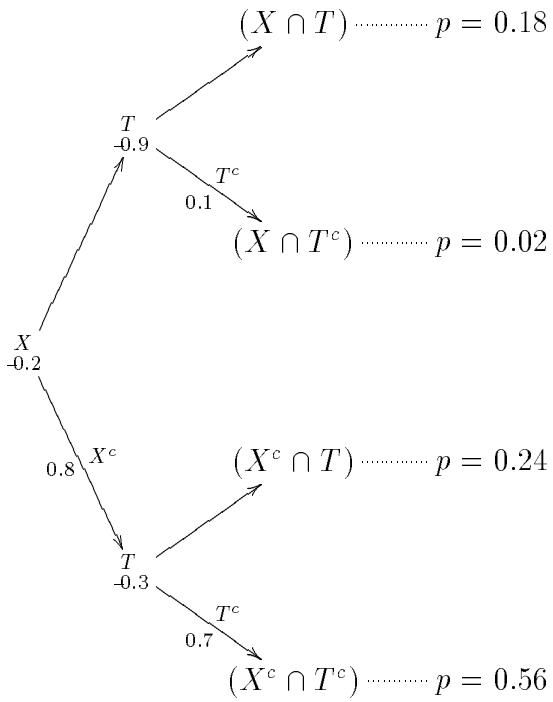
$$P(A|B) = \frac{P(B|A)}{P(B)}P(A) = \frac{0.6 \times 0.4}{0.8} = 30\%.$$

One can often represent a sample space as a sequence of choices or events that can be drawn as a tree. If one labels the branches of the tree with conditional probabilities, the probability of reaching a leaf in the tree is the product of the labels along the path from the root to the leaf (cf. Multiplication Theorem). We illustrate this in the following example.

Example 8 Suppose 20% of population has gene X (call this event X). There is a test for having the gene, which if the subject has gene X, gives positive answer 90% of the time, and if the subject does not have gene X, gives negative answer 70% of the time. In other words, the test is 90% reliable if the subject has the gene, and 70% reliable on the subjects without the gene. (This is a common way of measuring reliability of a test.)

If Joe has tested positive, what is the probability he has the gene?

Let T be the event that the test gives a positive answer. Then we are asking what $P(X|T)$ is. We are given $P(T|X) = 0.9$ and $P(T^c|X^c) = 0.7$. We can draw the tree as before.



By Bayes Rule we have

$$P(X|T) = \frac{P(T|X)P(X)}{P(T)} = \frac{0.9 \times 0.2}{0.18 + 0.24} = 3/7.$$

This is less than 1/2! Even though Joe tested positive for gene X, and test is correct more then half of the times, he is still more likely *not* to have the gene! The reason is that a priori probability of having the gene is relatively low (20%).

One can view the probability as "degree of belief" that something is true. What does it mean that "Joe has a 3/7 chance of having gene X"? Joe either does or does not have the gene. However, since we cannot know for certain whether he does, we use probability to measure our belief that he does have the gene. In this light, the positive test is evidence that he does have the gene. How convincing is the evidence? Not entirely (the probability is still less than 1/2), but it does increase one's belief in the proposition that Joe has the gene.

Odds and Bayes Rule

For any event A we define

$$\text{Odds}(A) = \frac{P(A)}{P(A^c)}.$$

For example the odds of a fair coin giving heads are $\frac{1}{2}/\frac{1}{2} = 1$. The odds of a die rolling a six are $\frac{1/6}{5/6} = 1/5$ (or 1 to 5). Since

$$\text{Odds}(A) = \frac{P(A)}{1 - P(A)}$$

we see that

$$P(A) = \frac{\text{Odds}(A)}{1 + \text{Odds}(A)}$$

If the odds for an event are x/y then it's a fair bet to put up x of your dollars (on A) against y of your opponents (on A^c).

From Bayes Rule we have that:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

and

$$P(A^c \mid B) = \frac{P(B \mid A^c)P(A^c)}{P(B)}$$

Thus

$$\begin{aligned} \frac{P(A \mid B)}{P(A^c \mid B)} &= \frac{P(B \mid A)}{P(B \mid A^c)} \cdot \frac{P(A)}{P(A^c)} \quad \text{or} \\ \text{Odds}(A \mid B) &= \frac{P(B \mid A)}{P(B \mid A^c)} \cdot \text{Odds}(A) \end{aligned}$$

This shows how to update the odds based on "new information". Odds(A) is called the *initial odds*, Odds$(A \mid B)$ denotes the *final odds* and $P(B \mid A)/P(B \mid A^c)$ is called the *likelihood ratio*.

For example suppose you have two coins in your pocket. Coin 1 is fair while coin 2 is double headed. You pick a coin at random and (without looking) then proceed to toss it. Let A be the event that you picked the double headed coin and B be the event that the toss came up heads. The initial odds are 1 and the likelihood ratio is 2. This the final odds are 2. If B is the event that the toss comes up tails, the likelihood ratio is 0 and so the final odds are 0.

Suppose we toss the coin again. Can we multiply by the likelihood ratio again to come up with new final odds? Should the odds go up to 4 when we toss the coin again? Yes, if the tosses are independent:

$$\text{Odds}(A|BB) = \frac{P(BB \mid A)}{P(BB \mid A^C)}\text{Odds}(A)$$

Now $P(BB \mid A) = P(B \mid A)P(B \mid A)$ since the tosses are independent. Thus

$$\text{Odds}(A \mid BB) = \frac{P(B \mid A)}{P(B \mid A^c)}\frac{P(B \mid A)}{P(B \mid A^c)}\text{Odds}(A)$$

This is just the likelihood ratio times Odds$(A \mid B)$.

Returning the example of medical testing. Let's say that 1% of the population has disease A. Thus Odds$(A) = 1/99$. There is a test for A that has a 5% error rate. If we denote by $+$ the event that the test comes up positive we see that

$$\begin{aligned} \text{Odds}(A \mid +) &= \frac{P(+ \mid A)}{P(+ \mid A^c)}\text{Odds}(A) \\ &= \frac{95}{5}\frac{1}{99} \\ &= 19/99 \end{aligned}$$

So the odds have increased to 19/99. If we perform the test again (independently with same error)

$$\text{Odds}(A \mid ++) = 19^2 1/99 = 3.646$$

and so $P(A \mid ++) = 3.646/4.646 = 0.7847$

The odds formulation is quite convenient since we only need to multiply by the likelihood ratio each time to update the odds.