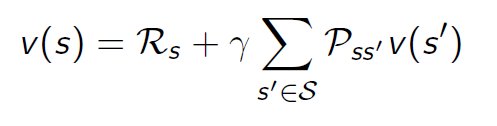
## Introduction to Markov Decision Processes – The Value Function

I originally expected this to be another easy week. But I couldn’t find a good place to stop. So it’s more challenging than I originally expected.

Again, there are multiple ways to cover the same material.

1. Watch Silver, [Markov Decision Processes](https://youtu.be/lfHX2hHRMVQ?t=13m), 13:00 – 42:54. These are the [key slides](https://drive.google.com/file/d/10Y6nRwMLwJkbwb9WAsMQoDIovNdpvRdH/view?usp=sharing).

There is an important difference between how Silver defines the *value* of a state and how the other readings (and lectures) define the *value* of a state. In Silver’s version, the value of a state includes all possible transitions that might occur from that state. Each transition has some defined probability of being taken. So the value of a state is the average of the future rewards for all the possible transitions weighted by the probability of that transition occurring.



This definition depends on Pss’, the probability of making the transition from state s to state s’. Silver points out that this system of equations can be solved because it is n equations in n unknowns.

In the other definitions, the assumption is made that one takes the *best* action. No probability is involved. (It’s still possible, though, that once an action is selected, it may or may not do what one expects. There may still be stochasticity in the environment!)

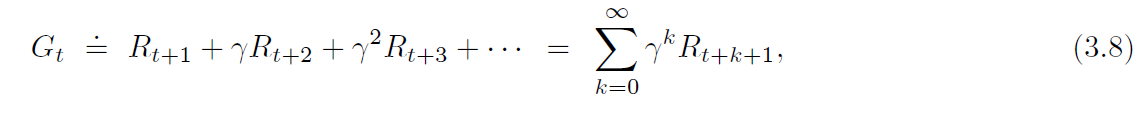
1. Watch videos 2.13-2.19 of the [Udacity course](https://classroom.udacity.com/courses/ud600/lessons/4100878601/concepts/6512308700923). Charles and Michael use the term “utility” (at least up to this point) for what the other courses refer to as “value.” The *utility of a state* (in their terms) is the same as the *value of a state* in the others. This is the Udacity course’s definition of the utility (value) of a state. Compare it to Silver’s.

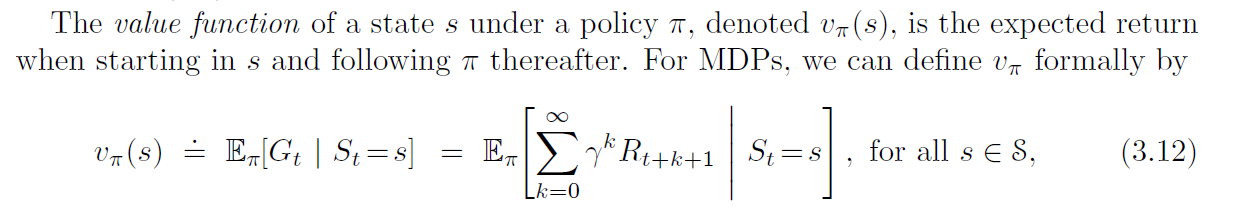


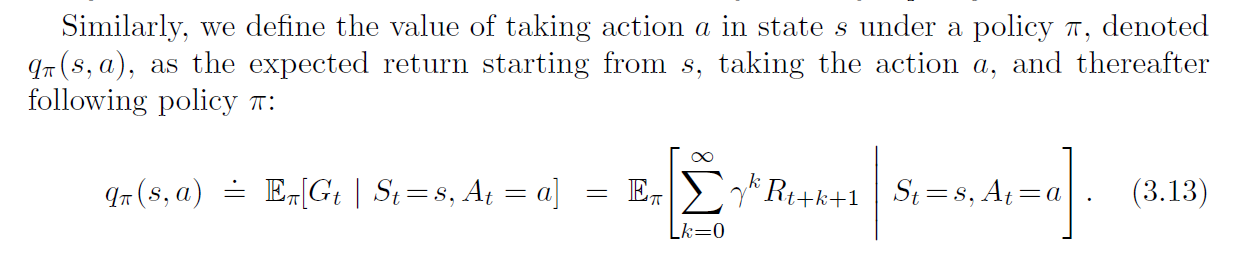
The *max* is taken over all possible actions. By using *max* one assumes that one takes the best action. Charles points out that his system of equations would be solvable were it not for the *max* function.

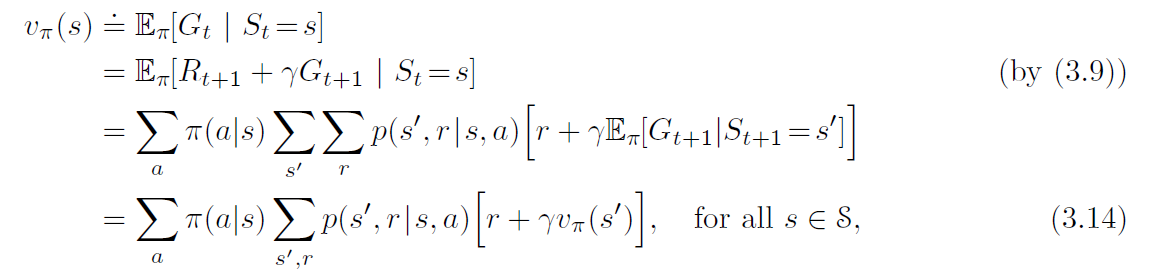
1. Read [Sutton and Bartow](http://incompleteideas.net/book/the-book-2nd.html), 3.3-3.5. Here is Sutton’s definition of the value of a state. (Note that he defines *value* explicitly in terms of a policy π. Charles and Michael use the implicit policy of taking the best action.)

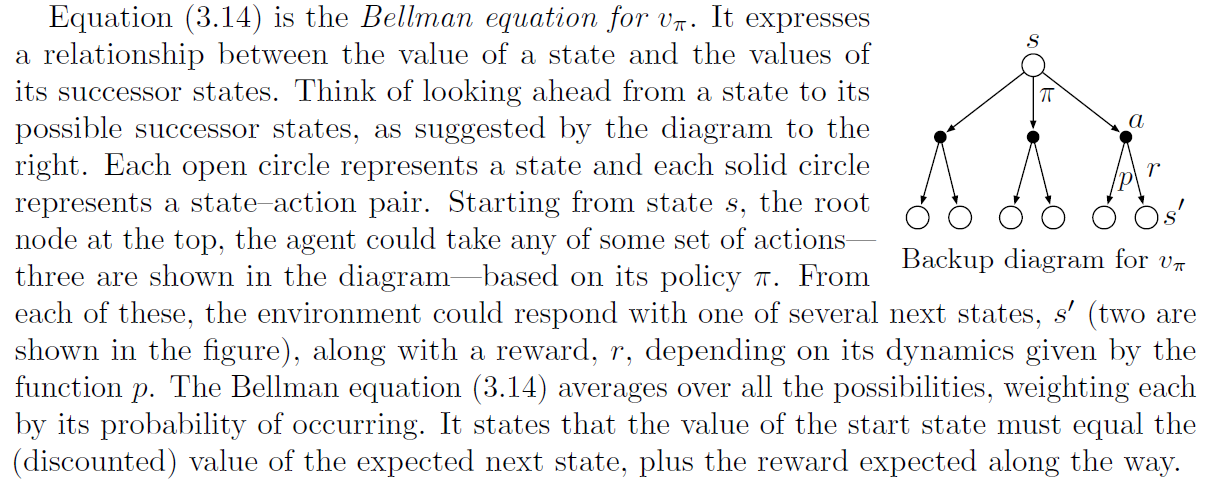
Define Gt as the expected future return from time t onward, i.e., the discounted sum of future rewards.







Even though it looks complicated, **you should be able to read (and understand and explain) the previous equations.** Given that, you should understand this paragraph toward the bottom of p 59.



* **Do the following exercises, and submit your results to CSNS. Exercises 3.7** (Notice that this problem does not mention discounting), **3.8**, **3.9**, **3.10**, (Charles proves this in one of the lectures.), and **3.14**. But instead of solving just the center square in Exercise 3.14, write code that generates all the numbers in the grid on the right hand side of the figure in Example 3.5 (p 60). (Here is [my code](https://drive.google.com/file/d/1jYbtK1ft9-VZQAgJXRuyckXy1qh7B8p-/view?usp=sharing) for this problem, in the style of the CS188 code.)