Stochastic Control with Applications to Finance (MATH69122)

Neil Walton

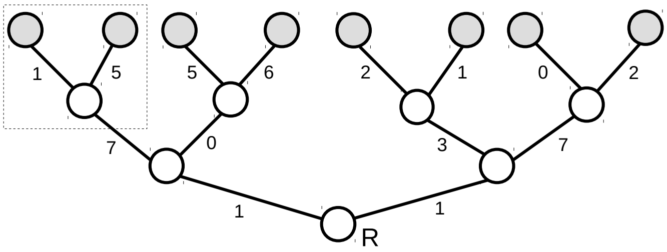
January 26, 2019

Contents

1 Dynamic Programming

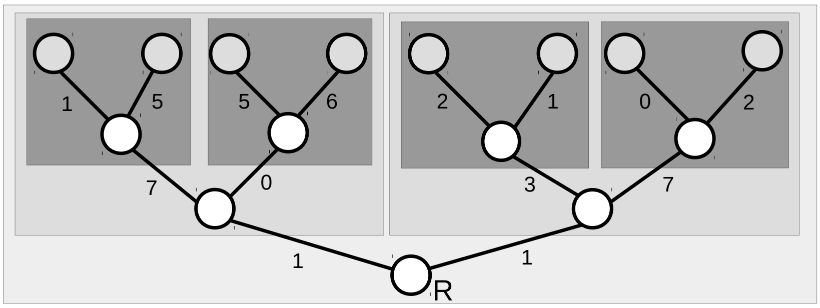
An Introductory Example

In the figure below there is a tree consisting of a root node labelled *R* and two leaf nodes colored grey. For each edge, there is a cost. Your task is to find the lowest cost path from the root node to a leaf.



There are a number of ways to solve this, such as enumerating all paths. However, we are interested in one approach where the problem is solved backwards, through a sequence of smaller subproblems. Specifically, once we reach the penultimate node on the left (in the dashed box) then it is clearly optimal to go left with a cost of 1. This solves an easier sub problem and, after solving each sub problem, we can then attack a slightly bigger problem. If we solve for each leaf in this way we can solve the problem for the antepenultimate nodes (the node before the penultimate node).

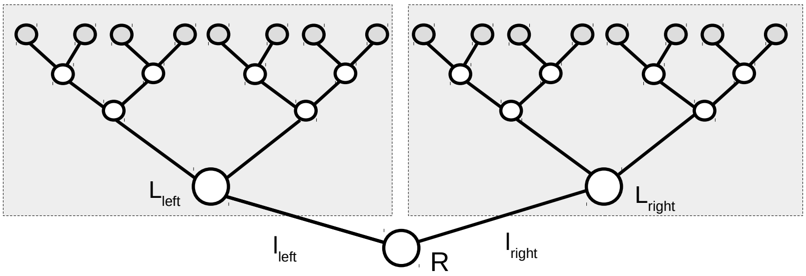
Thus the problem of optimizing the cost of the original tree can be broken down to a sequence of much simpler optimizations given by the shaded boxed below.



From this we see the optimal path has a cost of 5 and consists of going right, then left, then right.

Let's consider the problem a little more generally in the next figure. The tree on the righthand-side has a lowest cost path of value L_{rhs} and the lefthand-side tree has lowest cost L_{lhs} and the edges leading to each, respective tree, have costs l_{rhs} and l_{lhs}. Once the decision to go left or right is made (at cost l_{rhs} or l_{lhs}) it is optimal to follow the lowest cost path (at cost L_{rhs} or L_{lhs}). So L, the minimal cost path from the root to a leaf node satisfies

$$L = \min_{a \in \{lhs,rhs\}} \{l_a + L_a\} .$$



Similarly, convince yourself that the same argument applies from any node x in the tree network that is

$$L_x = \min_{a \in \{lhs,rhs\}} \left\{l_a + L_{x(a)}\right\} .$$

where L_x is the minimum cost from x to a leaf node and where for $a \in \{lhs, rhs\}$ $x(a)$ is the node to the lefthand-side or righthand-side of x. The equation above is an example of the *Bellman equation* for this problem, and in our example, we solved this equation recursively starting from leaf nodes and working our way back to the root node.

The idea of solving a problem from back to front and the idea of iterating on the above equation to solve an optimisation problem lies at the heart of dynamic programming.

Definition

We now give a general definition of a *dynamic programming*:

Time is discrete $t = 0, 1, ..., T$; $x_t \in \mathcal{X}$ is the state at time t; $a_t \in \mathcal{A}_t$ is the action at time t; The state evolves according to functions $f_t : \mathcal{X} \times \mathcal{A}_t \to \mathcal{X}$. Here

$$x_{t+1} = f(x_t, a_t). \tag{Plant eq}$$

This is called the Plant Equation. A policy π choses an action π_t at each time t. The (instantaneous) reward for taking action a in state x at time t is $r_t(a, x)$ and $r_T(x)$ is the reward for terminating in state x at time T.

Def 1 (Dynamic Program)**.** *Given initial state x_0, a dynamic program is the optimization*

$$\begin{aligned} V(x_0) := \textit{Maximize} \quad & R(x_0, \mathbf{a}) := \sum_{t=0}^{T-1} r_t(x_t, a_t) + r_T(x_T) & & \text{(DP)} \\ \textit{subject to} \quad & x_{t+1} = f(x_t, a_t), & & t = 0, ..., T-1 \\ \textit{over} \quad & a_t \in \mathcal{A}_t, & & t = 0, ..., T-1 \end{aligned}$$

Further, let $R_\tau(x_\tau, \mathbf{a})$ (Resp. $V_\tau(x_\tau)$) be the objective (Resp. optimal objective) for (DP) *when the summation is started from $t = \tau$, rather than $t = 0$.*

Thrm 2 (Bellman's Equation)**.** *$V_T(x) = r_T(x)$ and for $t = T-1, ..., 0$*

$$V_t(x_t) = \max_{a_t \in \mathcal{A}_t} \{ r_t(x_t, a_t) + V_{t+1}(x_{t+1}) \}, \tag{Bell eq}$$

where $x_t \in \mathcal{X}$ and $x_{t+1} = f_t(x_t, a_t)$.

Proof 2 Let $\mathbf{a}_t = (a_t, ..., a_{T-1})$. Note that $R_t(x_t, \mathbf{a}_t) = r_t(x_t, a_t) + R_{t+1}(x_{t+1}, \mathbf{a}_{t+1})$.

$$\begin{aligned} V_t(x_t) &= \max_{\mathbf{a}_t} \{ R_t(x_t, \mathbf{a}_t) \} = \max_{a_t} \max_{\mathbf{a}_{t+1}} \{ r_t(x_t, a_t) + R_{t+1}(x_{t+1}, \mathbf{a}_t) \} \\ &= \max_{a_t} \left\{ r_t(x_t, a_t) + \max_{\mathbf{a}_{t+1}} R_{t+1}(x_{t+1}, \mathbf{a}_t) \right\} = \max_{a_t \in \mathcal{A}_t} \{ r_t(x_t, a_t) + V_{t+1}(x_{t+1}) \}. \end{aligned}$$

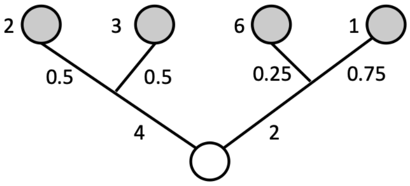
□

2 Markov Decision Processes

Markov decision processes are essentially the randomized equivalent of a dynamic program. Let's first consider how to randomize the tree example introduced in Section 1.

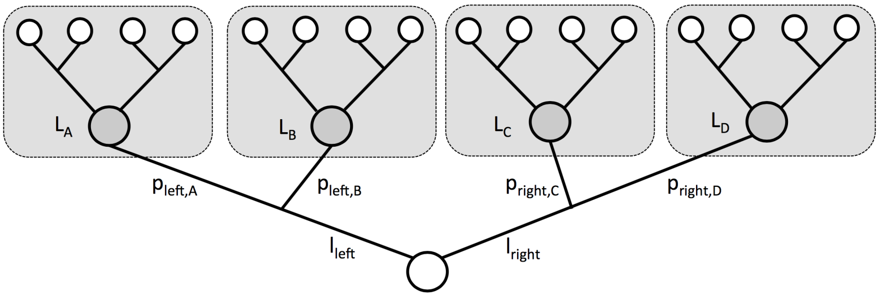
A Random Example

Below is a tree with a root node and four leaf nodes colored grey. At the route node you choose to go left or right. This incurs costs 4 and 2, respectively. Further, after making this decision there is a probability for reaching a leaf node. Namely, after going left the probabilities are 0.5 & 0.75, and for turning right, the probabilities are 0.25 & 0.75. For each leaf node there is there is a cost, namely, $2, 3, 6,$ and 1.



Given you only know the probabilities (and not what happens when you choose left or right), you'd want to take the decision with lowest expected cost. The expected cost for left is $4 + 0.5 \times 2 + 0.3 \times 3 = 5.5$ and for right is $2 + 0.25 \times 6 + 0.75 \times 1 = 4.25$. So go right.

Below we now replace the numbers above with symbols. At the route node you can choose the action to go left or right. These, respective, decisions incur costs of l_{left} and l_{right}. After choosing left, you will move to state A with probability $p_{\text{left},A}$ or to state B with probability $p_{\text{left},B}$ and similarly choosing right states C & D can be reached with probabilities $p_{\text{left},C}$ & $p_{\text{left},D}$. After reaching node A (resp. B,C,D) the total expected cost thereafter is L_A (resp. L_B, L_C, L_D).



Show that the optimal expected cost from the route node, L_R satisfies

$$L_R = \min_{a\in\{\text{left},\text{right}\}} \{l_a + \mathbb{E}_a\left[L_{X_a}\right]\}.$$

where here the random variable X_a denotes the state in $\{A, B, C, D\}$ reached after action is taken. The cost from choosing “left" is :

$$l_{\text{left}} + p_{\text{left},A}L_A + p_{\text{left},B}L_B = l_{\text{left}} + \mathbb{E}_{\text{left}}[L_{\text{left}}]$$

and the cost for choosing “right" is:

$$l_{\text{right}} + p_{\text{right},A}L_A + p_{\text{right},B}L_B = l_{\text{right}} + \mathbb{E}_{\text{right}}[L_{\text{right}}].$$

The optimal cost is the minimum of these two is

$$L_R = \min_{a\in\{\text{left},\text{right}\}} \{l_a + \mathbb{E}_a\left[L_{X_a}\right]\}.$$

where here the random variable X_a denotes the state in $\{A, B, C, D\}$ reached after action is taken. Notice how we abstracted away the future behaviour after arriving at A, B, C, D. Into a single cost for each state: L_A, L_B, L_C, L_D. And we can propagate this back to get the costs at the route state R. I.e. we can essentially apply the same principle as dynamic programming here.

Definitions

A Markov Decision Process (MDP) is a Dynamic Program where the state evolves in a random (Markovian) way.

Def 3 (Markov Decision Process)**.** *Like with a dynamic program, we consider discrete times* $t = 0, 1, ..., T$*, states* $x \in \mathcal{X}$*, actions* $a \in \mathcal{A}_t$ *and rewards* $r_t(a, x)$*. However, the plant equation and definition of a policy are slightly different.*

Like with a Markov chain, the state evolves as a random function of the current state and action, $f : \mathcal{X} \times \mathcal{A}_t \times [0, 1] \to \mathcal{X}$*. Here*

$$X_{t+1} = f(X_t, a_t; U_t) \equiv f(X_t, a_t)$$

where $(U_t)_{t\geq 0}$ *are IIDRVs uniform on* $[0, 1]$*. This is called the Plant Equation.*

A policy π *choses an action* π_t *at each time* t *as a function of past states* $x_0, ..., x_t$ *and past actions* $\pi_0, ..., \pi_{t-1}$*. We let* $\mathcal{P}$ *be the set of policies. A policy, a plant equation, and the resulting sequence of states and rewards describe a Markov Decision Process.*

As noted in the equivalence above, we will usually suppress dependence on U_t. Also, we will use the notation

$$\mathbb{E}_{x_t,a_t}[G(X_{t+1})] = \mathbb{E}[G(F_t(x_t, a_t; U))] \quad \text{and} \quad \mathbb{E}_{x,a}[G(\hat{X})] = \mathbb{E}[G(f(x, a; U))]$$

where here and here after we use $\hat{X}$ to denote the next state (after taking action a in state x). Notice in both equalities above, the term on the right depends on only one random variable, U.

Objective is to find a process that optimizes the expected reward.

Def 4 (Markov Decision Problem)**.** *Given initial state* x_0*, a Markov Decision Problem is the following optimization*

$$\begin{aligned} V(x_0) = &\textit{Maximize} \quad R_T(x_0, \Pi) := \mathbb{E}\left[\sum_{t=0}^{T-1} r_t(X_t, \pi_t) + r_T(X_T)\right] \qquad \text{(MDP)} \\ &\textit{over} \qquad \Pi \in \mathcal{P}. \end{aligned}$$

Further, let $R_\tau(x_\tau, \Pi)$ *(Resp.* $V_\tau(x_\tau)$*) be the objective (Resp. optimal objective) for* (MDP) *when the summation is started from time* $t = \tau$ *and state* $X_\tau = x_\tau$*, rather than* $t = 0$ *and* $X_0 = x_0$*. We often call* V *to value function of the MDP.*

The next result shows that the Bellman equation follows essentially as before but now we have to take account for the expected value of the next state.

Thrm 5 (Bellman Equation)**.** *Setting* $V_T(x) = r_T(x)$ *for* $t = T-1, T-2, ..., 0$

$$V_t(x_t) = \max_{a_t \in \mathcal{A}_t} \left\{ r_t(x_t, a_t) + \mathbb{E}_{x_t, a_t} \left[V_{t+1}(X_{t+1}) \right] \right\}. \qquad \text{(Bell eq.)}$$

The above equation is Bellman's equation for a Markov Decision Process.

Proof. Let $\mathcal{P}_t$ be the set policies that can be implemented from time t to T. Notice it is the product actions at time t and the set of policies from time $t+1$ onward. That is $\mathcal{P}_t = \{(\pi_t, \Pi) : \Pi \in \mathcal{P}_{t+1}, \pi_t : \mathcal{X}^t \times \prod_{\tau=0}^{t-1} \mathcal{A}_\tau \to \mathcal{A}_t\}$.

$$\begin{aligned} V_t(x_t) &= \max_{\Pi_t \in \mathcal{P}_t} \mathbb{E}_{x_t \pi_t} \left[\sum_{t=0}^{T-1} r_t(X_t, \pi_t) + r_T(X_T) \right] \\ &= \max_{\pi_t} \max_{\Pi \in \mathcal{P}_{t+1}} \left\{ r_t(x_t, \pi_t) + \mathbb{E}_{x_t \pi_t} \left[\mathbb{E}_{X_{t+1} \pi_{t+1}} \left[\sum_{\tau=t+1}^{T-1} r_\tau(X_\tau, \pi_\tau) + r_T(X_T) \right] \right] \right\} \\ &= \max_{a \in \mathcal{A}} \left\{ r_t(x_t, a) + \mathbb{E}_{x_t a} \underbrace{\left[\max_{\Pi \in \mathcal{P}_{t+1}} \mathbb{E}_{X_{t+1} \pi_{t+1}} \left[\sum_{\tau=t+1}^{T-1} r_\tau(X_\tau, \pi_\tau) + r_T(X_T) \right] \right]}_{= V_{t+1}(x_{t+1})} \right\} \end{aligned}$$

2nd equality uses structure of $\mathcal{P}_t$, takes the r_t term out and then takes conditional expectations. 3rd equality takes the supremum over $\mathcal{P}_{t+1}$, which does not depend on π_t, inside the expectation and notes the supremum over π_t is optimized at a fixed action $a \in \mathcal{A}$ (i.e. the past information did not help us.) □

Ex 6. *You need to sell a car. At every time $t = 0, \dots, T-1$, you set a price p_t and a customer then views the car. The probability that the customer buys a car at price p is $D(p)$. If the car isn't sold be time T then it is sold for fixed price V_T, $V_T < 1$. Maximize the reward from selling the car and find the recursion for the optimal reward when $D(p) = (1-p)_+$.*

Ex 7 (Call Option)**.** *You own a call option with strike price p. Here you can buy a share at price p making profit $X_t - p$ where x_t is the price of the share at time t. The share must be exercised by time T. The price of stock X_t satisfies*

$$X_{t+1} = X_t + \epsilon_t$$

for ϵ_t IIDRV with finite expectation. Show that there exists a decresing sequence $\{a_t\}_{0\leq t\leq T}$ such that it is optimal to exercise whenever $X_s \geq a_s$ occurs.

Ex 8. *You own an expensive fish. Each day you are offered a price for the fish according to a distribution density $f(x)$. You make the accept or reject this offer. With probability $1-p$ the fish dies that day. Find the policy that maximizes the profit from selling fish.*

Ex 9 (MDPs with Random Rewards)**.** *Consider an MDP where rewards are now random, i.e. after we have specified the state and action the reward $r(x,a)$ is still an independent random variable. Argue that this is the same as a MDP with non-random rewards given by*

$$\bar{r}(x,a) = \mathbb{E}[r(x,a)]$$

Ex 10 (Rewards that depend on the transition made)**.** *Consider an MDP where rewards depend both on the current state and the next state, as well as the action taken, i.e. rewards are of the form $r(x,\hat{x},a)$. Show that this is the same as a MDP with rewards*

$$\bar{r}(x,a) = \mathbb{E}_{x,a}[r(x,\hat{X},a]$$

Ans 10. *Apply [9].*

Ex 11. *Indiana Jones is trapped in a room in a temple. There are n passages that he can try and escape from. If he attempts to escape from passage $i \in \{1,...,n\}$ then either: he esacapes with probability p_p; he dies with probability q_i; or with probability $r_i = 1 - p_i - q_i$ the passage is a deadend and he returns to the room which he started from. Determine the order of passages which Indiana Jones must try in order to maximize his probability of escape.*

Ex 12. *You invest in properties. The total value of these properties is x_t in year $t = 1,...,T$. Each year t, you gain rent of rx_t and you choose to consume a proportion $a_t \in [0,1]$ of this rent. The remaining proportion is reinvested in buying new property. Further you pay mortgage payments of mx_t which are deducted from your consumed wealth. Here $m < r$. Your objective is to maximize the wealth consumed over T years.*

Briefly explaining why we can express this problem as a finite time dynamic program with

$$f(x,a) = x + rx(1-a), \qquad r_t(x,a) = x(ra - m), \qquad r_T(x) = 0\ ,$$

prove that if $W_{T-s}(x) = x\rho_s$ for some constant ρ_s then

$$\rho_s = \max\{r - m + \rho_{s-1}, (1+r)\rho_{s-1} - m\}.$$

3 Infinite Time Horizon

Thus far we have considered finite time Markov decision processes. We now want to solve MDPs of the form

$$V(x) = \underset{\Pi \in \mathcal{P}}{\text{maximize}} \quad R(x, \Pi) := \mathbb{E}_{x_0}\left[\sum_{t=0}^{\infty} \beta^t r(X_t, \pi_t)\right].$$

We can generalize Bellman's equation to infinite time, a correct guess at the form of the equation would, for instance, be

$$V(x) = \max_{a \in \mathcal{A}} \left\{ r(x, a) + \beta \mathbb{E}_{x,a}\left[V(\hat{X})\right] \right\}, \qquad x \in \mathcal{X}.$$

Previously we solved Markov Decisions Processes inductively with Bellman's equation. In infinite time, we can not directly apply induction; however, we see that Bellman's equation still holds and we can use this to solve our MDP. For now we will focus on the case of discounted programming: here we assume that

$$\max_{x \in \mathcal{X}, a \in \mathcal{A}} |r(x, a)| < \infty \qquad \text{and} \qquad \beta \in (0, 1).$$

We will cover cases where $\beta = 1$ later.

At this point it is useful define the concept of a Q-factor. A Q-factor of a policy π is the reward that arises when we take action a from state x and then follow policy π.

Def 13 (Q-Factor)**.** *The Q-factor of reward function $R(\cdot)$ is the value for taking action a in state x*

$$Q_R(x, a) = \mathbb{E}_{x,a}[r(x, a) + \beta R(\hat{X}))].$$

Similarly the Q-factor for a policy π, denoted by $Q_\pi(x, a)$, is given by the above expression with $R(x) = R(x, \pi)$. The Q-factor of the optimal policy is given by

$$Q^*(x, a) = \max_{\pi} Q_\pi(x, a).$$

The following result shows that if we have solved the Bellman equation then the solution and its associated policy is optimal.

Thrm 14. *For a discounted program, the optimal policy $V(x)$ satisfies*

$$V(x) = \max_{a \in \mathcal{A}} \left\{ r(x, a) + \beta \mathbb{E}_{x,a}\left[V(\hat{X})\right] \right\}.$$

Moreover, if we find a function $R(x)$ such that

$$R(x) = \max_{a\in\mathcal{A}} \left\{ r(x,a) + \beta\mathbb{E}_{x,a}\left[R(\hat{X})\right]\right\}$$

then $R(x) = V(x)$, i.e. the solution to the Bellman equation is unique, and we find a function $\pi(x)$ such that

$$\pi(x) \in \operatorname*{argmax}_{a\in\mathcal{A}} \left\{ r(x,a) + \beta\mathbb{E}_{x,a}\left[R(\hat{X})\right]\right\}$$

Then π is optimal and $R(x,\pi) = R(x) = V(x)$ the optimal value function.

Proof. We know that $R_t(x,\Pi) = r(x,\pi_0) + \beta\mathbb{E}[R_{t-1}(\hat{X},\hat{\Pi})]$. Applying limits as $t \to \infty$ on both sides and bounded convergence theorem gives that

$$\begin{aligned} R(x,\Pi) &= r(x,\pi_0) + \beta\mathbb{E}_{x,\pi_0}\left[R(\hat{X},\hat{\Pi})\right] \\ &\leq r(x,\pi_0) + \beta\mathbb{E}_{x,\pi_0}\left[V(\hat{X})\right]. \end{aligned}$$

For the inequality, above, we maximize $R(\hat{X},\hat{\Pi})$ over $\hat{\Pi}$. Now maximizing the left hand side over Π gives

$$V(x) \leq \sup_{\pi_0\in\mathcal{A}} \left\{ r(x,\pi_0) + \beta\mathbb{E}_{x,\pi_0}\left[V(\hat{X})\right]\right\}.$$

At this point we have that the Bellman equation but with an inequality. We need to prove the inequality in the other direction. For this, we let π_ϵ be the policy that chooses action a and then, from the next state $\hat{X}$, follows a policy $\hat{\Pi}_\epsilon$ which satisfies

$$R(\hat{X},\hat{\Pi}_\epsilon) \geq V(\hat{\Pi}) - \epsilon.$$

We have that

$$\begin{aligned} V(x) \geq R(x,\pi_\epsilon) &= r(x,a) + \beta\mathbb{E}_{x,a}\left[R(\hat{X},\hat{\Pi}_\epsilon)\right] \\ &\geq r(x,a) + \beta\mathbb{E}_{x,a}\left[V(\hat{X})\right] - \epsilon\beta \end{aligned}$$

The first inequality holds by the sub-optimality of Π_ϵ and the second holds by the assumption on $\hat{\Pi}_\epsilon$. Maximizing over $a \in \mathcal{A}$, and taking $\epsilon \to 0$ gives

$$V(x) \geq \max_{a\in\mathcal{A}} \left\{ r(x,a) + \beta\mathbb{E}_{x,a}\left[V(\hat{X})\right]\right\}.$$

Thus we now have that

$$V(x) = \max_{a\in\mathcal{A}} \Big\{ r(x,a) + \beta\mathbb{E}_{x,a}\Big[V(\hat{X}) \Big] \Big\}.$$

So at this point we know that the optimal value function satisfies the Bellman equation. For the next part of the result we need to show that the solution to this recursion is unique.

Suppose that $R(x)$ is another solution to the Bellman equation. From the definition of a Q-factor and the Bellman recursion, $R(x) = \max_a Q_R(x,a)$ and $V(x) = \max_a Q_V(x,a)$. Thus note that

$$Q_V(x,a) - Q_R(x,a) = \beta\mathbb{E}[V(\hat{X}) - R(\hat{X})] = \beta\mathbb{E}[\max_{a'} Q_V(\hat{X},a) - \max_{a'} Q_R(\hat{X},a')]$$

Thus

$$||Q_V - Q_R||_\infty \leq \beta \max_{\hat{x}} |\max_{a'} Q_V(\hat{x},a) - \max_{a'} Q_R(\hat{x},a')| \leq \beta||Q_V - Q_R||_\infty\,.$$

Since $0 < \beta < 1$ the only solution to this inequality is $Q_V = Q_R$ and thus

$$R(x) = \max_a Q_R(x,a) = \max_a Q_V(x,a) = V(x)\,.$$

So solutions to the Bellman equation are unique for discounted programming. Finally we must show that if we can find a policy that solves the Bellman equation, then it is optimal.

If we find a function $R(x)$ and a function $\pi(x)$ such that

$$R(x) = \max_{a\in\mathcal{A}} \Big\{ r(x,a) + \beta\mathbb{E}_{x,a}\Big[R(\hat{X}) \Big] \Big\}, \quad \pi(x) \in \operatorname*{argmax}_{a\in\mathcal{A}} \Big\{ r(x,a) + \beta\mathbb{E}_{x,a}\Big[R(\hat{X}) \Big] \Big\}$$

then note that the MDP induced by π is a Markov chain (with transition matrix $P^{\pi(x)}_{xy}$). Both $R(x,\pi)$ and $R(x)$ solve the equation $R(x) = r(x,\pi(x)) + \beta\mathbb{E}_{x,\pi(x)}[R(\hat{X})]$. So by Prop 58, $R(x) = R(x,\pi)$. □

It is worth collating together a similar result for Q-factors. Given the facts accrued about value function and Bellman's equation. The following Proposition should not be too great a surprise (and can be skipped on first reading).

Prop 15. *a) Stationary Q-factors satisfy the recursion*

$$Q_\pi(x,a) = \mathbb{E}_{x,a}[r(x,a) + \beta Q_\pi(\hat{X},\pi(\hat{X}))]\,.$$

b) Bellman's Equation can be re-expressed in terms of Q-factors as follows

$$Q^*(x,a) = \mathbb{E}_{x,a}[r(x,a) + \beta \max_{\hat{a}} Q^*(\hat{X},\hat{a}))]\,.$$

The optimal value function satisfies

$$V(x) = \max_{a \in \mathcal{A}} Q^*(x,a).$$

c) The operation

$$F_{x,a}(\boldsymbol{Q}) = \mathbb{E}_{x,a}[r(x,a) + \beta Q_\pi(\hat{X}, \pi(\hat{X}))]$$

is a contraction with respect to the supremum norm, that is,

$$\|\boldsymbol{F}(\boldsymbol{Q}_1) - \boldsymbol{F}(\boldsymbol{Q}_2)\|_\infty \leq \|\boldsymbol{Q}_1 - \boldsymbol{Q}_2\|_\infty .$$

Proof. a) We can think of extending the state space of our MDP to include states $\mathcal{X}_0 = \{(x,a) : x \in \mathcal{X}, a \in \mathcal{A}\}$ as well as $\mathcal{X}$. In this new MDP we can assume that initially the MDP starts in state (x,a) then moves to the state $\hat{X} \in \mathcal{X}$ according to the transition probabilities $P^a_{x\hat{x}}$. There after it remains in $\mathcal{X}$ moving according to policy π. Thus by Prop 58

$$Q_\pi(x,a) = \mathbb{E}_{x,a}[r(x,a) + \beta R(\hat{X}, \pi)]$$

where $R(x,\pi)$ is the reward function of policy π . Further since $Q_\pi(x,a)$ is the value from taking a instead of following policy π to should also be clear that

$$Q_\pi(x,\pi) = \mathbb{E}_{x,\pi(x)}[r(x,\pi(x)) + \beta R(\hat{X},\pi)] = R(x,\pi)$$

Thus, as required,

$$Q_\pi(x,a) = \mathbb{E}_{x,a}[r(x,a) + \beta Q_\pi(\hat{X}, \pi(\hat{X}))] .$$

b) Further it should be clear that the optimal value function for the extended MDP discussed has a Bellman equation of the form

$$\begin{aligned} Q^*(x,a) &= \mathbb{E}_{x,a}[r(x,a) + \beta V(\hat{X})] \\ V(x) &= \max_{a \in \mathcal{A}} \mathbb{E}_{x,a}[r(x,a) + \beta V(\hat{X})] \end{aligned}$$

Comparing the first equation above with the second, it should be clear that $V(x) = \max_a Q^*(x,a)$ and substituting this back into the first equation gives as required

$$Q^*(x,a) = \mathbb{E}_{x,a}[r(x,a) + \beta \max_{\hat{a} \in \mathcal{A}} Q^*(\hat{X}, \hat{a}))] .$$

c) The proof of this part is already embedded in the previous Theorem. Note that

$$F_{x,a}(\boldsymbol{Q}_1) - F_{x,a}(\boldsymbol{Q}_2) = \beta \mathbb{E}[\max_{a'} Q_V(\hat{X}, a) - \max_{a'} Q_R(\hat{X}, a')]$$

Thus

$$\|\boldsymbol{F}(\boldsymbol{Q}_1) - \boldsymbol{F}(\boldsymbol{Q}_2)\|_\infty \leq \beta \max_{\hat{x}} |\max_a Q_1(\hat{x}, a) - \max_{a'} Q_2(\hat{x}, a')| \leq \beta \|\boldsymbol{Q}_1 - \boldsymbol{Q}_2\|_\infty \,,$$

as required. □

4 Algorithms for MDPs

For infinite time MDPs, we cannot apply to induction on Bellman's equation from some initial state – like we could for finite time MDP. So we need some algorithms to solve MDPs.

At a high level, for a Markov Decision Processes (where the transitions P^a_{xy} are known), an algorithm solving a Markov Decision Process involves two steps:

- (Policy Improvement) Here you take your initial policy π_0 and find a new improved new policy π, for instance by solving Bellman's equation:

$$\pi(x) \in \operatorname*{argmax}_{a\in\mathcal{A}} \left\{ r(x,a) + \beta \mathbb{E}_{x,a}\left[R(\hat{X}, \pi_0) \right] \right\}$$

- (Policy Evaluation) Here you find the value of your policy. For instance by finding the reward function for policy π:

$$R(x,\pi) = \mathbb{E}^{\pi}_{x}\left[\sum_{t=0}^{\infty} \beta r(X_t, \pi(X_t)) \right]$$

Value iteration

Value iteration provides an important practical scheme for approximating the solution of an infinite time horizon Markov decision process.

Def 16 (Value iteration). *Take* $V_0(x) = 0\ \forall x$ *and recursively calculate*

$$\pi_{s+1}(x) \in \operatorname*{argmax}_{a\in\mathcal{A}} \left\{ r(x,a) + \beta \mathbb{E}_{x,a}\left[V_s(\hat{X}) \right] \right\}$$
$$V_{s+1}(x) = \max_{a\in\mathcal{A}} \left\{ r(x,a) + \beta \mathbb{E}_{x,a}\left[V_s(\hat{X}) \right] \right\}$$

for $s = 1, 2, ..$ *this is called* value iteration*.*

We can think of the two display equations above, respectively, as the policy improvement and policy evaluation steps. Notice, that we don't really need to do the policy improvement step to do each iteration. Notice the policy evaluation step evalutes one action under the new policy π afterwards the value is $V_s(\hat{X})$.

The following result shows that Value Iteration converges to the optimal policy.

Thrm 17. *For positive programming, i.e. where all rewards are positive and the discount factor β belongs to the interval $(0, 1]$, then*

$$0 \leq V_s(x) \leq V_{s+1}(x) \nearrow V(x), \quad \text{as} \quad s \to \infty\,.$$

Here $V(x)$ is the optimal value function.

The following lemma is the key property for value iterations convergence, as well as a number of other algorithms.

Lemma 1. *For reward function $R(x)$ define*

$$\mathcal{L}R(x) = \max_{a\in\mathcal{A}} \left\{ r(x,a) + \beta\mathbb{E}_{x,a}\left[R(\hat{X})\right]\right\}.$$

Show that if $R(x) \geq \tilde{R}(x)$ for all $x \in \mathcal{X}$ then $\mathcal{L}R(x) \geq \mathcal{L}\tilde{R}(x)$ for all $x \in \mathcal{X}$

Proof. Clearly,

$$r(x,a) + \beta\mathbb{E}_{x,a}\left[R(\hat{X})\right] \geq r(x,a) + \beta\mathbb{E}_{x,a}\left[\tilde{R}(\hat{X})\right].$$

Now maximize both sides over $a \in \mathcal{A}$. □

Proof of Thrm 17. Note that $V_1(x) = \max_a r(x,a) \geq 0 = V_0(x)$. Now, since $V_{s+1}(x) = \mathcal{L}V_s(x)$, repeatedly applying Lemma 1 to the inequality $V_1(x) \geq V_0(x)$ gives that

$$V_{s+1}(x) \geq V_s(x)\,.$$

Since $V_s(x)$ is increasing $V_s(x) \nearrow V_\infty(x)$ for some function V_∞. We must show that V_∞ is the optimal value function from the MDP.

Next note that $V_s(\cdot)$ is the optimal value function for the finite time MDP with rewards $r(x,a)$ and duration s. So $V(x) \geq V_s(x)$ and thus $V(x) \geq V_\infty(x)$. Further, for any policy Π,

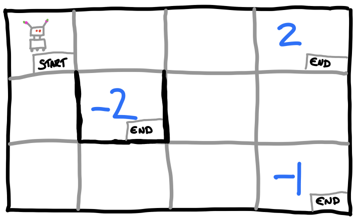
$$V_s(x) \geq R_s(x,\Pi)\,.$$

Now take limits $V_\infty(x) \geq R(x,\Pi)$. Now maximize over Π to see that $V_\infty(x) \geq V(x)$. So $V_\infty(x) = V(x)$ as required. □

Ex 18. *Show that for discounted programming,*

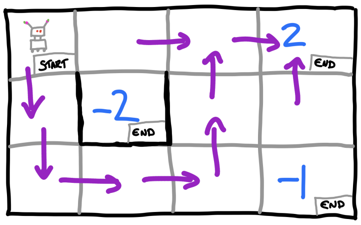
$$V_s(x) + \frac{\beta^{s+1} r_{\max}}{1-\beta} \geq V(x) \geq V_s(x) - \frac{\beta^{s+1} r_{\min}}{1-\beta}$$

Ex 19 (GridWorld)**.** *A robot is placed on the following grid.*



The robot can chose the action to move left, right, up or down provided it does not hit a wall, in this case it stays in the same position. (Walls are colored black.) With probability 0.8, the robot does not follow its chosen action and instead makes a random action. The rewards for the different end states are colored above. Write a program that uses, Value Iteration to find the optimal policy for the robot.

Ans 19. *Notice that the robot does not just take the shortest root. (I.e. some forward planning is required)*



Policy Iteration

We consider a discounted program with rewards $r(x,a)$ and discount factor $\beta \in (0,1)$.

Def 20 (Policy Iteration). *Given the stationary policy* Π, *we may define a new (improved) stationary policy,* $\mathcal{I}\Pi$, *by choosing for each* x *the action* $\mathcal{I}\Pi(x)$ *that solves the following maximization*

$$\mathcal{I}\Pi(x) \in \operatorname*{argmax}_{a\in\mathcal{A}} \; r(x,a) + \beta\mathbb{E}_{x,a}\left[R(\hat{X},\Pi)\right]$$

where $R(x,\Pi)$ *is the value function for policy* Π. *We then calculate* $R(x,\mathcal{I}\Pi)$. *Recall that, by Prop 58 for each* x *this solves the equation*

$$R(x,\mathcal{I}\Pi) = r(x,\mathcal{I}\Pi(x)) + \beta\mathbb{E}_{x,a}\left[R(\hat{X},\mathcal{I}\Pi)\right]$$

Policy iteration is the algorithm that takes

$$\Pi_{n+1} = \mathcal{I}\Pi_n$$

Starting from a stationary policy Π_0.

Thrm 21. *Under Policy Iteration*

$$R(x,\Pi_{n+1}) \geq R(x,\Pi_n)$$

and, for bounded programming,

$$R(x,\Pi_n) \nearrow V(x) \qquad as \quad n \to \infty$$

Proof. By the optimality of $\mathcal{I}\Pi$ with respect to Π we have

$$R(x,\Pi) = r(x,\pi(x)) + \beta\mathbb{E}_{x,\pi(x)}\left[R(\hat{X},\Pi)\right] \leq r(x,\mathcal{I}\Pi(x)) + \beta\mathbb{E}_{x,\mathcal{I}\pi(x)}\left[R(\hat{X},\Pi)\right]$$

Thus from the last part of Thrm 58, we know that $R(x,\Pi) \leq R(x,\mathcal{I}\Pi)$. This show that Policy iteration improves solutions. Now we must show it improves to the optimal solution.

First note that

$$\begin{aligned} r(x,a) + \beta\mathbb{E}_{x,a}\left[R(\hat{X},\Pi)\right] &\leq r(x,\mathcal{I}\pi(x)) + \beta\mathbb{E}_{x,\mathcal{I}(x)}\left[R(\hat{X},\Pi)\right] \\ &\leq r(x,\mathcal{I}\pi(x)) + \beta\mathbb{E}_{x,\mathcal{I}\Pi}\left[R(\hat{X},\mathcal{I}\Pi)\right] = R(x,\mathcal{I}\Pi). \end{aligned}$$

We can use the above inequality to show that the following process is a supermartingale

$$M_t = \beta^t R(X_t, \Pi_{T-t}) + \sum_{s=0}^{t-1} \beta^s r(X_s, \pi^*(X_s))$$

where $\pi^*(x)$ is the optimal policy.[1] To see taking expectations with respect to the optimal policy π^* gives

$$\begin{aligned}
&\mathbb{E}^* \left[M_{t+1} - M_t | \mathcal{F}_t \right] \\
&= \beta^t \mathbb{E}^* \left[\beta R(X_{t+1}, \Pi_{T-t-1}) + r(X_t, \pi^*(X)t) - R(X_t, \Pi_{T-t}) \middle| \mathcal{F}_t \right] \\
&= \beta^t \mathbb{E}^* \left[\beta \mathbb{E}^*_{X_t, \pi^*(X_t)} \left[\beta R(\hat{X}, \Pi_{T-t-1}) + r(X_t, \pi^*(X_t)) - R(X_t, \Pi_{T-t}) \right] \middle| \mathcal{F}_t \right] \\
&\leq 0\,.
\end{aligned}$$

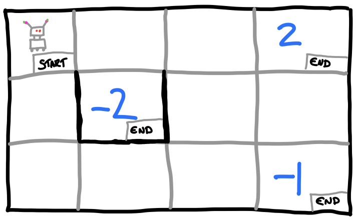
Since M_t is a supermartingale:

$$R(x, \Pi_T) = \mathbb{E}^*_x \left[M_0 \right] \geq \mathbb{E}^*_x \left[M_T \right] = \underbrace{\mathbb{E}^*_x \left[\beta^T R(X_T, \Pi_0) \right]}_{\xrightarrow[T\to\infty]{} 0} + \underbrace{R_T(x, \Pi^*)}_{\xrightarrow[T\to\infty]{} V(x)}$$

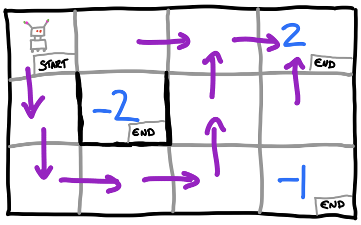
Therefore, as required, $\lim_{T\to\infty} R(x, \Pi_T) \geq V(x)$. □

[1] Note we are implicity assuming an optimal stationary policy exists. We can remove this assumption by considering a ϵ-optimal (non-stationary) policy. However, the proof is a little cleaner under our assumption.

Ex 22 (GridWorld, again)**.** *Write a program that uses, Policy iteration to find the optimal policy for the robot in [19].*



Ans 22.



5 Optimal Stopping

- Optimal Stopping Problems; One-Step-Look-Ahead Rule
- The Secretary Problem.
- Infinite Time Stopping

An Optimal Stopping Problem is an Markov Decision Process where there are two actions: $a = 0$ meaning to stop, and $a = 1$ meaning to continue. Here there are two types of costs

$$c(x, a) = \begin{cases} \kappa(x), & \text{for } a = 0 \quad \textit{(the stopping cost)} \\ c(x), & \text{for } a = 1 \quad \textit{(the continuation cost),} \end{cases}$$

This defines a *stopping problem.*

Assuming that time is finite, the Bellman equation is

$$C_s(x) = \min\left\{k(x), c(x) + \mathbb{E}_x[C_{s-1}(\hat{X})]\right\}$$

for $s \in \mathbb{N}$ and $C_0(x) = k(x)$.

Def 23 (OLSA rule). *In the one step lookahead (OSLA) rule we stop when ever $x \in S$ where*

$$S = \{x : k(x) \leq c(x) + \mathbb{E}_x[k(\hat{X})]\}.$$

We call S the stopping set. In words, you stop whenever it is better stop now rather than continue one step further and then stop.

Def 24 (Closed Stopping Set). *We say the set $S \subset \mathcal{X}$ is closed, it once inside that said you cannot leave, i.e.*

$$P_{xy} = 0, \qquad \forall x \in S, y \notin S.$$

Prop 25. *If, for the finite time stopping problem, the set S given by the one step lookahead rule is closed then the one step lookahead rule is an optimal policy.*

Proof. Given the set S is closed, we argue that if $C_{s-1}(x) = k(x)$ for $x \in S$ then $C_s(x) = k(x)$:If $x \in S$ then since S is closed $\hat{X} \in S$. In otherwords $C_{s-1}(\hat{X}) = k(\hat{X})$. Therefore, in this case, Bellman's equation becomes

$$C_s(x) = \min\{k(x), c(x) + \mathbb{E}_x[C_{s-1}(\hat{X})]\} = \min\{k(x), c(x) + \mathbb{E}_x\Big[k(\hat{X})\Big]\} = k(x).$$

The last inequality above follows by the definition of $x \in S$.

We now proceed by induction. The OSLA rule is optimal for $s = 1$ steps, since OSLA is exactly the optimal policy for one step.

Suppose that the result is holds for upto $s-1$ steps. Now consider the Optimal Stopping Problem with s steps. If $x \in S$ then $C_s(x) = k(x)$. So it is better to stop. If $x \notin S$, then clearly it's better to continue. □

Ex 26 (The Secretary Problem)**.** *There are N candidates for a secretary job. You interview candidates sequentially. After each interview, you must either accept or reject the candidate. We assume each candidate has the rank: $1, 2, ..., N$ And arrive for interview uniformly at random. Find the policy that maximises the probability that you hire the best candidate.*

Ans 26. *All that matters at each time is if the current candidate is the best so far. At time t let*

$$x_t = \begin{cases} 1, & \textit{if candidate is best so far,} \\ 0, & \textit{otherwise,} \end{cases} \qquad a_t = \begin{cases} 1, & \textit{if candidate accepted,} \\ 0, & \textit{if reject.} \end{cases}$$

Since is uniform random where the best candidate is

$$\mathbb{P}(t \textit{ is the best} \,|x_t = 1) = \frac{1/N}{1/t} = \frac{t}{N}.$$

Under the chosen policy, we let

$$R_t = \mathbb{P}(\textit{ select the best candidate} \mid \textit{rejected first } t)$$

be our reward function. Now

$$x_t = 1, \ \textit{w.p.} \ \frac{1}{t} \implies \begin{cases} a_t = 1 & \textit{gives expected reward } \frac{t}{N} \\ a_t = 0 & \textit{gives expected reward } R_{t+1} \end{cases}$$

$$x_t = 0, \ \textit{w.p.} \ 1 - \frac{1}{t} \implies \begin{cases} a_t = 1 & \textit{gives expected reward } 0 \\ a_t = 0 & \textit{gives expected reward } R_{t+1}. \end{cases}$$

Thus the Bellman equation for the above problem is

$$\begin{aligned}
R_t &= \underbrace{\frac{1}{t}\max\left\{\frac{t}{N}, R_{t+1}\right\}}_{\max\{1/N, R_{t+1}/t\}} + \left(1-\frac{1}{t}\right)\underbrace{\max\left\{0, R_{t+1}\right\}}_{R_{t+1}} \\
&= \max\left\{\frac{1-t}{t}R_{t+1} + \frac{1}{N}R_{t+1}\right\} \\
&= \begin{cases} R_{t+1}, & \textit{if } R_{t+1} \geq \frac{t}{N}, \\ \frac{t-1}{t}R_{t+1} + \frac{1}{N}, & \textit{if } R_{t+1} < \frac{t}{N}. \end{cases}
\end{aligned}$$

Notice that $R_t \geq R_{t+1}$. *Let* t^* *be the smallest* t *such that* $R_{t^*} \geq (t^*-1)/N$. *Starting from* $R_{N+1} = 0$ *note that so long as* $R_{t+1} < \frac{t}{N}$ *holds in second case in the above expression, we have that*

$$\frac{R_t}{t-1} = \frac{R_{t-1}}{t} + \frac{1}{N}\frac{1}{t-1}$$

Thus

$$\frac{R_{t^*}}{t^*-1} = \sum_{t=t^*}^{N} \frac{R_t}{t-1} - \frac{R_{t+1}}{t} = \sum_{t=t^*}^{N} \frac{1}{N}\frac{1}{t-1}.$$

Thus our condition for the optimal t^* *is to take the smallest* t^* *such that*

$$\sum_{t=t*}^{N} \frac{1}{t-1} \geq 1.$$

In other words, the optimal policy is to interview the first t^* *candidates and then accept the next best candidate.*

Ex 27 (The Secretary Problem, continued)**.** *Argue that as* $N \to \infty$, *the optimal policy is to interview* e^{-1} *of the candidates and then to accept the next best candidate.*

Ans 27. *From [26], the optimal condition is*

$$\sum_{t=t*}^{N} \frac{1}{t-1} \geq 1.$$

We know that as $N \to \infty$

$$\sum_{t=t^*}^{N} \frac{1}{t-1} \sim \int_{t^*}^{N} \frac{1}{t}dt = \log N - \log t^*.$$

Thus $\log N - \log t^* \geq 1$ *for* $t^*/N = e^{-1}$.

Ex 28. *You look for a parking space on street, each space is free with probability $p = 1 - q$. You can't tell if space is free until you reach it. Once at space you must decide to stop or continue. From position s (s spaces from your destination), the cost of stopping is s. The cost of passing your destination without parking is D.*

Ans 28. *Let*

$$x_s = \begin{cases} 1, & \text{if space } s \text{ is free,} \\ 0, & \text{otherwise.} \end{cases}$$

Here stopping means take the next free parking space. The Bellman equation is

$$\begin{aligned} C_s(0) &= pC_{s-1}(1) + qC_{s-1}(0) \\ C_s(1) &= \min\{s, pC_{s-1}(1) + qC_{s-1}(0)\} \end{aligned}$$

Consider the stopping set

$$S = \{s : s \leq k(s-1)\}$$

where $k(s)$ is the cost of taking the next available space. Note that

$$k(s) = ps + qk(s-1) \quad \text{and} \quad k(0) = qD.$$

Recursion of this form have solution

$$k(s) = -\frac{q}{p} + s + (D + \frac{1}{p})q^{s+1}$$

Therefore

$$S = \{s : (Dp + 1)q^s \geq 1\}.$$

Since s is decreasing, this set if clearly closed. Therefore the optimal policy is to take the next available space once $(Dp + 1)q^s \geq 1$ holds.

Ex 29. *In a game show a contestant is asked a series of 10 questions. For each question $q = 1, ..., 10$ there is a reward r_q for answering the question correctly. With probability p_q the contestant answers the question correctly. After correctly answering a question, the contestant can choose to stop and take their total winnings home or they can continue to the next question $q + 1$. However, if the contestant answers a question incorrectly then the contestant looses all of their winnings. The probability of winning each round is decreasing and is such that the expected reward from each round, $p_q r_q$, is constant.*

i) Write down the Bellman equation for this problem.

ii) Using the One-Step-Look-Ahead rule, or otherwise, find the optimal policy of the contestant.

Ans 29. *The Bellman equation for this problem is*

$$R_t(x) = \max\{x, p_t R_{t+1}(x + r_t)\}$$

The stopping set for this problem is

$$S = \{(x,t) : x \geq p_t(x + r_t)\} = \{(x,t) : x \geq r_t p_t/(1 - p_t)\}$$

Since by assumption $r_t p_t = c$ and $p_t \searrow 0$ (therefore $c/1 - p_t$) is decreasing) and $x_t \nearrow$ the set S is closed. Thus the OSLA rule is optimal for this problem.

Ex 30. *A burglar robs houses over N nights. At any night the burglar may choose to retire and thus take home his total earnings. On the tth night house he robs has a reward r_t where r_t is an iidrv with mean $\bar{r}$. Each night the probability that he is caught is p and if caught he looses all his money. Find the optimal policy for the burglar's retirement. (Hint: OLSA)*

Ex 31 (Bruss' Odds Algorithm)**.** *You sequentially treat patients $t = 1, ..., T$ with a new trail treatment. The probability of success is $p_t = 1 - q_t$. We must minimize the number of unsuccessful treatments while treating all patients for which the trail is will be successful. (i.e. if we label 1 for success and 0 for failure, we want to stop on the last 1). Argue, using the One-Step-Look-Ahead rule that the optimal policy is the stop treating at t^* the largest integer such that*

$$\frac{p_{t^*}}{q_{t^*}} + ... + \frac{p_T}{q_T} \geq 1.$$

This procedure is called Bruss' Odds Algorithm.

Ex 32. *You own an asset that must be sold in T days. Each day you are offered a price for the asset according to a probability distribution density $f(x)$. You may the accept any offer that you have received so far. Once the asset is sold the money is invested in a bank account which multiplies the invested money by β^{-1} each day. Here $\beta \in (0,1)$. Your task is the maximize your profit at time T.*

Ans 32. *The stopping set for this problem is*

$$S = \{z : z \geq \beta\mathbb{E}[z \vee X]\} = \{z : 0 \geq \mathbb{E}[((\beta-1)z) \vee (X-z)]\}$$

Not that since $X-z$ and $(\beta-1)z$ are both decreasing the expectation above is decreasing and so the set S is closed. Consequently the OSLA rule is optimal. Thus it is optimal to stop at the first offer z for which $z \geq \alpha$ where α solves

$$\beta^{-1}\alpha = \alpha\mathbb{P}(X \leq \alpha) + \int_\alpha^\infty xf(x)dx$$

Optimal stopping in infinite time

We now give conditions for the one step look ahead rule to be optimal for infinite time stopping problems.

Prop 33. *If the following two conditions hold*

- $K = \max_x k(x) < \infty$
- $C = \min_x c(x) > 0$

then the One-Step-Lookahead-Rule is optimal.

Proof. Suppose that the optimal policy π stops at time τ then

$$(s+1)C\mathbb{P}(\tau > s) \leq \mathbb{E}\left[\left\{\sum_{t=0}^{\tau-1} c(x_t) + k(x_\tau)\right\}\mathbb{I}[\tau > s]\right] \leq k(x_0) \leq K.$$

Therefore if we follow optimal policy π but for the s time horizon problem and stop at s if $\tau \geq s$ then

$$L(x) \leq L_s(x) \leq L(x) + K\mathbb{P}(\tau > s) \leq L(x) + \frac{K^2}{C(s+1)} \xrightarrow[s\to\infty]{} L(x)$$

Thus $L_s(x) \to L(x)$.

As before (for the finite time problem), it is no optimal to stop if $x \notin S$ and for the finite time problem $L_s(x) = k(x)$ for all $x \in S$. Therefore, since $L_s(x) \to L(x)$, we have that $L(x) = k(x)$ for all $x \in S$ and there for it is optimal to stop for $x \in S$. □

Ex 34. *You own a "toxic" asset its value, x_t at time t, belongs to $\{1, 2, 3, ...\}$. The daily cost of holding the asset is x_t. Every day the value moves up to $x + 1$ with probability $1/2$ or otherwise remains the same at x. Further the cost of terminating the asset after holding it for t days is $C(1 - \alpha)^t$. Find the optimal policy for terminating the asset.*

The one step lookahead rule is not always the correct solution to an optimal stopping problem.

Def 35 (Concave Majorant)**.** *For a function $r : \{0, ..., N\} \to \mathbb{R}_+$ a concave majorant is a function G such that*

- $G(x) \geq \frac{1}{2}G(x - 1) + \frac{1}{2}G(x + 1)$
- $G(x) \geq r(x)$.

Prop 36 (Stopping a Random Walk)**.** *Let X_t be a symmetric random walk on $\{0, ..., N\}$ where the process is automatically stopped at 0 and N. For each $x \in \{0, ..., N\}$, there is a positive reward of $r(x)$ for stopping. We are asked to maximize*

$$\mathbb{E}[r(X_T)]$$

where T is our chosen stopping time. The optimal value function $V(x)$ is the minimal concave majorant, and that it is optimal to stop whenever $V(x) = r(x)$.

Proof. The Bellman equation is

$$R(x) = \max\left\{r(x), \frac{1}{2}R(x - 1) + \frac{1}{2}R(x + 1)\right\}$$

with $R(x) = r(0)$ and $R(N) = r(N)$. Thus the optimal value function is a concave majorant.

We will show that the optimal policy is the minimal concave majorant of $r(x)$. We do so by, essentially applying induction on value iteration. First $R_0(x) = 0 \leq G(x)$ for any concave majorant of $r(x)$. Now suppose that R_{s-1}, the function reached after $s - 1$ value iterations, satisfies $R_{s-1}(x) \leq G(x)$ for all x, then

$$\begin{aligned} R_s(x) &= \max\left\{r(x), \frac{1}{2}R_{s-1}(x - 1) + \frac{1}{2}R_{s-1}(x + 1)\right\} \\ &\leq \max\left\{r(x), \frac{1}{2}G(x - 1) + \frac{1}{2}G(x + 1)\right\} \\ &\leq \max\{r(x), G(x)\} = G(x). \end{aligned}$$

Since value iteration converges $R_s(x) \nearrow V(x)$, where $V(x)$ satisfies $V(x) \leq G(x)$, as required.

Finally observe that from the Bellman equation the optimal stopping rule is to stop whenever $V(x) = r(x)$ for the minimal concave majorant. □

6 Continuous Time Dynamic Programming

Discrete time Dynamic Programming was given in Section 1. We now consider the continuous time analogue.

Time is continuous $t \in \mathbb{R}_+$; $x_t \in \mathcal{X}$ is the state at time t; $a_t \in \mathcal{A}$ is the action at time t; Given function $f : \mathbb{R}_+ \times \mathcal{X} \times \mathcal{A}_t \to \mathcal{X}$, the state evolves according to a differential equation

$$\frac{dx_t}{dt} = f_t(x_t, a_t). \tag{1}$$

This is called the Plant Equation. A policy π chooses an action π_t at each time t. The (instantaneous) reward for taking action a in state x at time t is $r_t(a, x)$ and $r_T(x)$ is the reward for terminating in state x at time T.

Def 37 (Dynamic Program)**.** *Given initial state x_0, a dynamic program is the optimization*

$$\begin{aligned}
L(x_0) := \textit{Minimize} \quad & C(\mathbf{a}) := \int_0^T e^{-\alpha t} c_t(x_t, a_t) dt + e^{-\alpha T} c_T(x_T) && \text{(DP)}\\
\textit{subject to} \quad & \frac{dx_t}{dt} = f_t(x_t, a_t), && t \in \mathbb{R}_+\\
\textit{over} \quad & a_t \in \mathcal{A}, && t \in \mathbb{R}_+
\end{aligned}$$

Further, let $C_\tau(\mathbf{a})$ (Resp. $L_\tau(x_\tau)$) be the objective (Resp. optimal objective) for (6) *when the summation is started from $t = \tau$, rather than $t = 0$.*

When a minimization problem where we minimize loss given the costs incurred is replaced with a maximization problem where we maximize winnings given the rewards received. The functions L, C and c are replaced with notation W, R and r.

Def 38 (Hamilton-Jacobi-Bellman Equation)**.** *For a continuous-time dynamic program* (6)*, the equation*

$$0 = \min_{a \in \mathcal{A}} \{ c_t(x, a) + \partial_t L_t(x) + f_t(x, a) \partial_x L_t(x) - \alpha L_t(x). \} \tag{HJB}$$

is called the Hamilton-Jacobi-Bellman equation. It is the continuous time analogoue of the Bellman equation [2].

A Heuristic Derivation of the HJB Equation

We now argue why the Hamiliton-Jacobi-Bellman equation is a good candidate for the Bellman equation in continuous time.

A good approximation to the plant equation (1) is

$$x_{t+\delta} - x_t = \delta f_t(x_t, a_t) \tag{2}$$

for $\delta > 0$ small, and a good approximation for the objective is

$$C(\mathbf{a}) := \sum_{t \in \{0,\delta,\dots,(T-\delta)\}} (1-\alpha\delta)^{t/\delta} c_t(x_t,a_t)\delta + (1-\alpha\delta)^{t/\delta} c_T(x_T) \tag{3}$$

This follows from the definition of the Riemann Integral and we further use the fact that $(1-\alpha\delta)^{t/\delta} \to e^{-\alpha t}$ as $\delta \to 0$.

The Bellman equation for the discrete time dynamic program with objective (3) and plant equation (2) is

$$L_t(x) = \min_{a \in \mathcal{A}} \left\{ c_t(x,a)\delta + (1-\alpha\delta) L_{t+\delta}(x_t + \delta f_t(x,a)) \right\}$$

If we minus $L_t(x)$ from each side in this Bellman equation and then divide by δ and let $\delta \to 0$ we get that

$$0 = \min_{a \in \mathcal{A}} \left\{ c_t(x,a) + \partial_t L_t(x) + f_t(x,a)\partial_x L_t(x) - \alpha L_t(x)\,, \right\}$$

where here we note that, by the Chain rule,

$$\frac{(1-\alpha\delta)L_{t+\delta}(x+\delta f) - L_t(x)}{\delta} \xrightarrow[\delta \to 0]{} \partial_t L_t(x) + f_t(x,a)\partial_x L_t(x) - \alpha L_t(x).$$

Thus we derive the HJB equation as described above.

The following result shows that if we solve the HJB equation then we have an optimal policy.

Thrm 39 (Optimality of HJB). *Suppose that a policy Π has a value function $C_t(x,\Pi)$ that satisfies the HJB-equation for all t and x then, Π is an optimal policy.*

Proof. Using shorthand $C = C_t(\tilde{x}_t, \Pi)$:

$$\begin{aligned} -\frac{d}{dt}\left(e^{-\alpha t} C_t(\tilde{x}_t,\Pi)\right) &= e^{-\alpha t}\left\{ c_t(\tilde{x}_t,\tilde{\pi}_t) - \left[c_t(\tilde{x}_t,\tilde{\pi}_t) - \alpha C + f_t(\tilde{x}_t,\tilde{\pi}_t)\partial_x C + \partial_t C \right] \right\} \\ &\leq e^{-\alpha t} c_t(\tilde{x}_t,\tilde{\pi}_t) \end{aligned}$$

The inequality holds since the term in the square brackets is the objective of the HJB equation, which is *not* maximized by $\tilde{\pi}_t$. □

Linear Quadratic Regularization

Def 40 (LQ problem)**.** *We consider a dynamic program of the form*

$$\begin{aligned}
&\textit{Minimize} && \int_0^T [x_t Q x_t + a_t R a_t]\, dt + x_T Q_T x_T && \text{(LQ)}\\
&\textit{subject to} && \frac{dx_t}{dt} = A x_t + B a_t, && t \in \mathbb{R}_+\\
&\textit{over} && a_t \in \mathbb{R}^m, && t \in \mathbb{R}_+.
\end{aligned}$$

Here $x_t \in \mathbb{R}^n$ and $a_t \in \mathbb{R}^m$. A and B are matrices. Q and R symmetric positive definite matrices. This an Linear-Quadratic problem (LQ problem).

Def 41 (Riccarti Equation)**.** *The differential equation with*

$$\dot{\Lambda}(t) = -Q - \Lambda(t)A - A^\top \Lambda(t) + \Lambda(t)BR^{-1}B^\top \Lambda(t) \quad \textit{and} \quad \Lambda(T) = Q_T. \quad \text{(RicEq)}$$

is called the Riccarti equation.

Thrm 42. *For each time t, the optimal action for the LQ problem is*

$$a_t = -R^{-1}B^\top \Lambda(t) x_t\,,$$

where $\Lambda(t)$ is the solution to the Riccarti equation.

Proof. The HJB equation for an LQ problem is

$$0 = \min_{a \in \mathbb{R}^m} \{x^\top Q x + a^\top R a + \partial_t L_t(x) + (Ax + Ra)^\top \partial_x L_t(x)\}$$

We now “guess" that the solution to above HJB equation is of the form $L_t(x) = x^\top \Lambda(t) x$ for some symmetric matrix $\Lambda(t)$. Therefore

$$\partial_x L_t(x) = 2\Lambda(t)x \quad \text{and} \quad \partial_t L_t(x) = x^\top \dot{\Lambda}(t) x$$

Substituting into the Bellman equation gives

$$0 = \min_{a \in \mathbb{R}^n} \left\{ x^\top Q x + a^\top R a + x^\top \dot{\Lambda}(t) x + 2x^\top \Lambda(x)(Ax + Ba) \right\}.$$

Differentiating with respect to a gives the optimality condition

$$2Ra + 2x^\top \Lambda(t) B = 0$$

which implies

$$a = -R^{-1}B^\top \Lambda(t) x\,.$$

Finally substituting into the Bellman equation, above, gives the expression

$$0 = x^\top \left[Q + \dot{\Lambda}(t) + \Lambda(t)A + A^\top \Lambda(t) - \Lambda(t)BR^{-1}B^\top \Lambda(t) - Q \right] x \, .$$

Thus the solution to the Riccarti equation has a cost function that solves the Bellman equation and thus by Theorem 39 the policy is optimal. □

7 Diffusion Control Problems

We consider a continuous time analogue of Markov Decision Processes from Section 2.

Time is continuous $t \in \mathbb{R}_+$; $X_t \in \mathbb{R}^n$ is the state at time t; $a_t \in \mathcal{A}$ is the action at time t.

Def 43 (Plant Equation). *Given functions* $\mu_t(X_t, a_t) = (\mu_t^i(X_t, a_t) : i = 1, .., n)$ *and* $\sigma_t(X_t, a_t) = (\sigma_t^{ij}(X_t, a_t) : i = 1, .., n, j = 1, ..., m)$, *the state evolves according to a stochastic differential equation*

$$dX_t = \mu_t(X_t, a_t)dt + \sigma_t(X_t, a_t) \cdot dB_t$$

where B_t *is an* m*-dimensional Brownian motion. This is called the Plant Equation.*

A policy π chooses an action π_t at each time t. (We assume that π_t is adapted and previsible.) Let $\mathcal{P}$ be the set of policies. The (instantaneous) cost for taking action a in state x at time t is $c_t(a, x)$ and $c_T(x)$ is the cost for terminating in state x at time T.

Def 44 (Diffusion Control Problem). *Given initial state* x_0, *a dynamic program is the optimization*

$$L(x_0) := \underset{\Pi \in \mathcal{P}}{minimize}\ C(x_0, \Pi) := \mathbb{E}_{x_0}\left[\int_0^T e^{-\alpha t} c_t(X_t, \pi_t)dt + e^{-\alpha T} c_T(X_T)\right] \quad \text{(DCP)}$$

Further, let $C_\tau(x, \Pi)$ *(Resp.* $L_\tau(x)$*) be the objective (Resp. optimal objective) for* (DCP) *when the integral is started from time* $t = \tau$ *with* $X_t = x$, *rather than* $t = 0$ *with* $X_0 = x$.

Def 45 (Hamilton-Jacobi-Bellman Equation). *For a Diffusion Control Problem* (DCP), *the equation*

$$0 = \min_{a \in \mathcal{A}} \left\{ c_t(x, a) + \partial_t L_t(x) + \mu_t(x, a) \cdot \partial_x L_t(x) + \frac{1}{2}[\sigma^T \sigma] \cdot \partial_{xx} L_t(x) - \alpha L_t(x). \right\} \quad \text{(HJB)}$$

is called the Hamilton-Jacobi-Bellman equation.[2] *It is the continuous time analogue of the Bellman equation [2].*

[2] Here $[\sigma^{\mathsf{T}}\sigma] \cdot \partial_{xx} L_t(x)$ is the dot-product of the Hessian matrix $\partial_{xx} L_t(x)$ with $\sigma^{\mathsf{T}}\sigma$. I.e. we multiply component-wise and sum up terms.

Heuristic Derivation of the HJB equation

We heuristically develop a Bellman equation for stochastic differential equations using our knowledge of the Bellman equation for Markov decision processes, in Section 2 (Theorem 5) and our heuristic derivation of the stochastic integral in Section B. This is analogous to continuous time control in Section 6.

Perhaps the main thing to remember is that (informally) the HJB equation is

$$0 = \min_{\text{actions}} \{\text{"instantaneous cost"} + \text{"Drift term from Ito's Formula"}\}.$$

Here Ito's formula is applied to the optimal value function at time t, $L_t(x)$. This is much easier to remember (assuming you know Ito's formula).

We suppose (for simplicity) that X_t belongs to $\mathbb{R}$ and is driven by a one-dimensional Brownian motion. The plant equation in Def 43 is approximated by

$$X_{t+\delta} - X_t = \mu_t(X_t, \pi_t)\delta + \sigma_t(X_t, \pi_t)(B_{T+\delta} - B_t)$$

for small δ (recall (12)). Similarly the cost function in (DCP) can be approximated by

$$C_t(x, \Pi) \approx \mathbb{E}\Bigg[\sum_{t\in\{0,\delta,\dots,T-\delta\}} (1-\alpha\delta)^{\frac{t}{\delta}} c_t(X_t, \pi_t)\delta + (1-\alpha\delta)^{\frac{T}{\delta}} c_T(X_T)\Bigg].$$

This follows from the definition of a Riemann Integral and since $(1-\alpha\delta)^{\frac{t}{\delta}} \to e^{-\alpha t}$. The Bellman equation for this objective function and plant equation is satisfies

$$L_t(x) = \min_{a\in\mathcal{A}} \{c_t(x,a)\delta + (1-\alpha\delta)\mathbb{E}_{x,a}\left[L_{t+\delta}(X_{t+\delta})\right]\}.$$

or, equivalently,

$$0 = \min_{a\in\mathcal{A}} \left\{ c_t(x,a) + \frac{1}{\delta}\mathbb{E}_{x,a}\left[L_{t+\delta}(X_{t+\delta}) - L_t(x)\right] - \alpha\mathbb{E}_{x,a}\left[L_{t+\delta}(X_{t+\delta})\right] \right\}.$$

Now by Ito's formula $L_t(X_t)$ can be approximated by

$$\begin{aligned} &L_{t+\delta}(X_{t+\delta}) - L_t(X_t) \\ \approx &\left[\partial_t L + \mu_t(X_t, \pi_t)\cdot\partial_x L + \frac{\sigma_t(X_t,\pi_t)^2}{2}\partial_{xx} L\right]\delta + \partial_x L \cdot \sigma_t(X_t, \pi_t)\cdot(B_{t+\delta} - B_t)\end{aligned}$$

Thus

$$\frac{1}{\delta}\mathbb{E}_{x,a}\left[L_{t+\delta}(X_{t+\delta}) - L_t(x)\right] = \partial_t L + \mu_t(X_t, \pi_t) \cdot \partial_x L + \frac{\sigma_t(X_t, \pi_t)^2}{2}\partial_{xx} L$$

Substituting in this into the above Bellman equation and letting $\delta \to 0$, we get, as required,

$$0 = \min_{a \in \mathcal{A}} \left\{ c_t(x,a) + \partial_t L + \mu_t(x,a) \cdot \partial_x L + \frac{\sigma_t(x,a)^2}{2}\partial_{xx} L - \alpha L_t(x) \right\}.$$

The following gives a rigorous proof that the HJB equation is the right object to consider for a diffusion control problem.

Thrm 46 (Davis-Varaiya Martingale Prinicple of Optimality)**.** *Suppose that there exists a function $L_t(x)$ with $L_T(x) = e^{-\alpha T} c_T(x)$ and such that for any policy Π with states X_t*

$$M_t = L_t(X_t) + \int_0^t e^{-\alpha\tau} c_\tau(X_\tau, \Pi) d\tau$$

is a sub-martingale and, moreover that for some policy Π^*, M_t is a martingale then Π^* is optimal and

$$L_0(X_0) = \min_{\Pi \in \mathcal{P}} \mathbb{E}\left[\int_0^T e^{-\alpha\tau} c_\tau(X_\tau, \pi_\tau) d\tau + c_T(X_T)\right].$$

Proof. Since M_t is a sub-martingale for all Π, we have

$$L_0(X_0) = M_0 \le \mathbb{E}[M_T] = \underbrace{\mathbb{E}_{X_0}\Big[\int_0^T e^{-\alpha s} c_\tau(X_\tau, \Pi_\tau) d\tau + \underbrace{L_T(X_T)}_{C_T(X_T)}\Big]}_{C(x_0,\Pi)}$$

Therefore $L_0(X_0) \le C(X_0, \Pi)$ for all policies Π.

If M_t is a Martingale for policy Π^*, then by the same argument $L_0(X_0) = C(X_0, \Pi^*)$. Thus

$$C(X_0, \Pi^*) = L_0(X_0) \le C(X_0, \Pi)$$

for all policies Π and so Π^* is optimal, and it holds that

$$L_0(X_0) = \min_{\Pi \in \mathcal{P}} \mathbb{E}\left[\int_0^T e^{-\alpha\tau} c_\tau(X_\tau, \pi_\tau) d\tau + c_T(X_T)\right].$$

□

8 Merton Portfolio Optimization

- HJB equation for Merton Problem; CRRA utility solution; Proof of Optimality.
- Multiple Assets; Dual Value function Approach.

We consider a specific diffusion control problem. We focus on setting where there is one risky asset and one riskless asset, though we will see that much of the analysis passes over to multiple assets.

Def 47 (The Merton Problem – Plant Equation)**.** *In the Merton problem you wish to optimise your long run consumption. You may invest your wealth in a bank account receiving riskless interest r, or in a risky asset with value S_t obeying the following SDE*

$$dS_t = S_t\{\sigma dB_t + \mu dt\}$$

where each $B = (B_t : t \geq 0)$ is an independent standard Brownian motion.

Wealth $(W_t : t \geq 0)$ obeys the SDE

$$dW_t = \underbrace{r(W_t - n_t S_t)\,dt}_{\substack{\text{Wealth in}\\ \text{bank}}} + \underbrace{n_t dS_t}_{\substack{\text{Wealth in}\\ \text{asset}}} - \underbrace{c_t dt}_{\text{consumption}} \tag{4a}$$

$$= r(W_t - n_t \cdot S_t)\,dt + n_t \cdot dS_t - c_t dt \tag{4b}$$

You can control c_t your rate of consumption at time t and n_t the number of stocks the risky asset at time t. Also, we define $\theta_t = n_t S_t$ to be the wealth in the risky asset at time t.

Def 48 (The Merton Problem – Objective)**.** *Given the above plant equation,* (4)*, the objective is to maximize the long-term utility of consumption*

$$V(w_0) = \max_{(n_t,c_t)_{t\geq 0}\in\mathcal{P}(w_0)} \mathbb{E}\left[\int_0^\infty e^{-\rho t} u(c_t)dt\right].$$

Here ρ is a positive constant and $u(c)$ is a concave increasing utility function. The set $\mathcal{P}(w_0)$ is the set of policies given initial wealth w_0. Further, let $V(w,t)$ be the optimal objective with the integral starting for time t with $w_t = w$.

Prop 49. *The HJB equation for the Merton Problem can be written as*

$$0 = \max_c \{u(c) - c\partial_w V\} + \max_\theta \left\{\theta(\mu - r)\partial_w V + \frac{1}{2}\sigma^2\theta^2\partial_{ww}V\right\} - \rho V + rw\partial_w V$$

Here the optimal θ is given by

$$\theta^* = -\frac{\partial_w V}{\partial_{ww}V}\sigma^{-2}(\mu - r)$$

Proof. First we note that we can rewrite the SDE for W_t as follows:

$$\begin{aligned} dW_t &= r\left(W_t - n_t \cdot S_t\right)dt + \underbrace{n_t \cdot dS_t}_{=n_tS_t\mu dt + n_tS_t\sigma dB_t} - c_t dt \\ &= \left(rW_t + (\mu - r)\theta_t - c_t\right)dt + \theta_t\sigma dB_t\,. \end{aligned}$$

Further note that if we shift the value function time by τ, a factor $e^{-\rho t}$ comes out,

$$V(w,\tau) = \max_{(n_t,c_t)_{t\geq\tau}} \mathbb{E}\left[\int_\tau^\infty e^{-\rho t}u(c_t)dt\right] = \max_{(n_t,c_t)_{t\geq 0}} \mathbb{E}\left[\int_0^\infty e^{-\rho(t+\tau)}u(c_t)dt\right] = e^{-\rho\tau}V(w)\,.$$

So $V(w,t) = e^{-\rho t}V(x)$.

Recall that informally the HJB equation is

$$0 = \max_{\text{actions}} \{\text{"instantaneous cost"} - \rho V + \text{"Drift term from Ito's Formula"}\}\,.$$

Notice that if we apply Ito's formula to $V(W_t)$ we get that

$$\begin{aligned} dV(W_t) &= \partial_w V(W_t)dW_t + \frac{1}{2}\partial_{ww}V(W_t)d[W]_t \\ &= \partial_w V(W_t)\left[r\left(W_t - n_t \cdot S_t\right)dt + n_t \cdot dS_t - c_t dt\right] \\ &\quad + \partial_t V(W_t)dt + \frac{\theta^2\sigma^2}{2}\partial_{ww}V(W_t)dt \end{aligned}$$

Applying this to the above term gives as required

$$\begin{aligned} 0 &= \max_{\theta,c}\left\{u(c) - \rho V + \left(rw + \theta\cdot(\mu - r) - c\right)\partial_w V + \frac{1}{2}\sigma^2\theta^2\partial_{ww}V\right\} \\ &= \max_c \{u(c) - c\partial_w V\} + \max_\theta\left\{\theta(\mu - r) + \frac{1}{2}\sigma^2\theta^2\partial_{ww}V\right\} - \rho V + rw\partial_w V \end{aligned}$$

Differentiating the HJB equation w.r.t. θ gives

$$\sigma^2\theta\partial_{ww}V = -(\mu - r)\partial_w V.$$

Now rearrange for θ^*. □

Merton for CRRA Utility

We focus on the case of CRRA utility, that is:

$$u(c) = \frac{c^{1-R}}{1-R}$$

for $R > 0$. (See the discussion on utility functions, Section C.) Thus we wish to solve for

$$V(w_0) = \max_{(n_t,c_t)_{t\geq 0}} \mathbb{E}\left[\int_0^\infty e^{-\rho t}\frac{c_t^{1-R}}{1-R}dt\right].$$

Prop 50. *For a CRRA utility it holds that:*
a) The Value function takes the form

$$V(w) = \gamma\frac{w^{1-R}}{1-R}$$

for some position constant $\gamma > 0$.
b) The HJB equation is optimized by

$$\theta^* = \frac{w}{R}\sigma^{-2}(\mu - r),$$

$$c^* = \gamma^{-\frac{1}{R}}w \quad \textit{and} \quad \sup_c\{u(c) - c\partial_w V\} = \frac{R}{1-R}\gamma^{1-\frac{1}{R}}w^{1-R}.$$

c) The HJB equation is satisfied by parameters

$$\gamma^* = R^{-1}\left\{\rho + (R-1)\left(r + \frac{1}{2}\frac{\kappa^2}{R}\right)\right\}$$

where

$$\kappa = \sigma^{-1}(\mu - r).$$

Proof. a) Note that having a policy for initial wealth λw_0 is the same as having a policy of wealth w_0 and then multiplying each amount invested by λ:

$$\begin{aligned} V(\lambda w) &= \max_{(n_t,c_t)_{t\geq 0}\in\mathcal{P}(\lambda w_0)} \mathbb{E}\left[\int_0^\infty e^{-\rho t}\frac{c_t^{1-R}}{1-R}dt\right] \\ &= \max_{(n_t,c_t)_{t\geq 0}\in\mathcal{P}(w_0)} \mathbb{E}\left[\int_0^\infty e^{-\rho t}\frac{(\lambda c_t)^{1-R}}{1-R}dt\right] = \lambda^{1-R}V(w). \end{aligned}$$

Letting $\lambda = w^{-1}$ and $\gamma = (1-R)V(1)$ gives the result.

b) By part a), $\partial_w V(w) = \gamma w^{-R}$ and $\partial_{ww} V(w) = -R\gamma w^{-xR-1}$. So, by Prop 49,

$$\theta^* = -\frac{\partial_w V}{\partial_{ww} V}\sigma^{-2}(\mu - r) = \frac{w}{R}\sigma^{-2}(\mu - r)$$

Also,

$$\sup_c \{u(c) - c\partial_w V\} \implies u'(c) = \partial_w V = \gamma w^{-R}$$

which since $u'(c) = c^{-R}$, gives that $c^* = \gamma^{-\frac{1}{R}} w$. Further,

$$\begin{aligned}
\sup_c \{u(c) - c\partial_w V\} &= \frac{c^{*1-R}}{1-R} - c^* \partial_w V(c^*) \\
&= \frac{(\gamma^{-\frac{1}{R}} w)^{1-R}}{1-R} - (\gamma^{-\frac{1}{R}} w)\gamma w^{-R} \\
&= \frac{R}{1-R}\gamma^{1-\frac{1}{R}} w^{1-R},
\end{aligned}$$

as required.

c) Applying a) and b) to the HJB equation in Prop 49 gives

$$\begin{aligned}
0 &= \frac{R}{1-R}\gamma^{1-\frac{1}{R}} w^{1-R} - \frac{1}{2}\sigma^{-2}(\mu - r)^2 \frac{(\partial_w V)^2}{\partial_{ww} V} - \rho V + rw\partial_w V \\
&= \frac{R}{1-R}\gamma^{1-\frac{1}{R}} w^{1-R} - \frac{1}{2}\sigma^{2}(\mu - r)^2 \frac{(\gamma w^{-r})^2}{(-R\gamma w^{-R-1})} - \rho\gamma \frac{w^{1-R}}{1-R} \\
&= \gamma w^{1-R}\left[\frac{R}{1-R}\gamma^{\frac{1}{R}} + \frac{1}{2}\sigma^2 \frac{(\mu - r)^2}{R} - \frac{\rho}{1-R} + r\right].
\end{aligned}$$

Cancelling γw^{1-R} and rearranging gives the required for for γ. □

To summarize: we notice we have shown that the parameters

$$\theta^* = \frac{w}{R}\sigma^{-2}(\mu - r), \quad c^* = \gamma^{-\frac{1}{R}} w, \tag{5a}$$

$$\gamma^* = R^{-1}\left\{\rho + (R-1)\left(r + \frac{1}{2}\frac{\kappa^2}{R}\right)\right\}, \quad \kappa = \sigma^{-1}(\mu - r). \tag{5b}$$

give a solution to the HJB equation for the Merton problem. (Although we have not yet proven them to be optimal.)

We now give rigorous argument for the optimality of parameters c^*, θ^* and γ^* for the Merton problem with CRRA utility. (This section can be skipped if preferred.)

Thrm 51. *The parameters in* (5)*, above, are optimal for the Merton problem.*

Proof. Since $u(y)$ is concave, $u(y) \leq u(x) + (y - x)u'(x)$. Thus for $\zeta_t = e^{-\rho t}u'(c_t^*) \propto e^{-\kappa B_t-(r+\frac{1}{2}|\kappa|^2)t}$ we have that

$$\begin{aligned}\mathbb{E}\left[\int_0^\infty e^{-\rho t}u(c_t)dt\right] &\leq \mathbb{E}\left[\int_0^\infty e^{-\rho t}\left\{u(c_t^*) + (c_t - c_t^*)u'(c_t^*)\right\}dt\right]\\ &= \mathbb{E}\left[\int_0^\infty e^{-\rho t}u(c_t^*)dt\right] + \mathbb{E}\left[\int_0^\infty (c_t - c_t^*)\,\zeta_t dt\right] \qquad (6)\end{aligned}$$

Next we show that

$$Y_t = \zeta_t W_t + \int_0^t \zeta_s c_s ds$$

is a positive local martingale. It is clear that the function Y_t is positive. Note that

$$\zeta_t = e^{-\rho t}u'(c_t^*) = De^{-\kappa B_t-(r+\frac{\kappa^2}{2})t} \qquad \text{where} \qquad \gamma w_0.$$

Define function

$$f_t(W, B) = W\exp\left\{-\kappa B - \left(r + \frac{\kappa}{2}\right)\right\}$$

and note that $\zeta_t w_t = Df_t(W_t, B_t)$. Now lets apply Ito's formula to $f_t(W_t, B_t)$. By Ito's formula:

$$df = \partial_t f dt + \partial_w f dW_t + \partial_B f dB_t + \frac{1}{2}\partial_B B f d[B]_t + \partial_{Bw} f d[BW]_t + \frac{1}{2}\partial_{ww} f d[W]_t.$$

Now lets check terms.

$$\begin{aligned}&\partial_t f = -\Big(r + \frac{1}{2}\kappa^2\Big)We^{-\kappa B_t-(r+\frac{1}{2}\kappa^2)t} \quad \partial_B f = -\kappa We^{-\kappa B_t-(r+\frac{1}{2}\kappa^2)t} \quad \partial_w f = e^{-\kappa B_t-(r+\frac{1}{2}\kappa^2)t}\\ &\partial_{BB} f = \kappa^2 We^{-\kappa B_t-(r+\frac{1}{2}\kappa^2)t} \quad \partial_{wB} f = -\kappa e^{-\kappa B_t-(r+\frac{1}{2}\kappa^2)t} \quad \partial_{ww} f = 0\\ &\qquad d[B]_t = dt \quad d[W, B]_t = \theta\sigma dt\end{aligned}$$

Substituting these into Ito's formula above gives,

$$\begin{aligned}df = e^{-\kappa B_t-(r+\frac{1}{2}\kappa^2)t}\Bigg[&-W\Big(r + \frac{1}{2}\kappa^2\Big)dt + \left\{rw_t - c_t + \theta(\mu - r)\right\}dt\\ &+ \theta\sigma dB_t - W\kappa dB_t + \frac{W}{2}\kappa^2 dt - \theta\sigma\kappa dt\Bigg]\end{aligned}$$

Cancelling and rearrganging we get

$$df + e^{-\kappa B_t - (r + \frac{\kappa^2}{2} t} c_t dt = e^{-\kappa B_t - (r + \frac{1}{2}\kappa^2)t} \left[\theta\sigma - W\kappa dB_t\right]$$

So

$$\zeta_t W_t + \int_0^t \zeta_s W_s ds = Df_t(W_t, B_T) + \int_0^t e^{\kappa B_t - (r + \frac{\kappa^2}{2})t} dt$$

is a local-Martingale. Recall from stochastic integration theory that every positive local martingale is a supermartingale.

Doob's Martingale Convergence Theorem applied to Y_t gives

$$\zeta_0 w_0 = Y_0 \geq \mathbb{E} Y_\infty = \mathbb{E}\left[\int_0^\infty \zeta_s c_s ds\right]$$

Since $\zeta_t = e^{-\rho t} u'(c_t^*) = e^{-\rho t}(c_t^*)^{-R}$ and by the definition of $V(w_0)$:

$$\mathbb{E}_{w_0}\left[\int_0^\infty \zeta_s c_s^* ds\right] = \mathbb{E}_{w_0}\left[\int_0^\infty e^{-\rho t}(c_t^*)^{1-R} ds\right] = (1-R)V(w_0) {=} \gamma w_0^{1-R} = \zeta_0 w_0$$

The last equality holds since $\zeta_0 = (c_0^*)^{-R}$ and $c_s^* = \gamma^{1/R} w_s^*$.

Combining the last equality and the inequality before that, we see that

$$\mathbb{E}\left[\int_0^\infty (c_t - c_t^*)\, \zeta_t dt\right] \leq 0\,.$$

Applying this to (6) we see that

$$\mathbb{E}\left[\int_0^\infty e^{-\rho t} u(c_t) dt\right] \leq \mathbb{E}\left[\int_0^\infty e^{-\rho t} u(c_t^*) dt\right]$$

and, as required, c_t^* is optimal. □

Merton Portfolio Optimization with Multiple Assets

We now note how the above results extend to the case where there aren't many assets. Now suppose that there are d assets that can be in invested in. These obey the Stochastic Differential Equation

$$dS_t^i = S_t^i \left\{ \sum_{j=1}^N \sigma^{ij} dB_t^j + \mu^i dt \right\}, \quad i = 1, \ldots, d$$

where B^j_t is an independent Brownian motion for each $j = 1, ..., N$.

Wealth now evolves according the SDE

$$dW_t = r\,(W_t - \boldsymbol{n}_t \cdot \boldsymbol{S}_t)\,dt + \boldsymbol{n}_t \cdot d\boldsymbol{S}_t - c_t dt$$

where $\boldsymbol{n}_t = (n^i_t : i = 1, ..., d)$ gives the amount of each asset $\boldsymbol{S}_t = (S^i_t : i = 1, ..., d)$ held in the portfolio at time t. Also we define $\boldsymbol{\theta}_t = (n^i_t S^i_t : i = 1, ..., d)$ as the wealth in each asset. As given in Def 48, our task is the maximize the objective

$$V(w) = \max_{(\boldsymbol{n}_t, c_t)_{t\geq 0} \in \mathcal{P}(w_0)} \mathbb{E}\left[\int_0^\infty e^{-\rho t} u(c_t) dt\right].$$

We now proceed through exercises that are very similar to the case with a single risky asset. We go through the proofs somewhat quickly.

Lemma 2. *Show that the HJB equation for the Merton Problem can be written as*

$$0 = \max_c \{u(c) - c\partial_w V\} + \max_{\boldsymbol{\theta}} \left\{\boldsymbol{\theta} \cdot (\boldsymbol{\mu} - \boldsymbol{r})\partial_w V + \frac{1}{2}|\sigma\boldsymbol{\theta}|^2 \partial_{ww} V\right\} - \rho V + rw\partial_w V.$$

where $\boldsymbol{r} = (r : i = 1, ..., d)$.

Proof. The proof follows more-or-less identically to Prop 49. Note that in this case Ito's formula applied to $V(W_t)$ gives

$$dV(W_t) = \partial_w V(W_t) dW_t + \frac{1}{2}\partial_{ww} V(W_t) d[W]_t$$

where

$$dW_t = \Big(rW_t - \underbrace{r\boldsymbol{n}_t \cdot \boldsymbol{S}_t}_{\boldsymbol{\theta}\cdot\boldsymbol{r}}\Big)dt + \underbrace{\boldsymbol{n}_t \cdot d\boldsymbol{S}_t}_{\boldsymbol{\theta}^\top[\sigma d\boldsymbol{B}_t + \boldsymbol{\mu} dt]} - c_t dt = (rW_t + \boldsymbol{\theta}_t(\boldsymbol{\mu} - \boldsymbol{r}) - c_t)\,dt + \boldsymbol{\theta}_t^\top \sigma d\boldsymbol{B}_t$$

$$d[W]_t = \sum_{ij} (\boldsymbol{\theta}_t^\top \sigma)_i (\boldsymbol{\theta}_t^\top \sigma)_j d[B^i_t, B^j_t] = |\boldsymbol{\theta}_t \sigma|^2 dt.$$

Thus

$$dV(W_t) = \left[(rW_t + \boldsymbol{\theta}_t(\boldsymbol{\mu} - \boldsymbol{r}) - c_t)\,\partial_w V(W_t) + \frac{1}{2}|\boldsymbol{\theta}\sigma|^2 \partial_{ww} V(W_t)\right] dt + \boldsymbol{\theta}_t^\top \sigma d\boldsymbol{B}_t.$$

This is the drift term applied in the HJB equation. Thus recalling that

$$0 = \max_{\text{actions}} \{\text{"instantaneous cost"} + \text{"Drift term from Ito's Formula"}\}.$$

This gives the require HJB equation. □

Lemma 3. *Show the optimal asset portfolio in the HJB equation is given by*

$$\boldsymbol{\theta}^* = -\frac{\partial_w V}{\partial_{ww} V}(\sigma\sigma^\top)^{-1}(\boldsymbol{\mu} - \boldsymbol{r})$$

and

$$\max_{\boldsymbol{\theta}} \left\{ \boldsymbol{\theta} \cdot (\boldsymbol{\mu} - \boldsymbol{r})\partial_w V + \frac{1}{2}(\boldsymbol{\theta}^\top \sigma^\top \sigma \boldsymbol{\theta})\partial_{ww} V \right\} = -\frac{1}{2}|\boldsymbol{\kappa}|^2 \frac{(\partial_w V)^2}{\partial_{ww} V}$$

Proof. Considering Lemma 2 we have that

$$\max_{\boldsymbol{\theta}} \left\{ \boldsymbol{\theta} \cdot (\boldsymbol{\mu} - \boldsymbol{r})\partial_w V + \frac{1}{2}(\boldsymbol{\theta}^\top \sigma^\top \sigma \boldsymbol{\theta})\partial_{ww} V \right\} \implies (\boldsymbol{\mu} - \boldsymbol{r})\partial_w V + \partial_{ww} V(\sigma^\top \sigma)\boldsymbol{\theta}^* = 0.$$

Solving for $\boldsymbol{\theta}^*$ and substituting back into the maximization gives the answer. □

Lemma 4. *Show that for a CRRA utility the optimal solution to the HJB equation is given by*

$$\boldsymbol{\theta}^* = \frac{w}{R}(\sigma\sigma^\top)^{-1}(\boldsymbol{\mu} - \boldsymbol{r}), \quad c^* = \gamma^{-\frac{1}{R}} w$$

where

$$\gamma^{-\frac{1}{R}} = R^{-1} \left\{ \rho + (R-1)(r + \frac{1}{2}\frac{|\boldsymbol{\kappa}|^2}{R}) \right\} \qquad \boldsymbol{\kappa} = \sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{r}).$$

Proof. 4 (In this proof when we refer to Prop 50 we mean that the argument which was applied in the single-asset setting is identical in the multiple asset setting.)

By Prop 50a)

$$V(w) = \gamma \frac{w^{1-R}}{1-R}$$

for some constant γ. Differentiating twice gives

$$\boldsymbol{\theta}^* = -\frac{\partial_w V}{\partial_{ww} V}(\sigma\sigma^\top)^{-1}(\boldsymbol{\mu} - \boldsymbol{r}) = \frac{w}{R}(\sigma\sigma^\top)^{-1}(\boldsymbol{\mu} - \boldsymbol{r}).$$

By Prop 50b), $c^* = \gamma^{-\frac{1}{R}} w$. Substituting these solutions into the HJB equation gives

$$\begin{aligned} 0 &= \max_c \{u(c) - c\partial_w V\} + \max_{\boldsymbol{\theta}} \left\{ \boldsymbol{\theta} \cdot (\boldsymbol{\mu} - \boldsymbol{r})\partial_w V + \frac{1}{2}|\sigma\boldsymbol{\theta}|^2 \partial_{ww} V \right\} - \rho V + rw\partial_w V \\ &= \frac{R}{1-R}\gamma^{1-\frac{1}{R}} w^{1-R} + \frac{1}{2}|\boldsymbol{\kappa}|^2 \cdot \frac{\gamma}{R} w^{1-R} - \rho\gamma \frac{w^{1-R}}{1-R} + r\gamma w^{1-R} \end{aligned}$$

Rearranging and solving for γ gives the required solution for γ^*. □

Def 52 (Merton Portfolio and Market Price Risk Vector)**.** *As given above,*

$$\boldsymbol{\theta}^* = \frac{w}{R}(\sigma\sigma^\top)^{-1}(\boldsymbol{\mu} - \boldsymbol{r}),$$

is called the Merton Portfolio *and*

$$\boldsymbol{\kappa} = \sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{r}).$$

is called the Market Price Risk Vector*.*

Dual value function approach

We could solve the CRRA utility case because it had a special structure. We now give a method for solving for general utilities $u(t, c)$.

Here we assume that $u(t, c)$ is continuous in t and c, concave in c and satisfies

$$\lim_{c\to\infty} u'(t, c) = 0$$

The HJB equation for the Merton problem is

$$0 = \max_{\theta, c}\left\{u(t, c) + \partial_t V + (rw + \theta \cdot (\mu - r) - c)\,\partial_w V + \frac{1}{2}|\sigma^\top\theta|^2\partial_{ww} V\right\}.$$

We take the LF transform of u,

$$u^*(t, z) = \max_c\{u(t, c) - zc\}$$

Further we define

$$J(t, z) = V(t, w) - wz$$

where w is such that $z = \partial_w V(t, w)$.

Thrm 53. *The HJB equation can be written as*

$$0 = u^*(t, c) + \partial_t J - rz\partial_z J + \frac{1}{2}|\kappa|^2 z^2 \partial_{zz} J$$

Moreover if we suppose that $u(t, x) = e^{-\rho t}u(x)$, for $u(x)$ concave and increasing, the HJB equation becomes

$$0 = u^*(y) - \rho j(y) + (\rho - r)yj'(y) + \frac{1}{2}|\kappa|^2 y^2 j''(y)$$

Noticed in the first HJB equation above that we have got rid of the maximization and in the second we have a linear ODE, which can be solved using standard methods.

Proof. First we will show that

$$\partial_z J = -w, \qquad \partial_{zz} J = -\frac{1}{\partial_{ww} V}, \qquad \partial_t J = \partial_t V \tag{7}$$

We can ignore the dependence of t for the first two expressions i.e. take $J(t,z) = J(z)$. Now $J(z) = V((V')^{-1}(z)) - z(V')^{-1}(z)$, so

$$J'(z) = \frac{d}{dz}(V')^{-1}(z) \cdot \underbrace{V'((V')^{-1}(z))}_{=z} - z\frac{d}{dz}(V')^{-1}(z) - (V')^{-1}(z) = -(V')^{-1}(z) = -w$$

and

$$J''(z) = -\frac{d}{dz}(V')^{-1}(z) = -\frac{1}{V''((V')^{-1}(z))} = -\frac{1}{V''(w)}.$$

Now reintroducing dependence on t,

$$\frac{\partial J}{\partial t} = \partial_t V(t,w) + \frac{dw}{dt}\underbrace{\partial_w V}_{=z} - \frac{dw}{dt}z = \partial_t V$$

This gives the required derivatives in (7).

Substituting the expressions in (7), the HJB equation is

$$\begin{aligned}
0 &= \max_c \{u(t,z) - c\partial_w V\} + \max_{\boldsymbol{\theta}} \left\{\boldsymbol{\theta} \cdot (\boldsymbol{\mu} - \boldsymbol{r})\partial_w V + \frac{1}{2}|\sigma\boldsymbol{\theta}|^2 \partial_{ww} V\right\} + \partial_t V + rw\partial_w V \\
&= u^*(t, \partial_w V) - \frac{1}{2}|\boldsymbol{\kappa}|^2 \frac{(\partial_w V)^2}{\partial_{ww} V} + \partial_t V + rw\partial_w V \\
&= u^*(t,z) + \frac{1}{2}|\boldsymbol{\kappa}|^2 z^2 \partial_{zz} J + \partial_t J - rz\partial_z J.
\end{aligned}$$

Now is we suppose that $u(t,x) = e^{-\rho t}u(x)$, for $u(x)$ concave and increasing, then by the same argument as Prop 50a) we have that

$$V(t,w) = e^{-\rho t}V(w).$$

Defining $j(z) = V(w) - wz$ where w is such that $z = \partial_w V(t,w)$, the following are straightforward calculations:

$$\begin{aligned}
J(t,z) &= e^{-\rho t} j(y), \qquad & \partial_t J &= -\rho e^{-\rho t} j(y) + \rho e^{-\rho t} y j'(y) \\
\partial_z J &= j'(y), & \partial_{zz} J &= e^{\rho t} j''(y)
\end{aligned}$$

where $y = ze^{\rho t}$. Now substituting these terms into the HJB equation gives the result:

$$0 = \underbrace{u^*(t,z)}_{=e^{-\rho t}u^*(y)} + \underbrace{\partial_t J}_{=-\rho e^{-\rho t}j(y)+\rho e^{-\rho t}yj'(y)} - \underbrace{rz\partial_z J}_{re^{-\rho t}yj'(y)} + \underbrace{\frac{1}{2}|\kappa|^2 z^2 \partial_{zz} J}_{\frac{1}{2}|\kappa|^2 y^2 e^{-\rho t} j''(y)} .$$

□

9 Principles of Reinforcement Learning

First we discuss at a high level a few of the key concepts in Reinforcement learning. These will then be discussed in more precise mathematical detail for specific examples and algorithms in subsequent sections.

Reinforcement Learning: Reinforcement Learning is the setting where we do not know the transition probabilities of a Markov Decision Process (or we might want to approximate a control problem with MDP). For instance, you might be able to simulate a problem with states, actions and rewards but you do not have access to the underlying dynamics of the simulation. Enough information must be gathered to approximate the optimal action for each state.

Policy Evaluation and Policy Improvement: When we look at reinforcement learning algorithms the same principles that applied to MDPs (with known transition probabilities) apply. I.e. we might want to think of the steps of the algorithm either improving the policy:

$$\pi(x) \in \operatorname*{argmax}_{a \in \mathcal{A}} \left\{ r(x,a) + \beta \mathbb{E}_{x,a} \left[R(\hat{X}, \pi_0) \right] \right\}$$

or evaluating the reward function of the current policy

$$R(x,\pi) = \mathbb{E}_x^{\pi} \left[\sum_{t=0}^{\infty} \beta r(X_t, \pi(X_t)) \right] .$$

Although algorithms might be subject to more noisy estimates.

Exploration-Exploitation trade-off: Because transition probabilities are unknown, when you are at a state, say x, there is a question of whether you should perform the best action a^* given the available information and thus attempt to implement the best known policy; or if you should chose a different (possibly random) action and thus get better information about the value of that action. I.e. there is a trade-off between doing what is myopically best given the available information (exploitation) and trying something new incase it might be better (exploration). (This is similar to policy evaluation and improvement, but here we are interested in finding the statistical properties of each action rather than performing computations on a function.) Problems that investigate exploration and exploitation tradeoff in isolation are often called Multi-armed Bandit problems, and there is a vast recent literature on these topics as well as a very well developed theoretical basis preceding this.

Model Free Control: Here we are especially interested in methods that are *model free*. A method is model free when it does not require an explicit estimation of the system dynamics, specifically, we don't try to estimate the transition probabilities P^a_{xy} for each action. For instance, if we perform policy improvement based on an estimation of the value function to V,

$$\pi(x) \in \operatorname*{argmax}_{a\in\mathcal{A}} \quad r(x,a) + \sum_{\hat{x}} P^a_{x\hat{x}} V(\hat{x})$$

then this is not model free, because we need to estimate $P^a_{x\hat{x}}$ in addition to our estimate of the value function V. Instead we might consider the Q-function of the MDP. This is the function $Q(x,a)$ which gives the value function for taking action a in state x and then afterward follow the optimal policy. If we perform policy improvement based on an estimation of the Q-function

$$\pi(x) \in \operatorname*{argmax}_{a\in\mathcal{A}} \quad Q(x,a)$$

then this is model free. We will discuss this in more detail in the next section.

Function Approximation: If the set of state and actions is moderately small then we can store functions of interest such as the Q-function $Q(x,a)$ as a table (or matrix) in computer memory. These algorithms are often called table based methods. But for larger problems or of problems with continuous state spaces and action spaces, then it is not possible to store this information. Further the likelihood of revisiting exactly the same state twice is vastly reduced. So often we have to infer relationships between states that are "close" and hope that the value function is suitably continuous that this forms a good approximation. So here we might for instance replace the value function $Q(x,a)$ with some approximation $Q_w(x,a)$ which is of lower dimension than $Q(x,a)$. Here w represents a weights that we use to parameterize our approximation (e.g. we could approximate continuous real valued function with a polynomial). Then we might look to find the best approximation:

$$\min_{w} \mathbb{E}[(\hat{Q}(x,a) - Q_w(x,a))^2] \,.$$

Here we let $\hat{Q}(x,a)$ be the Q-values of the current policy as observed from the data seen so far and we look to find the weights that give the best approximation. Above we minimize the mean-squared-error of the loss function, but we could consider other metrics and we could approximate other functions e.g. policies $\pi_w(x) \approx \pi_w(x)$.

10 Q-learning

Q-learning is an algorithm, that contains many of the basic structures required for reinforcement learning and acts as the basis for many more sophisticated algorithms. The Q-learning algorithm can be seen as an (asynchronous) implementation of the Robbins-Munro procedure for finding fixed points. For this reason we will require results from Section D when proving convergence.

A key ingredient is the notion of a Q-factor as described in Section 3. Recall that optimal Q-factor, $Q(x,a)$, is the value of starting in state x taking action a and thereafter following the optimal policy. In Prop 15 we showed that this solved the recursion:

$$Q(x,a) = \mathbb{E}_{x,a}[r(x,a) + \beta \max_{\hat{a}} Q(\hat{X},\hat{a}))]\,. \tag{8}$$

Def 54 (Q-learning)**.** *Given a state x, an action a, its reward $r(x,a)$ and the next state $\hat{x}$, Q-learning performs the update*

$$Q(x,a) \overset{\alpha}{\longleftarrow} r(x,a) + \beta \max_{a' \in \mathcal{A}} Q(\hat{x},a') - Q(x,a)$$

where α positive (learning rate) parameter. Recall $x \overset{\alpha}{\longleftarrow} dx$ means reset x with x' such that $x' = x + \alpha dx$.

To implement this as an algorithm, we assume that we have a sequence of state-action-reward-next_state quadruplets $\{(x_t,a_t,r_t,\hat{x}_t)\}_{t=0}^{\infty}$ and we apply the above update to each of the terms in this sequence.

Thrm 55. *For a sequence of state-action-reward triples $\{(x_t,a_t,r_t,\hat{x}_t)\}_{t=0}^{\infty}$ Consider the Q-learning update for $(x,a,r,\hat{x}) = (x_t,a_t,r_t,\hat{x}_t)$*

$$Q_{t+1}(x,a) = Q_t(x,a) + \alpha_t(x,a)\Big(r + \max_{a'} Q_t(x',a') - Q_t(x,a)\Big)$$

if the sequence of state-action-reward triples visits each state and action infinitely often, and if the learning rate $\alpha_t(x,a)$ is an adapted sequence satisfying the Robbins-Munro condition

$$\sum_{t=1}^{\infty} \alpha_t(x,a) = \infty, \qquad \sum_{t=1}^{\infty} \alpha_t^2(x,a) < \infty$$

then, with probability 1,

$$Q_t(x,a) \to Q^*(x,a)$$

where $Q^*(x,a)$ is the optimal value function.

Proof. We essentially show that the result is a consequence of Theorem 73 in Section D. We note that the optimal Q-function, $\boldsymbol{Q} = (Q(x,a) : x \in \mathcal{X}, a \in \mathcal{A})$ satisfies a fixed point equation

$$\boldsymbol{Q} = \boldsymbol{F}(\boldsymbol{Q}),$$

with

$$F_{x,a}(\boldsymbol{Q}) = \mathbb{E}_{x,a}[r(x,a) + \beta \max_{\hat{a}} Q(\hat{X}, \hat{a})],$$

for each $x \in \mathcal{X}$ and $a \in \mathcal{A}$. We know from Prop 15 that for discounted programming $\boldsymbol{F}(\cdot)$ is a contraction. I.e.

$$||\boldsymbol{F}(\boldsymbol{Q}_1) - \boldsymbol{F}(\boldsymbol{Q}_2)||_\infty \leq \beta ||\boldsymbol{Q}_1 - \boldsymbol{Q}_2||_\infty .$$

Now notice that the Q-learning algorithm performs the update

$$Q_{t+1}(x,a) = Q_t(x,a) + \alpha_t(x,a)(F(\boldsymbol{Q})(x,a) - Q_t(x,a) + \epsilon(x,a)),$$

where

$$\epsilon(x,a) = r + \beta \max_{\hat{a}} Q(\hat{X}, \hat{a}) - \mathbb{E}_{x,a}[r(x,a) + \beta \max_{\hat{a}} Q(\hat{X}, \hat{a})]$$

for $(x_t, a_t, r_t, \hat{x}_t) = (x, a, r, \hat{x})$. The update above is a Robbin's Munro update. FurtherbNotice $Q(x', a')$ remains the same for all other values of x, a, the update is asynchronous. It is not hard to see that when we condition on $\mathcal{F}_t$ the set of previous actions and states that

$$\mathbb{E}[\epsilon_t(x_t, a_t) | \mathcal{F}_t] = 0$$

and, a quick calculation shows,[3] that

$$\mathbb{E}[\epsilon_t(x_t, a_t)^2 | \mathcal{F}_t] \leq 2r_{\max}^2 + 2\beta^2 \max_{x,a} Q_t(x,a)^2 .$$

From this we see that we are working in the setting of Theorem 73 and that the condtions of that theorem are satisfied. Thus it must be that

$$Q_t(x,a) \xrightarrow[t\to\infty]{} Q^*(x,a)$$

where $Q^*(x,a)$ satisfies $\boldsymbol{Q}^* = \boldsymbol{F}(\boldsymbol{Q}^*)$. In otherwords, as required, it satisfies the Bellman equation for the optimal Q-function and thus is optimal. □

[3] Note $(x+y)^2 \leq 2x^2 + 2y^2$

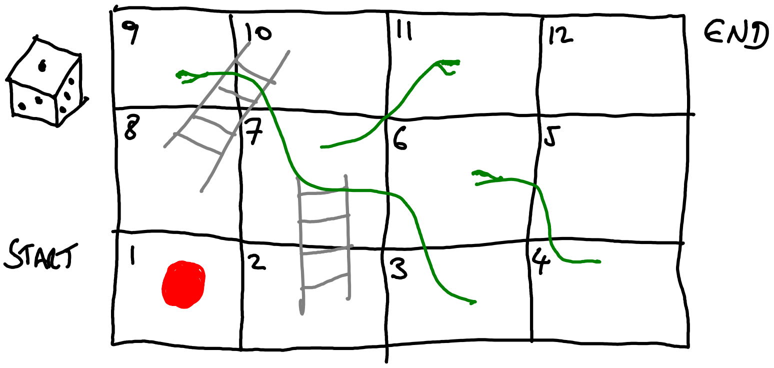
A Markov Chains: A Quick Review

This section is intended as a brief introductory recap of Markov chains. A much fuller explanation and introduction is provided in standard texts e.g. Norris [1], Bremaud [2], or Levin & Peres [3].

Introductory example: snakes and ladders

We highlight some of the key properties of Markov chains: how to calculate transitions, how the past effects the current movement of the processes, how to construct a chain, what the long run behavior of the process may (or may not) look like. We give an initial example to better position our intuition.

Below in Figure A, we are given a game of snakes and ladders (or shoots and ladders in the US). Here a counter (coloured red) is placed on the board at the start. You roll a dice. You move along the numbered squares by an amount given by the dice. The objective is to get to the finish. If the counter lands on a square with a snake's head, you must go back to the square at the snakes tail and, if you land on a square at the bottom of a ladder, go to the top of the ladder.



We let X_t be the position of the counter on the board after the dice has been thrown t times. The processes $X = (X_t : t \in \mathbb{Z})$ is a discrete time Markov chain. Two things to note: First, note that given the counter is currently at a state, e.g. on square 5, the next square reached by the counter – or indeed the sequence of states visited by the counter after being on square 5 – is not effected by the path that was used to reach the square. I.e. This is called the Markov

Property. Second, notice each movement of the counter from one state is a function of two pieces of information the current state and the independent random roll of the dice. In this way we can construct (or simulate) the random process.

Definitions

Let $\mathcal{X}$ be a countable set. An initial distribution

$$\lambda = (\lambda_x : x \in \mathcal{X})$$

is a positive vector whose components sums to one. A transition matrix $P = (P_{xy} : x, y \in \mathcal{X})$ is a postive matrix whose rows sum to one, that is, for $x \in \mathcal{X}$

$$\sum_{y \in \mathcal{X}} P_{xy} = 1.$$

With an initial distribution λ and a transition matrix P, you can define a Markov chain. Basically λ_x determines the probability the process starts in state x Vand P_{xy} gives the probability of going to y if you are currently in state x.

Def 56 (Discrete Time Markov Chain)**.** *We say that a sequence of random variables $X = (X_t : t \in \mathbb{Z}_+)$ is a discrete time Markov chain, with initial distribution λ and transition matrix P if for $x_0, ..., x_{t+1} \in \mathcal{X}$,*

$$\mathbb{P}(X_0 = x_0) = \lambda_0$$

and

$$\begin{aligned} \mathbb{P}(X_{t+1} = x_{t+1} | X_t = x_t, ..., X_0 = x_0) &= \mathbb{P}(X_{t+1} = x_{t+1} | X_t = x_t) \qquad \text{(Markov)} \\ &= P_{x_t x_{t+1}} \end{aligned}$$

The condition (Markov) is often called the Markov property and is the key defining feature of a Markov chain or, more generally, Markov process. It states that the past $(X_1, ..., X_{t-1})$ and future X_{t+1} are conditionally independent of the present X_t. Otherwise stated, is says that, when we know the past and present states $(X_1, ..., X_t) = (x_0, ..., x_t)$, the distribution of the future states $X_{t+1}, X_{t+2}, ...$ is only determined by the present state $X_t = x_t$. Think of a board game like snakes and ladders, where you go in the future is only determined by where you are now and not how you got there; this is the Markov property.

The following proposition shows that the evolution of a Markov chain can be constructed from its the current state and an independent dice throw.

Prop 57 (Constructing Markov Chains). *Take a function $f : \mathcal{X} \times [0,1] \rightarrow \mathcal{X}$, X_0 a random variable on $\mathcal{X}$, and $(U_t)_{t\geq 0}$, independent uniform $[0,1]$ random variables. The sequence $(X_t)_{t\geq 0}$ constructed with the recursion*

$$X_{t+1} = f(X_t, U_t) \qquad \text{for} \quad t = 0, 1, 2, ..$$

is a discrete time Markov chain. Moreover all discrete time Markov chains can be constructed in this way.

The following proposition will be useful when we want to sum up a long sequence of rewards.

Prop 58 (Markov Chains and Potential Functions). *For $r : \mathcal{X} \rightarrow \mathbb{R}_+$ be a bounded function and for $\beta \in (0,1)$,*

$$R(x) = \mathbb{E}_x\left[\sum_{t=0}^{\infty} \beta^t r(X_t)\right]$$

is the unique solution to the equation

$$R(x) = \beta(PR)(x) + r(x), \qquad x \in \mathcal{X}.$$

Moreover, if function $\tilde{R} : \mathcal{X} \rightarrow \mathbb{R}_+$ satisfies

$$\tilde{R}(x) \geq \beta(P\tilde{R})(x) + r(x), \qquad x \in \mathcal{X}.$$

then $\tilde{R}(x) \geq R(x)$, $x \in \mathcal{X}$.

The following is an alternative formulation of the previous proposition.

Prop 59. *Let $\partial\mathcal{X}$ be a subset of $\mathcal{X}$ and let T be the hitting time on $\partial\mathcal{X}$ i.e. $T = \inf\{t : X_t \in \partial\mathcal{X}\}$ and take $f : \partial X \rightarrow \mathbb{R}_+$ argue that*

$$R(x) = \mathbb{E}_x\left[\sum_{t<T} r(X_t) + f(X_T)\mathbb{I}[T < \infty]\right]$$

solves the equation

$$R(x) = (PR)(x) + r(x), \qquad x \notin \partial\mathcal{X} \tag{9}$$

$$R(x) = f(x), \qquad x \in \mathcal{X}. \tag{10}$$

There is a close connection between Markov chains and Martingales that we will want to use later when considering Markov Decision Processes.

Prop 60 (Markov Chains and Martingale Problems)**.** *Show that a sequence of random variables $X = (X_t : t \in \mathbb{Z}_+)$ is a Markov chain if and only if, for all bounded functions $f : \mathcal{X} \to \mathbb{R}$, the process*

$$M_t^f = f(X_t) - f(X_0) - \sum_{\tau=0}^{t-1} (P - I) f(X_\tau)$$

is a Martingale with respect to the natural filtration of X. Here for any matrix, say Q, we define

$$Qf(x) := \sum_{y \in \mathcal{X}} Q_{xy} f(y).$$

References

[1] Norris, J.R., 1997. Markov chains. Cambridge University Press.

[2] Bremaud, P., 2013. Markov chains: Gibbs fields, Monte Carlo simulation, and queues (Vol. 31). Springer Science & Business Media.

[3] Levin, D.A. and Peres, Y., 2017. Markov chains and mixing times (Vol. 107). American Mathematical Soc..

B Stochastic Integration

What follows is a heuristic derivation of the Stochastic Integral, Stochastic Differential Equations and Itô's Formula.

First note that for $(B_t : t \geq 0)$ a standard Brownian motion argue that, for all T and for δ sufficiently small and positive,

$$\sum_{t\in\{0,\delta,..,T\}} (B_{t+\delta} - B_t) = B_T \quad \text{and} \quad \sum_{t\in\{0,\delta,..,T\}} (B_{t+\delta} - B_t)^2 \approx T \tag{11}$$

The 1st sum is an interpolating sum. By independent increments property of Brownian motion, the 2nd sum adds IIDRVs with each with mean δ. Thus the strong law of large numbers gives the approximation. From this it is reasonable to expect that

$$\sum_{t\in\{0,\delta,..,T\}} \sigma(X_t)\,(B_{t+\delta} - B_t) \approx \int_0^T \sigma(X_t)dB_t$$

and

$$\sum_{t\in\{0,\delta,..,T\}} \mu(X_t)\,(B_{t+\delta} - B_t)^2 \approx \int_0^T \mu(X_t)dt.$$

The first sum, above, is approximation from a Riemann-Stieltjes integral, i.e.

$$\int_0^T f(t)dg(t) \approx \sum_{t\in\{0,\delta,..,T\}} f(t)(g(t+\delta) - g(t)).$$

So one might expect a integral limit. (This is unrigorous because Riemann-Stieltjes Integration only applies to functions with finite variation – while Brownian motion does not have finite variation.) The second sum is a Riemann integral upon using the approximation $(B_{t+\delta} - B_t)^2 \approx \delta$. This is, very roughly, how a stochastic integral is defined.

We can also define stochastic differential equations. If we inductively define X_t by the recursion

$$X_{t+\delta} - X_t = \sigma(X_t)(B_{t+\delta} - B_t) + \mu(X_t)\delta, \qquad t = 0, \delta, 2\delta, ... \tag{12}$$

then, by summing over values of $t \in \{0, \delta, ..., T - \delta\}$, we expect X_t to approximately obey an equation of the form

$$X_T = X_0 + \int_0^T \sigma(X_t)dB_t + \int_0^T \mu(X_t)dt.$$

This gives a Stochastic Differential Equation.

Often in differential and integration, we apply chain rule, $\frac{df(x_t)}{dt} = f'(x_t)\frac{dx_t}{dt}$. Ito's the analogous result for Stochastic Integrals. Let X_t be as above. For a twice continuously differentiable function f and $\delta > 0$ small, we can apply a Taylor approximation

$$\begin{aligned} & f(X_{t+\delta}) - f(X_t) \\ =& f(X_t + \sigma(X_t)(B_{t+\delta} - B_t) + \mu(X_t)\delta) - f(X_t) \\ =& f'(X_t)\left\{\mu\delta + \sigma \cdot (B_{t+\delta} - B_t)\right\} + \frac{f''(X_t)}{2}\left\{\mu\delta + \sigma \cdot (B_{t+\delta} - B_t)\right\}^2 + o(\delta) \\ =& f'(X_t)\left\{\mu\delta + \sigma \cdot (B_{t+\delta} - B_t)\right\} + \frac{f''(X_t)}{2}\sigma^2 \cdot (B_{t+\delta} - B_t)^2 + o(\delta) \end{aligned}$$

In the last equality we use that $(B_{t+\delta} - B_t) = o(\delta^{1/2})$ (which follows from (11)). Thus we see that

$$f(X_{t+\delta}) - f(X_t) \approx \left[f'(X_t)\mu(X_t) + \frac{\sigma(X_t)^2}{2}f''(X_t)\right]\delta + f'(X_t)\sigma(X_t)\left(B_{t+\delta} - B_t\right).$$

Consequently we expecrt that $f(X_t)$ obeys the following Stochastic Differential Equation:

$$f(X_T) - f(X_0) = \int_0^T \left[f'(X_t)\mu(X_t) + \frac{\sigma(X_t)^2}{2}f''(X_t)\right]dt + \int_0^T f'(X_t)\sigma(X_t)dB_t.$$

This is Ito's formula.

C Utility Theory

- Utility functions; equivalence of utility functions
- Relative risk aversion; CRRA Utility and iso-elasticity.

A utility function $U(x)$ is used to quantify the value that you gain from an outcome x.

Def 61 (Utility Function)**.** *For $\mathcal{X} \subset \mathbb{R}^d$, a utility function is a function $U : \mathcal{X} \to \mathbb{R}$ that is increasing, i.e. if $x \leq y$ component-wise then $U(x) \leq U(y)$. The utility of a random variable X is then its expected utility, $\mathbb{E}U(X)$. A utility function creates an ordering where an outcome X is preferred to Y if $\mathbb{E}U(X) \geq \mathbb{E}U(Y)$.*

Jensen's inequality applies to a concave utility:

$$\mathbb{E}U(X) \leq U(\mathbb{E}X)$$

So we prefer a certain outcome $\mathbb{E}X$ rather than the risky outcome X that has the same mean – This is being risk averse.

Def 62 (Risk Aversion)**.** *If the function is concave then we also say that the function is risk averse. (Unless stated otherwise we assume that the utility function is risk averse).*

Def 63. *We say that two utility functions U and V are equivalent if they induce the same ordering. I.e. $\mathbb{E}U(X) \leq \mathbb{E}U(Y)$ iff $\mathbb{E}V(X) \leq \mathbb{E}V(Y)$.*

Ex 64. *Show that two utility functions are equivalent iff V the same as U up-to an affine transform, i.e.*

$$V(x) = aU(x) + b$$

for constants $a > 0$ and b.

Ans 64. *Define $\phi : \mathbb{R} \to \mathbb{R}$ s.t. $\phi(\mathbb{E}U(X)) = \mathbb{E}V(X)$. Let $X = x$ w.p. and $X = y$ w.p. $q = 1 - p$. Then*

$$\phi(pU(x) + qU(y)) = pV(x) + qV(y) = p\phi(U(x)) + (1 - p)\phi(U(y)).$$

This implies ϕ is linear.

Def 65 (Coefficient of Relative Risk Aversion). *For a utility function $U : \mathbb{R} \to \mathbb{R}$ (twice differentiable) the Coefficient of Relative Risk Aversion is*

$$-x\frac{U''(x)}{U'(x)}.$$

Ex 66. *You have utility function U. You are offered a bet that increases you wealth w multiplicatively by $(1 + X)$ here X is a "small" positive is a RV. Discuss why you would accept the bet iff*

$$\frac{2\mathbb{E}X}{\mathbb{E}X^2} \geq -x\frac{U''(x)}{U'(x)}$$

I.e. You accept the bet if you mean is large but a large variance makes this less likely, and the coefficient of relative risk aversion decides the threshold.

Ans 66. *Accept if*

$$0 \leq \mathbb{E}[U(w(1+X)) - U(w)] \overset{Taylor}{\approx} \mathbb{E}\left[U'(w)wX + \frac{1}{2}U''(w)w^2X^2\right].$$

Def 67 (CRRA Utility/Iso-elastic Utility). *A Constant Relative Risk Averse utility (CRRA) takes the form*

$$U(x) = \begin{cases} \frac{x^{1-R}}{1-R}, & R \neq 1, \\ \log x, & R = 1. \end{cases}$$

Def 68. *A utility function is Iso-elastic if it is unchanged under multiplication: for all $c > 0$,*

$$\mathbb{E}U(X) \geq \mathbb{E}U(Y) \qquad \textit{iff} \qquad \mathbb{E}U(cX) \geq \mathbb{E}U(cY).$$

I.e. the utility only cares about the relative magnitude of the risk.

Ex 69. *Show that a utility function is iso-elastic iff it is a CRRA utility (up-to an affine transform).*

Ans 69. *By [64], its immediate that CRRA implies isolastic. Further by [64], $\forall c$, $U(cx) = a_cU(x) + b_c$ for constants a_c and b_c. Differentiate twice w.r.t. x and divide gives*

$$\frac{cU''(cx)}{U'(cx)} = \frac{U''(x)}{U'(x)}$$

Set $x = 1$ and integrate twice w.r.t. c gives the required result.

D Robbins-Munro

We review a method for finding fixed points then extend it to slightly more general, modern proofs.

Often it is important to find a solution to the equation

$$0 = g(x^*)$$

by evaluating g at a sequence of points. For instance Newton's method would perform the updates $x_{n+1} = x_n - g(x_n)/g'(x_n)$. However, Robbins and Munro consider the setting where we cannot directly observe g but we might observe some random variable whose mean is $g(x)$. Thus we observe

$$y_n = g(x_n) + \epsilon_n \tag{13}$$

and hope solve for $g(x) = 0$. Notice in this setting, even if we can find $g'(x)$, Newton's method may not converge. The key idea of Robbins and Munro is to use a schema where

$$x_{n+1} = x_n - \alpha_n y_n \tag{RM}$$

where we chose the sequence $\{\alpha_n\}_{n=0}^{\infty}$ so that

$$\sum_n \alpha_n = \infty, \qquad \sum_n \alpha_n^2 < \infty .$$

Before proceeding here are a few different use cases:

- **Quartiles.** We want to find x such that $P(X \leq x) = p$ for some fixed p. But we can only sample the random variable X.
- **Regression.** We preform regression $g(x) = \beta_0 + \beta_1 x$, but rather than estimate β we want to know where $g(x) = 0$.
- **Optimization.** We want to optimize a convex function $f(x)$ whose gradient is $g(x)$. Assume that $f(x) = \sum_{k=1}^{K} f_k(x)$ for some large K. To find the optimum at $g(x) = 0$ we randomly sample (uniformly) $f_k(x)$ whose gradient, $g_k(x)$, is an bias estimate of $g(x)$.

The following result contains the key elements of the Robbins-Munro proof

Prop 70. *Suppose that z_n is a positive sequence such that*

$$z_{n+1} \leq z_n(1-a_n) + c_n \tag{14}$$

where a_n and c_n are positive sequences such that

$$\sum_n a_n = \infty, \qquad \text{and} \qquad \sum_n c_n < \infty \tag{15}$$

then $\lim_{n\to\infty} z_n = 0$.

Proof. We can assume that equality holds, i.e., $z_{n+1} = z_n(1-a_n)+c_n$. We can achieve this by increasing a_n or decreasing c_n in the inequality (14); neither of which effect the conditions on a_n and b_n, (15).

Now for all n we have the following lower-bound

$$-z_0 \leq z_n - z_0 = \sum_{k=1}^{n-1}(z_{k+1}-z_k) = \sum_{k=1}^{n-1} c_k - \sum_{k=1}^{n-1} a_k z_k$$

Since $\sum c_k < \infty$ it must be that $\sum a_k z_k < \infty$. Thus since both sums converge it must be that $\lim_n z_n$ converges. Finally since $\sum a_k = \infty$ and $\sum a_k z_k < \infty$ it must be that $\lim_n z_n = 0$. □

The following proposition is a Martingale version of the above result.

Prop 71 (Robbins-Siegmund Theorem). *If*

$$\mathbb{E}[Z_{n+1}|\mathcal{F}_n] \leq (1-a_n+b_n)Z_n + c_n \tag{16}$$

for positive adaptive RVs Z_n, a_n, b_n, c_n such that with probability 1,

$$\sum_n a_n = \infty, \qquad \sum_n b_n < \infty, \qquad \text{and} \qquad \sum_n c_n < \infty$$

then $\lim_{n\to\infty} z_n = 0$.

Proof. The results is some manipulations analogous to the Robbins-Munro proof and a bunch of nice reductions to Doob's Martingale Convergence Theorem.

First note the result is equivalent to proving the result with $b_n = 0$ for all n. If we divide both sides of (16) by $\prod_{m=0}^{n}(1-b_m)$ we get

$$\mathbb{E}[Z'_{n+1}|\mathcal{F}_n] \leq (1-a'_n)Z'_n + c'_n\,,$$

where $a'_n = a_n/(1+b_n)$, $c'_n = c_n/\prod_{m=0}^{n}(1+b_m)$ and $Z'_n = Z_n/\prod_{m=0}^{n}(1+b_m)$. Notice since $\sum b_n$ converges then so does $\prod(1+b_n)$. Thus a'_n, c'_n and Z'_n have the same convergence properties as those required for a_n, c_n and Z_n. Thus, we now assume $b_n = 0$ for all n.

Now notice

$$Y_n = Z'_n + \sum_{k=0}^{n-1} a'_k Z'_k - \sum_{k=0}^{n-1} c'_k$$

is a super-martingale. We want to use Doob's Martingale convergence theorem; however, we need to apply a localization argument to apply this. Specifically, let $\tau_C = \inf\{n \geq 0 : \sum_{k=1}^{n} c'_k > C\}$. This is a stopping time. Notice

$$Y_{n\wedge\tau_C} \geq - \sum_{k=0}^{n\wedge\tau_C - 1} c'_k \geq -C.$$

So $Y_{n\wedge\tau_C}$ is a super-martingale and below by $-C$. Thus by Doob's Martingale Convergnce Theorem, $\lim_{n\to\infty} Y_{n\wedge\tau_C}$ exists for each $C > 0$, and $\tau_C = \infty$ for some C, since $\sum c'_k < \infty$. Thus $\lim_{n\to\infty} Y_n$ exists.

Now notice

$$\sum_{k=1}^{n} c'_k - \sum_{k=1}^{n} a'_k Z'_k = Z'_{n+1} - Y_{n+1} \leq -Y_{n+1}.$$

So like in the last proposition, since $\lim Y_n$ and $\sum c'_k$ exists, we see that $\sum_{k=1}^{\infty} a'_k Z'_k$ converges. And thus Z'_{n+1} converges.

Finally since we assume $\sum_k a'_k = \infty$ and we know that $\sum_{k=1}^{\infty} a'_k Z'_k < \infty$ it must be that Z'_k converges to zero. □

Stochastic Gradient Decent

Suppose that we have some function $F : \mathbb{R}^p \to \mathbb{R}$

$$F(\theta) = \mathbb{E}_X[f(X;\theta)]$$

that we wish to minimize. We suppose that the function $f(X;\theta)$ is known and so is its gradient $g(\theta;X)$, where $\mathbb{E}[g(\theta;X)] = G(\theta)$ is the gradient of $F(\theta)$. The difficulty is that we do not have direct access to the distribution of X, but we can draw random samples $X_1, X_2, \dots$. We can use the Robbins-Munro Schema to optimize $F(\theta)$. Specifically we take

$$\begin{aligned} \theta_{n+1} &= \theta_n - \alpha_n g_n(\theta_n) \\ &= \theta_n - \alpha_n G(\theta_n) + \alpha_n \epsilon_n \end{aligned} \tag{SGD}$$

where $g_n(\theta) = g(\theta;X_n)$ and $\epsilon_n = G(\theta) - g_n(\theta)$. The above sequence is often referred to as *Stochastic Gradient Descent*. We chose the sequence $\{\alpha_n\}_{n=0}^{\infty}$ so that

$$\sum_n \alpha_n = \infty, \qquad \sum_n \alpha_n^2 < \infty\,.$$

(Note here we may assume that α_n is a function of previous parameters and observations $\theta_1, ..., \theta_{n-1}$ and $X_1, ..., X_{n-1}$.) We let $||\cdot||_2$ be the Euclidean norm. We can prove that convergence θ_n to the minimizer of $F(\theta)$.

Thrm 72 (Stochastic Gradient Descent)**.** *Suppose that θ_n, $G(\cdot)$, and ϵ_n in Stochastic Gradient Descent* (SGD) *satisfy the following conditions*

1. $\exists\theta^*$ *such that* $\forall\theta$, $G(\theta)\cdot(\theta-\theta^*) \geq \mu||\theta-\theta^*||_2^2$
2. $||G(\theta_n)||_2^2 \leq A + B||\theta_n||_2^2$
3. $\mathbb{E}[||\epsilon_n||_2^2|\mathcal{F}_n] \leq K$

Then $\lim_n \theta_n = \theta^*$ *where* $\theta^* = \operatorname{argmin}_\theta F(\theta)$ *and assuming α_n are deterministic then* $\lim \mathbb{E}[||\theta_n - \theta^*||_2^2] = 0$

Let's quickly review the conditions above. First consider Condition 1. Note condition 1 implies moving in the direction of θ^* always decreases the $F(\theta)$, so θ^* minimizes F. The statement $(G(\theta)-G(\phi))\cdot(\theta-\phi) \geq \mu||\theta-\phi||^2$ is equivalent to $F(\theta)$ being strongly convex. So this is enough to give Condition 1. Condition 2 can be interpreted as a gradient condition, or that the steps θ_n do not grow unboundedly. Condition 3 is natural given our analysis so far.

Proof.

$$\begin{aligned}&\|\theta_{n+1}-\theta^*\|_2^2-\|\theta_n-\theta^*\|_2^2\\=&-\alpha_n G(\theta_n)\cdot(\theta_n-\theta^*)-\alpha_n\epsilon_n\cdot(\theta_n-\alpha_n G(\theta_n)-\theta^*)+\alpha_n^2\|\epsilon_n\|_2^2+\alpha_n^2\|G(\theta_n)\|_2^2\end{aligned}$$

Taking expecations with respect to $\mathbb{E}[|\mathcal{F}_n]$ we get

$$\begin{aligned}&\mathbb{E}[\|\theta_{n+1}-\theta^*\|_2^2-\|\theta_n-\theta^*\|_2^2|\mathcal{F}_n]\\=&-\alpha_n G(\theta_n)\cdot(\theta_n-\theta^*)+\alpha_n^2\mathbb{E}[\|\epsilon_n\|_2^2|\mathcal{F}_n]+\alpha_n^2\|G(\theta_n)\|_2^2\\\leq&-\alpha_n\mu\|\theta_n-\theta^*\|_2^2+\alpha_n^2K+\alpha_n^2(A+B\|\theta_n-\theta^*\|_2^2)\end{aligned}$$

Thus, rearranging

$$\mathbb{E}[\|\theta_{n+1}-\theta^*\|_2^2|\mathcal{F}_n]\leq(1-\alpha_n\mu+\alpha_n^2B)\|\theta_n-\theta^*\|_2^2+\alpha_n^2(K+A)$$

Thus by Proposition 71, $\theta_{n+1}\to\theta^*$. Further taking expectations on both sides above we have

$$\mathbb{E}[\|\theta_{n+1}-\theta^*\|_2^2]\leq(1-\alpha_n\mu+\alpha_n^2B)\mathbb{E}[\|\theta_n-\theta^*\|_2^2]+\alpha_n^2(K+A)$$

We can apply Proposition 70 (note that $a_n=\alpha_n\mu+\alpha_n^2B$ will be positive for suitably large n), to give that $\mathbb{E}\|\theta_{n+1}-\theta^*\|_2^2]\to 0$ as $n\to\infty$ as required. □

Finally we remark that in the proof we analyzed $\|\theta_n-\theta^*\|_2$ but equally we could have analyzed $F(\theta_n)-F(\theta^*)$ instead.

Fixed Points and Asynchronous Update

We now consider Robbins-Munro from a slightly different perspective. Suppose we have a continuous function $F : \mathbb{R}^p \to \mathbb{R}^p$ and we wish to find a fixed point x^* such that $F(x^*) = x^*$. We assume that $F(\cdot)$ is a contraction namely that, for some $\beta \in (0, 1)$,

$$||F(x) - F(y)||_\infty \le \beta ||x - y||_\infty . \tag{17}$$

Here $||x||_\infty = \max_{i=1,...,p} |x_i|$. (Note this contraction condition implies the existence of a fixed point). (Note the previous analysis was somewhat restricted to euclidean space.) If we suppose that we do not observe $F(x)$ but instead some perturbed version whose mean is $F(x)$, then we can perform the Robbins-Munro update for each component $i = 1, ...p$:

$$x_i(t + 1) = x_i(t) + \alpha_i(t)(F_i(x(t)) - x_i(t) + \epsilon_i(t)) \tag{RM-Async}$$

where $\alpha_i(t)$ is a sequence such that for all i

$$\sum_t \alpha_i(t) = \infty, \qquad \sum_t \alpha_i^2(t) < C . \tag{RM step}$$

for some constant C. Further we suppose that $\epsilon_i(t-1)$ is measurable with respect to $\mathcal{F}_t$, the filtration generated by $\{\alpha_i(s), x_i(s)\}_{s\le t}$ measurable and

$$\mathbb{E}[\epsilon_i(t)|\mathcal{F}_t] = 0 . \qquad \text{and} \qquad \mathbb{E}[\epsilon_i^2(t)|\mathcal{F}_t] \le A + B \max_j |x_j(t)|^2 . \tag{18}$$

Note that in the above we let the step rule depend on i. For instance at each time t we could chose to update one component only at each step, e.g., to update component i only, we would set $\alpha_j(t) = 0$ for all $j \neq i$. Thus we can consider this step rule to be asynchronous.

We can analyze the convergence of this similarly

Thrm 73. *Suppose that $F(\cdot)$ is a contraction with respect to $||\cdot||_\infty$* (17), *suppose the vector $x(t)$ obeys the step rule* (RM-Async) *with step sizes satisfying* (RM step) *and further suppose the noise terms satisfy* (18), *then*

$$\lim_{t\to\infty} x(t) = x^*$$

where x^* is the fixed point $F(x^*) = x^*$.

We will prove the result under the assumption that $x(t)$ is bounded in t, this is the proposition, Prop 74, below. We then prove that $x(t)$ is bounded in t to complete the proof.

Prop 74. *If we further assume that*

$$\sup_t ||x(t)||_\infty < \infty$$

with probability 1 then Thrm 73 holds.

We may assume with out loss of generality that $x^* = 0$, since the recursion above is equivalent to

$$x_i(t+1) - x^* = x_i(t) - x^* + \alpha_i(t)(F_i(x(t)) - F_i(x^*) - x_i(t) + x^* + \epsilon_i(t)) .$$

Given the assumption above that $||x(t)||_\infty \leq D_0$ for all t, further define

$$D_{k+1} = \beta(1+2\epsilon)D_k$$

Here we choose ϵ so that $(1+2\epsilon)\beta < 1$ so that $D_k \to 0$. By induction, we will show that, given $||x(t)||_\infty < D_k$ for all $t \geq \tau_k$ for some τ_l, then there exists a τ_{k+1} such that for all $t \geq \tau_{k+1}$

$$||x(t)||_\infty < D_{k+1}$$

We use two recursions to bound the behavior of $x_i(t)$:

$$W_i(t+1) = (1-\alpha_i(t))W_i(t) + \alpha_i(t)w_i(t)$$
$$Y_i(t+1) = (1-\alpha_i(t))Y_i(t) + \alpha_i(t)\beta D_k .$$

for $t \geq \tau_k$, where $W_i(\tau_k) = 0$ and $Y(\tau_k) = 0$. We use $W_i(t)$ to summarize the effect of noise on the recursion for $x_i(t)$ and we use $Y_i(t)$ to bound the error arising from the function $F_i(x)$ in the recursion for $x_i(t)$. Specifically we show that

$$|x_i(t) - W_i(t)| \leq Y_i(t)$$

in Lemma 5 below. Further we notice that is a Robbin-Munro recursion for $W_i(t)$ to go to zero and $Y_i(t)$ to go to βD_k.

Lemma 5. $\forall t_0 \geq \tau_k$

$$|x_i(t) - W_i(t)| \leq Y_i(t)$$

Proof. We prove the result by induction. The result is clearly true for $t = \tau_k$.

$$\begin{aligned} x_i(t+1) &= (1-\alpha_i(t))x_i(t) + \alpha_i(t)F_i(\boldsymbol{x}^i(t)) + \alpha_i(t)w_i(t) \\ &\leq (1-\alpha_i(t))(Y_i(t) + W_i(t)) + \alpha_i(t)\beta D_k + \alpha_i(t)\epsilon_i(t) \\ &= Y_i(t+1) + W_i(t+1) \end{aligned}$$

In the inequality above with apply the induction hypothesis on $x_i(t)$ and bounds of F_i. The second equality just applies the definitions of Y_i and W_i. Similar inequalities hold in the other direction and give the result. □

Lemma 6.

$$\lim_{t\to\infty} |W_i(t)| = 0$$

Proof. Letting $W(t) = W(t,0)$, we know

$$\mathbb{E}[W(t+1)^2|\mathcal{F}_t] \leq (1-2\alpha(t)+\alpha^2)W(t)^2 + \alpha(t)^2\mathbb{E}[\epsilon(t)^2|\mathcal{F}_t].$$

From the Robbins-Siegmund Theorem (Prop 71), we know that

$$\lim_{t\to\infty} W(t) = 0.$$

□

Lemma 7.

$$Y_i(t) \xrightarrow[t\to\infty]{} \beta D_k$$

Proof. Notice

$$Y_i(t+1) - \beta D_k = (1-\alpha_i(t))(Y_i(t) - \beta D_k) = \dots = \left(\prod_{s=1}^{t}(1-\alpha_i(s))\right)(Y_i(0) - \beta D_k)$$

The result holds since $\sum \alpha_i(t) = \infty$. □

We can now prove Prop 74.

Proof of Prop 74. We know that $\|x(t)\|_\infty \leq D_0$ for all t and we assume $\|x(t)\|_\infty \leq D_k$ for all $t \geq \tau_k$. By Lemma 5 and then by Lemmas 6 and 7

$$|x_i(t)| \leq Y_i(t) + |W_i(t)| \xrightarrow[t\to\infty]{} \beta D_k$$

as required. Thus these exists τ_{k+1} such that $\sup_{t\geq\tau_{k+1}} \|x(t)\|_\infty \leq D_{k+1}$. Thus by induction we see that $\sup_{t\geq\tau_k} \|x(t)\|_\infty$ decreases through sequence of levels D_k as $k \to \infty$, thus $x(t)$ goes to zero as required. □

Proving Boundedness of $x(t)$

We now prove that $x(t)$ remains bounded

Prop 75.

$$\sup_t \|x(t)\|_\infty < \infty$$

To prove this proposition, we define a processes that bounds the max of $\|x(t)\|_\infty$ from above in increments of size $(1+\epsilon)$. Specfically we let

$$M(t) := \max_{\tau \le t} \|x(\tau)\|_\infty$$

and we define $G(0) = \max\{M(0), G_0\}$ and let

$$G(t+1) = \begin{cases} G(t) & \text{if } M(t+1) < (1+\epsilon)G(t), \\ (1+\epsilon)^k G(t) & \text{if } M(t+1) \ge (1+\epsilon)G(t). \end{cases}$$

where in the above k is the smallest integer such that $M(t+1) \le (1+\epsilon)^k G(t)$. Note that $G(t)$ is adapted. Further note $|F_i(x)| \le \gamma \max\{\max_j |x_j|, G_0\}$ for some G_0 and $\gamma < 1$, since $\beta y + c \le \gamma y \vee G_0$ for suitable choice of γ and G_0. We use G_0 in the definition of $G(t)$ above and we choose ϵ so that $(1+\epsilon)\gamma < 1$.

Also we define $\tilde{W}_i(t_0; t_0) = 0$ and

$$\tilde{W}_i(t+1; t_0) = (1-\alpha_i(t))\tilde{W}_i(t; t_0) + \alpha_i(t)\tilde{w}_i(t), \quad \text{where} \quad \tilde{w}_i(t) = \frac{w_i(t)}{G(t)}.$$

Notice, like before, $\tilde{W}$ is a Robbin-Munro recursion that goes to zero and $x_i(t)$ is bounded by a recursion of this type.

Lemma 8. *If $G(t) = G(t_0)$ for $t \ge t_0$ then*

$$|x_i(t)| \le G(t_0) + \tilde{W}_i(t; t_0)G(t_0).$$

Proof. The result is somewhat similar to Lemma 5. At t_0, we have that

$$|x_i(t_0)| \le M(t_0) \le G(t_0).$$

Now assuming that the bound is true at time t.

$$\begin{aligned} x_i(t+1) &= (1-\alpha_i(t))x_i(t) + \alpha_i(t)F_i(\boldsymbol{x}^i(t)) + \alpha_i(t)w_i(t) \\ &\le (1-\alpha_i(t))\{G(t_0) + \tilde{W}_i(t; t_0)G(t_0)\} + \alpha_i(t)\gamma G(t_0)(1+\epsilon) + \alpha_i(t)\tilde{w}_i(t)G(t_0) \\ &\le G(t_0) + [(1-\alpha_i)\tilde{W}_i(t; t_0) + \alpha_i\tilde{w}_i(t)]G(t_0). \end{aligned}$$

Above we bound $x(t)$ knowing the bound holds at time t and we bound F using the fact that we know $G(t)$ has not yet increased. We then use the fact we chose ϵ so that $\gamma(1+\epsilon) < 1$. A similar bound holds on the other side. □

Lemma 9.

$$\lim_{t_0\to\infty} \sup_{t\geq t_0} |\tilde{W}(t;t_0)| = 0$$

Proof. Since $G(t)$ is adapted and from our assumptions on $w_i(t)$, we have that

$$\mathbb{E}[\tilde{w}_i(t)|\mathcal{F}_t] = 0 \quad \text{and} \quad \mathbb{E}[\tilde{w}_i(t)^2|\mathcal{F}_t] \leq K\,.$$

We know that $\lim_{t\to\infty} |\tilde{W}_i(t;0)|$ – the argument for this is identical to Lemma 6. Further notice for all $t \geq t_0$ we have

$$\begin{aligned} W_i(t;0) - W_i(t;t_0) &= (1-\alpha_i(t))\Big[W_i(t-1;0) - W_i(t-1;t_0)\Big] \\ &= \cdots = \prod_{s=t_0}^{t}(1-\alpha(s)) \cdot W(t_0;0)\,. \end{aligned}$$

Thus,

$$|W_i(t;t_0)| \leq |W_i(t_0;0)| + |W_i(t;0)|\,.$$

As required both terms on the righthand-side go to zero. □

Proof of Prop 75. To prove the proposition we will show that $G(t)$ at some point must remain bounded. Suppose t_0 is a time just after $G(t)$ increased. Note if $x_i(t)$ grew unboundedly then we could chose t_0 as large as we like.

So we'd know by Lemma 9 that there exists a t_0 such that for all $t \geq t_0$, $|\tilde{W}_i(t;t_0)| \leq \epsilon/2$. Thus applying this to Lemma 8 we see that

$$|x_i(t)| \leq G(t_0) + \tilde{W}_i(t;t_0)G(t_0) \leq G(t_0)(1+\epsilon/2)\,.$$

Thus since $M(t) < G(t)(1+\epsilon)$ for all $t \geq t_0$. So $G(t)$ can not longer increase and thus we arrive at a contradiction. $x_i(t)$ must be bounded in t. □

E Exercises

Exercise Sheet 1

Ex 1.1. (Shortest Paths) Consider a directed graph $G = (V, E)$ each edge has a cost, c_{ij} for each $(i, j) \in E$. Take a vertex d. Let L_i be the length of the shorest path from i to d and let $L_i(t)$ be the shortest path from i to d that uses t steps.

i) In addition to satisfying $L_d = 0$, argue that L_i satisfies the equations

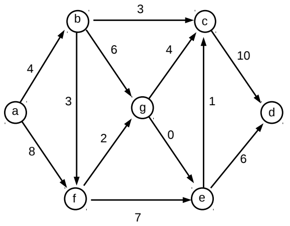
$$L_i = \min_{j:(i,j)\in E} \left\{ c_{ij} + L_j \right\}, \qquad \text{for } i \neq d.$$

ii) Argue that $L_i(t)$ satisfies, $L_d(t) = 0$ and

$$L_i(t+1) = \min_{j:(i,j)\in E} \left\{ c_{ij} + L_j(t) \right\}, \qquad \text{for } i \neq d.$$

(Here you may assume that $L_d(t) = 0$ for all $t \geq 0$ and $L_i(t) = \infty$ unless assigned a value in the above set of equalities.)

iii) Your answer to part ii) describes a algorithm called the *Bellman-Ford algorithm.* Use it to find the shortest path from node a to node d in the following graph



Ex 1.2. (Scheduling) There are N appointments that need to be sucessively scheduled over time. Each appointment $i = 1, ..., N$ requires t_i units of time and when completed has reward r_i. Given discount factor $\beta \in (0, 1)$, the total reward arragning the appointments in order $1, 2, ..., N$ is

$$R(1, ..., N) = r_1\beta^{t_1} + r_2\beta^{t_1+t_2} + ... + r_N\beta^{t_1+...+t_N},$$

i) Write down a dynamic program for the optimal discounted reward. (Here let $W(S)$ be the optimal reward when $S \subset \{1, ..., N\}$ is the remaining set of unassigned appointments.)

ii) Argue that it is optimal to order appointments so that the indices

$$G_i = \frac{r_i \beta^{t_i}(1-\beta)}{1-\beta^{t_i}}$$

indexed from highest to lowest.

Ex 1.3. A deck of cards is shuffled and you take out cards sequentially. You can bet at any point that the next card is red if correct you win £1000 and if incorrect you win nothing and the game stops. Calculate $F(r,b)$, where $F(r,b)$ gives the maximum expected winnings given there are r red and b black cards in the deck.

Ex 1.4. Prove the theorem from the course that the value function $W_t(x)$ of a Markov Decision Problem satisfies the Bellman equation:

$$W_t(x) = \sup_{a_t \in \mathcal{A}_t} \left\{ r_t(x_t, a_t) + \mathbb{E}_{x_t, a_t}\left[W_{t+1}(x_{t+1}) \right] \right\}$$

and $W_T(x) = r_T(x)$.

Exercise Sheet 2

The first questions here act as practice/revision on Markov Chains.

We let $\mathcal{X}$ be a countable set. Further, let $\lambda = (\lambda_x : x \in \mathcal{X})$ be a positive vector such that

$$\sum_{x\in\mathcal{X}} \lambda_x = 1.$$

Let $P = (P_{xy} : x, y \in \mathcal{X})$ be a positive matrix such that, for each $x \in \mathcal{X}$,

$$\sum_{y\in\mathcal{X}} P_{xy} = 1.$$

We say a sequence of random variables $X = (X_t : t \in \mathbb{Z}_+)$ is a *discrete time Markov chain* with initial distribution λ and transition matrix P if

$$\mathbb{P}\Big(X_0 = x\Big) = \lambda_x$$

and

$$\begin{aligned}\mathbb{P}\left(X_{t+1} = x_{t+1}\middle|X_t = x_t, ..., X_0 = x_0\right) &= \mathbb{P}\left(X_{t+1} = x_{t+1}\middle|X_t = x_t\right) \\ &= P_{x_t x_{t+1}}\end{aligned} \tag{19}$$

Remark. The condition (19) is called the Markov Property and is the key defining feature of a Markov chain or, more generally, Markov process. It states that the past $(X_1, ..., X_{t-1})$ and future X_{t+1} are conditionally independent of the present X_t. Otherwise stated, is says that, when we know the past and present states $(X_1, ..., X_t) = (x_0, ..., x_t)$, the distribution of the future states $X_{t+1}, X_{t+2}, ...$ is only determined by the present state $X_t = x_t$. Think of a board game like snakes and ladders, where you go in the future is only determined by where you are now and not how you got there; this is the Markov property.

Ex 2.1. (Constructing Markov Chains) Suppose that we are given a function $f : \mathcal{X} \times [0, 1] \to \mathcal{X}$. Suppose we take X_0 a random variable with values in $\mathcal{X}$ and we define

$$X_{t+1} = f(X_t, U_t) \qquad \text{for} \quad t = 0, 1, 2, .. \tag{20}$$

where each U_t is an independent uniform $[0, 1]$ random variable.

i) Show that $(X_t)_{t\geq 0}$ is a discrete time Markov chain.

ii) Show that all discrete time Markov chains can be constructed in this way. (I.e. for every (λ, P) there exists a function f such that X defined by (20) is a discrete time Markov chain with initial stats λ and transition matrix P.)

Ex 2.2. (Markov Chains and Martingale Problems) Show that a sequence of random variables $X = (X_t : t \in \mathbb{Z}_+)$ is a Markov chain if and only if, for all bounded functions $f : \mathcal{X} \to \mathbb{R}$, the process

$$M_t^f = f(X_t) - f(X_0) - \sum_{\tau=0}^{t-1} (P - I) f(X_\tau)$$

is a Martingale with respect to the natural filtration of X. Here for any matrix, say Q, we define

$$Qf(x) := \sum_{y \in \mathcal{X}} Q_{xy} f(y).$$

Ex 2.3. (Markov Chains and Potential Theory) Let $r : \mathcal{X} \to \mathbb{R}_+$ be a bounded function.

i) Argue that for $\alpha \in (0,1)$

$$R(x) = \mathbb{E}_x\left[\sum_{t=0}^{\infty} \alpha^t r(X_t)\right]$$

solves the equation

$$R(x) = \alpha(PR)(x) + r(x), \qquad x \in \mathcal{X}.$$

[You may note further that R is the unique bounded solution to this equation.]

Further argue that is $\tilde{R}$ is a different function that satisfies

$$\tilde{R}(x) \geq \alpha(P\tilde{R})(x) + r(x), \qquad x \in \mathcal{X}.$$

then $\tilde{R}(x) \geq R(x)$ for all x.

ii) Let $\partial\mathcal{X}$ be a subset of $\mathcal{X}$ and let T be the hitting time on $\partial\mathcal{X}$ i.e. $T = \inf\{t : X_t \in \partial\mathcal{X}\}$ and take $f : \partial X \to \mathbb{R}_+$ argue that

$$R(x) = \mathbb{E}_x\left[\sum_{t<T} r(X_t) + f(X_T)\mathbb{I}\left[T < \infty\right]\right]$$

solves the equation

$$R(x) = (PR)(x) + r(x), \qquad x \notin \partial\mathcal{X} \tag{21}$$
$$R(x) = f(x), \qquad x \in \mathcal{X}. \tag{22}$$

iii) Argue that

$$R_t(x) = \mathbb{E}_x\left[\sum_{\tau=0}^{t-1} r_\tau(X_\tau) + r_t(X_t)\right], \qquad t \in \mathbb{Z}_+$$

solves the equation

$$R_{t+1}(x) = (PR_t)(x) + r_t(x)$$

(Compare the above with Bellman's equation.)

Ex 2.4. (More Markov Chains and Potential Functions) [This question is Optional]

i) Let $R(x)$ be a solution to Ex.2.3i) but with $\alpha = 1$. Argue that

$$R(x) = (Gr)(x)$$

where

$$G = \sum_{n=0}^{\infty} P^n.$$

(The Matrix G is called the Green's function for our Markov chain.)

ii) Argue that

$$G_{xy} = \frac{h_x^y}{1 - f_y}$$

where h_x^y is the hitting probability of x from y and f_y is the return probability of y.

iii) Let $R(x)$ be the solution to Ex.2.3i) with $\alpha \in (0, 1)$. Argue that

$$R(x) = (\rho_\alpha)(x)$$

where

$$\rho_\alpha(x) = \sum_{n=0}^{\infty} \alpha^n P^n$$

The functions $(\rho_\alpha : 0 < \alpha < 1)$ is called the resolvant of our Markov chain.

iv) In part Ex.2.3i) take $\alpha = 1$ and $r = 0$ and, thus take, $H(x)$ a solution to

$$H(x) = (PH)(x), \qquad x \notin \partial\mathcal{X}, \qquad H(x) = f(x), \qquad x \in \mathcal{X} \tag{23}$$

where $H(x)$ is a Harmonic function. Show that $H(X_t)$ is a Martingale.

Exercise Sheet 3

Ex 3.1. Suppose that for a discounted program there exists a stationary policy π such that for all x

$$R(x,\pi) \geq \max_{a\in\mathcal{A}} \{r(x,a) + \beta\mathbb{E}\left[R(x,\pi)\right]\} - \epsilon$$

then prove that

$$R(x,\pi) \geq W(x) - \frac{\epsilon r^*}{1-\beta}.$$

where $r^* = \max|r(x,a)|$.

Ex 3.2. Each day a machine is either working or broken. If broken, then the day is spent repairing the machine at a cost $8c$. If the machine is working, then it can be either run unattended or attended at a cost 0 or c. In each case the probability of the machine breaking is p and q respectively. Costs are discounted by $\beta \in (0,1)$.

The objective is to minimize the infinite horizon discounted cost. Letting $F(0)$ and $F(1)$ be the minimal cost starting on a day were the machine starts broken or working, respectively. Show that it is optimal to run the machine unattended iff $7p + 8q \leq \beta^{-1}$.

Ex 3.3. Here we consider a positive programming problem

a) Let $\mathcal{V}C$ give the cost function reached after one iteration of the value iteration algorithm. Argue that if $\mathcal{V}C = C$ then $C(x)$ is optimal.

b) Let $\mathcal{I}$ give the policy reached after one iteration of the policy improvement algorithm. Argue that if $\mathcal{I}\pi = \pi$ then π is optimal.

Ex 3.4. Consider a symmetric random walk on $\mathbb{Z}$. We wish to choose a time to stop that minimizes the cost $k(x) = \exp\{-x\}$ where x gives the value of the walk when stopped. Argue that $W_s(x)$ the optimal value function for the s-time horizon problem is constant for each x. Argue that the

$$\lim_{s\to\infty} W_s(x) \neq W(x)$$

where $W(x)$ is the optimal value function for the infinite time optimal stopping problem.
(Note for this Negative program we have a solution to the Bellman equation that is not optimal.)

Exercise Sheet 4

Ex 4.1 A burglar robs houses over N nights. At any night the burglar may choose to retire and thus take home his total earnings. On the tth night house he robs has a reward r_t where r_t is an iidrv with mean $\bar{r}$. Each night the probability that he is caught is p and if caught he looses all his money. Find the optimal policy for the burglar's retirement. (Hint: OLSA)

Ex 4.2 You own a "toxic" asset its value, x_t at time t, belongs to $\{1, 2, 3, ...\}$. The daily cost of holding the asset is x_t. Every day the value moves up to $x + 1$ with probability $1/2$ or otherwise remains the same at x. Further the cost of terminating the asset after holding it for t days is $C(1 - \alpha)^t$. Find the optimal policy for terminating the asset.(Hint: OLSA)

Ex 4.3 (Bruss' Odds Algorithm) You sequentially treat patients $t = 1, ..., T$ with a new trail treatment. The probability of success is $p_t = 1 - q_t$. We must minimize the number of unsuccessful treatments while treating all patients for which the trail is will be successful. (i.e. if we label 1 for success and 0 for failure, we want to stop on the last 1). Argue, using the One-Step-Look-Ahead rule that the optimal policy is the stop treating at t^* the largest integer such that

$$\frac{p_{t^*}}{q_{t^*}} + ... + \frac{p_T}{q_T} \geq 1.$$

This procedure is called *Bruss' Odds Algorithm.*

Exercise Sheet 5

Ex 5.1 (Discrete time LQ-regularization) We consider discrete time LQ minimization, here you minimize the objective

$$\min_{a_0,..,a_{T-1}} \qquad x_T' R x_T + \sum_{t=0}^{T-1} x_t' R x_t + a_t' Q a_t \qquad \text{subject to} \qquad x_t = Ax_{t-1} + B_{t-1} a_{t-1}, \quad t = 1, ..., T$$

Here R and Q are positive semi-definate matrices.

1) Show that the Bellman equation for this dynamic program is

$$L_t(x) = \min_a \{x'Rx + a'Qa + L_{t+1}(Ax + Ba)\}$$

2) Assuming the solution is of the form $L_t(x) = x'\Lambda_t x$ find the action that a minimizes the above Bellman equation.

3) Using your answer to Part 2), show that

$$\Lambda_t = R + A'\Lambda_{t+1}A - (A'\Lambda_{t+1}B)(Q + B'\Lambda_{t+1}B)^{-1}B'\Lambda_{t+1}A.$$

This is the Riccarti Recursion (the discrete time analogue of the Riccarti equation).

Utility Theory

The point of the next part is to teach some utility theory. A utility function $U(x)$ is used to quantify the value that you gain from an outcome x.

Def 76 (Utility Function). *For $\mathcal{X} \subset \mathbb{R}^d$, a utility function is a function $U : \mathcal{X} \to \mathbb{R}$ that is increasing, i.e. if $x \leq y$ component-wise then $U(x) \leq U(y)$. The utility of a random variable X is then its expected utility, $\mathbb{E}U(X)$. A utility function creates an ordering where an outcome X is preferred to Y if $\mathbb{E}U(X) \geq \mathbb{E}U(Y)$.*

Jensen's inequality applies to a concave utility:

$$\mathbb{E}U(X) \leq U(\mathbb{E}X)$$

So we prefer a certain outcome $\mathbb{E}X$ rather than the risky outcome X that has the same mean – This is being risk averse.

Def 77 (Risk Aversion). *If the function is concave then we also say that the function is risk averse. (Unless stated otherwise we assume that the utility function is risk averse).*

Def 78. *We say that two utility functions U and V are equivalent if they induce the same ordering. I.e. $\mathbb{E}U(X) \leq \mathbb{E}U(Y)$ iff $\mathbb{E}V(X) \leq \mathbb{E}V(Y)$.*

Ex 5.2 i) Define $\phi : \mathbb{R} \to \mathbb{R}$ s.t. $\phi(\mathbb{E}U(X)) = \mathbb{E}V(X)$. Show that for $p \in (0,1)$

$$\phi(pU(x) + qU(y)) = p\phi(U(x)) + (1-p)\phi(U(y)).$$

ii) Argue that ϕ is linear.
iii) Show that two utility functions are equivalent iff V the same as U up-to an affine transform, i.e.

$$V(x) = aU(x) + b$$

for constants $a > 0$ and b.

Def 79 (Coefficient of Relative Risk Aversion). *For a utility function $U : \mathbb{R} \to \mathbb{R}$ (twice differentiable) the Coefficient of Relative Risk Aversion is*

$$-x\frac{U''(x)}{U'(x)}.$$

Ex 5.3 You have utility function U. You are offered a bet that increases you wealth w multiplicatively by $(1 + X)$ here X is a "small" positive is a RV. Discuss why you would accept the bet iff

$$\frac{2\mathbb{E}X}{\mathbb{E}X^2} \geq -x\frac{U''(x)}{U'(x)}$$

I.e. You accept the bet if you mean is large but a large variance makes this less likely, and the coefficient of relative risk aversion decides the threshold.

Def 80 (CRRA Utility/Iso-elastic Utility). *A Constant Relative Risk Averse utility (CRRA) takes the form*

$$U(x) = \begin{cases} \frac{x^{1-R}}{1-R}, & R \neq 1, \\ \log x, & R = 1. \end{cases}$$

Def 81. *A utility function is Iso-elastic if it is unchanged under multiplication: for all $c > 0$,*

$$\mathbb{E}U(X) \geq \mathbb{E}U(Y) \qquad \text{iff} \qquad \mathbb{E}U(cX) \geq \mathbb{E}U(cY).$$

I.e. the utility only cares about the relative magnitude of the risk.

Ex 5.4 Show that a CRRA utility is iso-elastic.

Ex 5.5 Now suppose that U is an iso-elastic utility. i) Use Ex 5.1 to argue that $\forall c$, $U(cx) = a_c U(x) + b_c$ for constants a_c and b_c.

ii) Show that

$$\frac{cU''(cx)}{U'(cx)} = \frac{U''(x)}{U'(x)}$$

iii) Integrate the above to show that any iso-elastic utility is a CRRA utility.

Exercise Sheet 6

The following questions use the same notation section of your notes on the infinite time Merton problem. The following questions ask you to do several calculations used in your notes.

Ex. 6.1 Show that the HJB equation for the Merton Problem is

$$0 = \sup_{\theta,c} \left\{ e^{-\rho t} \frac{c_t^{1-R}}{1-R} + \partial_t V + (rw + \theta \cdot (\mu - r) - c)\, \partial_w V + \frac{1}{2} |\sigma^{\mathsf{T}} \theta|^2 \partial_{ww} V \right\}.$$

Ex. 6.2 Show that the minimum over θ of the above optimization is given by

$$\theta = -\frac{\partial_w V}{\partial_{ww} V} (\sigma \sigma^{\mathsf{T}})^{-1} (\mu - r)$$

Ex. 6.3 For a concave increasing function u, we define the Legendre-Fenchel transform by

$$u^*(y) = \sup_c \{u(c) - cy\}$$

i) Show that for

$$u(c) = \frac{c^{1-R}}{1-R}$$

that

$$u^*(y) = \frac{y^{1-R^{-1}}}{1-R^{-1}}.$$

and that the above supreme is optimized by $c = y^{-\frac{1}{R}}$.

ii) [Optional slightly harder question] Now suppose u a concave differentiable function with derviative whose derivative has range $(0, \infty)$, argue that

$$(u^*)^*(c) = c$$

Ex. 6.4 Verify that

$$Y_t = \zeta_t W_t + \int_0^t \zeta_s c_s ds$$

(as given in the course) is a local-martingale.

Ex. 6.5 Verify that

$$w_0 = \mathbb{E} \left[\int_0^\infty \zeta_s c_s^* ds \right]$$

where ζ_t and c_t^* are as given in the course.

Exercise Sheet 7

Ex. 7.1 For the Interest Rate Risk example (as given in the course), show that the HJB eqution is

$$0 = \sup_{c,\theta} \Big\{ u(c) - \rho V + \frac{1}{2}\sigma^2\theta^2 \partial_{ww} V + \sigma\sigma_r \eta\theta\partial_{wr} V + \frac{1}{2}\sigma_r^2 \partial_{rr} V + (rw + \theta(\mu - r) - c)\, \partial_w V + \beta(\bar{r} - r)\partial_r V$$

Ex. 7.2 Show for the transaction costs example (as given in the course) that upto a local Martingale

$$\begin{aligned} e^{\rho t} dZ_t = & \Big\{ -\rho V + (Rx - c)\partial_x V + \mu y \partial_y V + \frac{1}{2}\sigma^2 y^2 \partial_{yy} V - u(c) \Big\} dt \\ & + [\partial y V - (1+\epsilon)\partial_x V]\, dL_t + \Big[(1-\epsilon)\partial_x V - \partial_y V\Big] dM_t . \end{aligned}$$

Ex. 7.3 For the example of optimization under drawdown constraints we found the HJB equations

$$\begin{aligned} & \sup_{\theta,c} \Big\{ u(c) - \rho V + \frac{1}{2}\sigma^2\theta^2\partial_{ww} V + (r(w-\theta) + \mu\theta - c)\partial_w V \Big\} = 0 \\ & \partial_{w^*} V = 0 \end{aligned}$$

Show that after optimization over θ and c the HJB equation becomes

$$\tilde{u}(\partial_w V) - \rho V + rw\partial_w V - \frac{1}{2}\kappa^2 \frac{(\partial_w V)^2}{\partial_{ww} V} = 0$$

where $\kappa = \sigma^{-1}(\mu - r)$.

Ex. 7.4 [Harder/Optional Question] Consider the insurance with Premium control example, here:

$$dW_t = rW_t dt + \theta_t \{\sigma dB_t + (\mu - r)dt\} - c_t dt + q_t p_t dt - dC_t$$

where for compound Poisson process Y we have that

$$C_t = Y\Big(\int_0^t q_s ds\Big).$$

i) Heuristically show that Itô's formula for a process (with jumps) is

$$dF(W) = \partial_w F(W) dW + \frac{1}{2}\partial_{ww} F(W) d[W] + [F(W) - F(W^-)] dJ(W)$$

Here $W_t^- = \lim_{s \nearrow t} W_s$ is the left limit of W_t, and here $J(W)_t$ gives the jumps of W, i.e.

$$J(W)_t = \sum_{s \leq t} (W_s - W_s^-)$$

(This sum is well defined since there are at most countably many jumps.)

ii) Form this argue that the HJB equation for the Insurance Preium Control example is

$$0 = \sup_{\theta,q,c} \Bigg\{ u(c) - \rho V(w) + \{rw + \theta(\mu - r) - cqp(q)\}\, \partial_w V + \frac{1}{2}\sigma^2\theta^2\partial_w V + q \int_0^\infty V(w - x) - V(w) dF(x)$$

Exercise Sheet 8

Ex. 8.1[Hoeffding's Inequality] Suppose that X is a bounded RV, i.e. $|X| < c$, with mean zero then show that

$$\mathbb{E}[\exp\{\theta X\}] \leq \cosh(c\theta) \leq \exp\{\theta^2 c^2\}$$

Ex. 8.2[Azuma-Hoeffding Inequality] Suppose that M_n is a discrete time Martingale such that $|M_n - M_{n-1}| \leq c$ for all c then show that

$$\mathbb{P}(M_n \geq x) \leq \exp\Big\{-\frac{x^2}{2nc}\Big\}.$$

[Hint: apply Markov's inequality to

$$\mathbb{P}(\exp\{\theta M_n\} \geq \exp\{\theta x\})$$

and then use Ex 8.1]

Ex. 8.3 In this question we consider the weighted majority algorithm:

i) Show that for r a RV in the interval $[0,1]$ and $\eta > 0$

$$\mathbb{E}e^{\eta r} \leq \mathbb{E}[(e^{\eta} - 1)r + 1] \leq \exp\{e^{\eta} - 1\mathbb{E}[r]\}$$

[Hint: find lines that bound e^x from above and below.]

ii) Following your proof for the weighted majority algorithm and applying the upper bound above show that

$$\exp\{\eta \max_i R(i,T)\} \leq N \exp\{(e^{\eta} - 1)\mathbb{E}R(\pi,T)\}$$

iii) Thus show that for η fixed the weighted majority algorithm satisfies the bound

$$R(\pi,T) \geq \frac{1}{1-e^{-\eta}}\Big\{\eta \max_i R(i,T) + \log N\Big\}.$$

Index