Continuous Time Finance, Spring 2018
NYU Courant Institute
Introduction to Stochastic Optimal Control

Monty Essid
DRAFT VERSION

April 11, 2018

Abstract

These notes provide an introduction to the theory of stochastic optimal control. Rather than focusing on the many (but interesting) regularity details of the theory, our treatment will remain rather formal and will focus on the main ideas, as well as examples.

For a brief lecture, it is recommended to read the most important Sections 1.1, 2 ,3.2, 4 and 5, which are highlighted with a *.

For a much better and more complete treatment, please review classical references such as [1],[2], or a more complete introduction found in [3],[4].

Contents

1 Examples

Before formally discussing the general problem, let us introduce it with some characterstic examples explaining our goal.

1.1 Optimal Investment/Consumption*

Assume that you have some wealth x\$ today, and that you can only invest in a risky asset (say a stock) S_t or a risk free asset (say a bond) B_t.

The goal is to find an investment and consumption strategy that maximizes a certain utility function.

This utility function represents the degree of satisfaction you get by spending you money between some starting time t_0 and final time T, as well as potentially a legacy you leave behind at T.

More formally, we will assume that the assets follow:

$$\begin{aligned} dB_t &= rB_t dt, \quad \text{Risk free asset with interest rate } r \\ dS_t &= \mu_t S_t dt + \sigma_t S_t dW_t, \quad \text{Risky asset} \end{aligned}$$

where W_t is a Brownian motion under a probability measure $\mathbb{P}$, generating a filtration $\mathcal{F}_t$. In these notes, we will restrict ourselves to μ_t, σ_t to be deterministic functions of t and S_t.

Denote by X_t the value of the portfolio at time t. We start at $X_{t_0} = x$.

The first **control/strategy/policy/decision** you can take is the proportion $\theta_t \in [0,1]$ of your portfolio invested in the risky asset.

The value of your portfolio at time t is thus given by:

$$X_t = \frac{\theta_t X_t}{S_t} S_t + \frac{(1-\theta_t)X_t}{B_t} B_t$$

where $\frac{\theta_t X_t}{S_t}, \frac{(1-\theta_t)X_t}{B_t}$ represents the number of shares of respectively S_t and B_t held in the portfolio at time t.

Obviously, the value of these controls at time t cannot depend on future values of S_t, B_t, otherwise the optimal strategy would require you to read the future to make a decision today; the controls must be **non-anticipating**.

In this example, we are only interested in 'self financing strategies'; between t and $t+dt$, the number of shares of S_t or B_t is held fixed, and there is no external injection of money.

If we do not consume anything, one would gets:

$$dX_t = \frac{\theta_t X_t}{S_t} dS_t + \frac{(1-\theta_t)X_t}{B_t} dB_t$$

The second control you can choose is the rate of consumption per unit time $c_t \geq 0$.

Our controls (θ_t, c_t) are both constrained to lie in the **admissible** set $A_t = [0,1] \times \mathbb{R}_+$, for all t.

The evolution of our wealth is hence given by:

$$dX_t = \frac{\theta_t X_t}{S_t} dS_t + \frac{(1-\theta_t)X_t}{B_t} dB_t - c_t dt$$

which is after rearranging the terms:

$$dX_t = \left[(\theta_t \mu_t + (1-\theta_t)r)X_t - c_t\right] dt + \theta_t \sigma_t X_t dW_t$$

The above equation is called the **state equation**: it quantifies the influence of the controls θ_t, c_t on the dynamics of our state X_t. Note that X_t depends on the controls θ, c and will sometimes be denoted as $X_t^{\theta,c}$.

Finally, we need to prescribe what our **objective function/value function/reward** is, which will quantify the value of each strategy.

For a given state t, x (i.e. $X_t = x$), and a strategy $\alpha = (\theta, c) \in A = (A_s)_{s\in[t,T]}$, one could consider:

$$V(t, x; \alpha) = \mathbb{E}_{X_t=x}\left[\int_t^T e^{-r(s-t)} c_s^\gamma ds + e^{-r(T-t)} \left(X_T^\alpha\right)^\gamma\right]$$

for some constant $\gamma \in (0, 1)$.

The ideas behind such an objective functions are:

1. The first term of the value function is an integral summing the instantaneous reward $e^{-r(s-t)}c_s^\gamma$ per unit time for every time step. This term encourages short sighted actions by increasing consumption, and thus tends to decrease the amount of money that we invest in the assets.

2. The second term is a final time rewards $e^{-r(T-t)}\left(X_T^\alpha\right)^\gamma$. This term favors long sighted action that consists in investing intelligently in the assets in order to maximize the legacy at he end, and thus decrease the instantaneous consumption.

3. All the quantities are discounted; It is more satisfying to consume something today than next year. The legacy is also discounted, as having X_T in the future is worth having $e^{-r(T-t)}X_T$ today.

4. The utility of consuming or the legacy is *increasing* and *concave*; increasing because consuming/leaving more generally yields more satisfaction. Concave because consuming/leaving an extra dollar on top of 2\$ usually yields a higher satisfaction than consuming an extra dollar on top of $10,000$\$.

5. The value function is taken as an expectation, due to the random fluctuations that the risky asset experiences.

The **stochastic optimal control problem** seeks the control $\alpha^* = (\theta^*, c^*)$ that maximizes the value function, as well as the maximum value function achievable:

$$u(t_0, x) = \max_{\alpha\in A} V(t_0, x; \alpha) = \max_{\alpha\in A} \mathbb{E}_{X_{t_0}=x}\left[\int_{t_0}^T e^{-r(s-t_0)} c_s^\gamma ds + e^{-r(T-t_0)}\left(X_T^\alpha\right)^\gamma\right]$$

Note that we didn't impose any banckrupcy constraint in this problem; to satisfy the constraint $X_t \geq 0$, we can replace the final time T by a stopping time $\tau = \min(T, \tau_0)$ where $\tau_0 = \inf\{s > t | X_s = 0\}$.

We can even get more creative and introduce a second random variable τ_D, independent of the process X_t, representing the death of the investor. For example, $\tau_D \sim \exp(\lambda)$.

This problem is solved in Section 8.

1.2 Vehicle steering

Let $X_t \in \mathbb{R}^d$ (say $d = 2$ or 3) be the position of a vehicle, say a boat on the sea or a rocket in the air.

Starting from some point x at t_0, one would like to arrive at to some point y at time T with minimum energy.

The **controls** for the vehicle are its direction and velocity $v_t \in \mathbb{R}^d$, as well as a stabilizer $\gamma_t \in (0, 1]$ than somewhat cancels the turbulences (due to a choppy sea, air pockets etc.).

The **state equation** is thus given by:

$$dX_t = v_t dt + \gamma_t \sigma_t dW_t$$

where σ_t is an external parameter modelling the roughness of the turbulence. We will assume that σ_t is a deterministic function of t and X_t. For example, one can take it constant.

Given a set of controls $\alpha = (v, \gamma) \in A$ for the **admissible set** $A = (A_s)_{s\in[t,T]}$, with $A_s = \mathbb{R}^d \times (0, 1]$, one can consider the **value function**:

$$V(t, x; \alpha) = \mathbb{E}_{X_t=x}\left[\int_t^T \left(m\frac{|v_s|^2}{2} + \frac{c_1}{\gamma_s}\right) ds + c_2|y - X_T^\alpha|^2\right]$$

for some constants $m, c_1, c_2 > 0$. The first term represents the sum of the kinetic energy, which is the energy required to steer the vehicle, as well as the energy required to stabilize it. The last term penalizes large displacements from the target y.

The **stochastic optimal control** problem seeks the optimal control α^* that minimizes the value function, i.e.

$$u(t_0, x) = \min_{\alpha \in A} V(t, x; \alpha) = \min_{\alpha \in A} \mathbb{E}_{X_{t_0}=x} \left[\int_{t_0}^{T} \left(m \frac{|v_s|^2}{2} + \frac{c_1}{\gamma_s} \right) ds + c_2 |y - X_T^{\alpha}|^2 \right]$$

1.3 Market Making problem

This toy example is taken from the excellent presentation D. Borden made [5]. All mistakes in this paragraph are my own.

More detailed work on this topic can be found in the paper by M. Avellaneda and S. Stoikov [6] for example.

Before setting up the optimal control problem, we will briefly review the setting.

In this simplified view, assume we are only trading one asset S_t, at high frequencies, with no market impact.

Because of the high frequency trading, we will model the stock by a standard Brownian motion with deterministic drift $\mu(t)$ and constant volatility σ;

$$dS_t = \mu(t)dt + \sigma dW_t \tag{S1}$$

Real markets do not store single prices; rather they store an *order book* recording all prices and their respective quantities at which someone willing to sell (*ask price*), as well as the prices and quantities at which someone is willing to buy (*bid price*).

We will assume that the highest bid price is less than the lowest ask price, otherwise the seller and buyers will be matched and these quantities will disappear from the book.

The dynamics above will represent the *mid-price*; it is a fictitious quantity, which is the average of the ask and bid prices.

Finally, assume in this overly simplistic model that the ask price $S_t^+ = S_t + \delta$, and the bid price is $S_t^- = S_t - \delta$, for some *constant* $\delta > 0$. The *spread* is the quantity $S_t^+ - S_t^- = 2\delta$, is constant here.

A trader can place two types of order;

1. Market order: this order buys at the ask price or sells at the bid price.

   In the case that you are buying at the ask, it means that you are willing to pay the spread and buy at $S_t^+ = S_t^- + 2\delta$ (compared to your closest rival willing to buy at S_t^-) in order to guarantee getting the asset right now.

   Similarly, if you were selling at the bid, you are also paying the spread to guarantee selling you asset right now.

2. Limit order: this order buys (resp. sells) as soon as the ask (resp. bid) crosses a pre-specified threshold. One might have to wait a long time for the limit order to get executed or *filled*; an example of a limit order would be to wait until the price of an iphone drops below 599$ to buy it.

   Say that you are issuing limit orders to buy at the bid, and sell at the ask, just for a small period of time dt. Contrary to the market orders, limit orders *earn* spread.

To complete the dynamical picture of the market, we will assume that:

- Market orders are filled at a constant rate f_0

- Limit orders are filled with rates:

$$\begin{aligned} d\nu_s &= \beta(\nu_0 - \nu_s)dt + \xi dW_t \quad \text{for buy orders} \\ d\nu_b &= \beta(\nu_0 - \nu_b)dt - \xi dW_t \quad \text{for sell orders} \end{aligned} \tag{S2}$$

  where $\beta, \xi > 0$ and $\nu_0 < f_0$.

  The ideas behind this simple execution rates model are

  - Limit order's fillings oscillate around a rate ν_0, with a characteristic amplitude of β
  - This rate ν_0 is smaller than f_0, as a market order has more chances of getting filled
  - Executed buy and sell limit orders have a negative correlation proportional to ξ^2; if a lot of your buy limit orders are filled, it means that a lot of people are selling the asset to you. This probably means that the bid price you entered is attractive to them, and hence not many people are willing to buy at your ask price (larger than your bid price).

We can finally start setting up the optimal control problem;

First our **controls** will be whether to set a buy $\lambda_b(t) \in \{0,1\}$ or sell $\lambda_s(t) \in \{0,1\}$ limt order at time t (and kill it after a short interval dt if it does not get filled), or set a buy $m_b \in \{0,1\}$ or sell $m_s \in \{0,1\}$ market order at time t.

Call $\alpha = (\lambda_s, \lambda_b, m_s, m_b)$ our control, which is in the **admissible** set A, meaning $\alpha_t \in A_t = \{0,1\}^4$ for all t.

Define Δ to be the number of shares of the asset we hold.

Then from the dynamics of the market, we have that the evolution of the number of shares of our portfolio is given by

$$d\Delta_t = \left[m_b(t)f_0 - m_s(t)f_0 + \lambda_b(t)\nu_b(t) - \lambda_s(t)\nu_s(t)\right] dt \tag{S3}$$

Our **state** is given by $X_t = (S_t, \nu_b(t), \nu_s(t), \Delta_t)$, and one can deduce a vector valued **state equation** of the type

$$dX_t = h(X_t, \alpha_t)dt + \Sigma dW_t$$

using equations (S1),(S2),(S3), for some determnisitic function h and constant Σ.

Given an initial state $X_t = x \in \mathbb{R}^4_+$ and a control $\alpha_t = (\lambda_s(t'), \lambda_b(t'), m_s(t'), m_b(t')) \in A_{t'}$, $\forall t' \in [t,T]$, one can define the **value function** to be

$$V(t,x;\alpha) = \mathbb{E}_{X_t=x}\left[\int_t^T \left(\underbrace{-m_b(t')(S_{t'}+\delta)f_0 + m_s(t')(S_{t'}-\delta)f_0 - \lambda_b(t')(S_{t'}-\delta)\nu_b(t') + \lambda_s t'(S_{t'}+\delta)\nu_s(t')}_{(1)} - \underbrace{\eta\sigma^2\Delta^2_{t'}}_{(2)}\right) dt' - \underbrace{\Delta^2_T}_{(3)}\right]$$

There are three terms in this value function:

Term (1) corresponds to the PnL of our strategy, summed over the interval $[t,T]$.

Term (2) represents risk aversion (η is a constant); holding a lot of stock can be very risky.

Term (3) represents the fact that we would like to hold the least possible amount of stock at the end of the day.

The market maker problem is to find a control α^*, that maximizes

$$u(t,x) = \max_{\alpha_{t'} \in A_{t'}, \forall t'} V(t,x;\alpha)$$

1.4 Grid World game

This is an example of a discrete time (and discrete space) optimal control problem, where the final time T is not fixed.

In this game, we are given a grid of points $G \subset \mathbb{Z}^2$.

The player starts at some position $x \in G$, and is free to move on G. In practical examples, the grid is bounded.

The goal is to reach a target point $o \in G$, the fastest possible.

To complicate the task, obstacles and pits are randomly positioned on the grid; obstacles do not allow you to select their position, but pits do at high cost.

Our **state** is given by our position on the grid x.

The **control** a here is *deterministic*, and represents the choice of direction one would like to take on the grid; up, right, left, down. One could also consider a *stochastic* problem where the player is confused and follows the control with probability p and chooses a random direction with probability $1-p$, which we won't do here.

Note that we cannot always choose any direction for the control a at any time; the available directions will depend on our state x, and more particularly on whether there are obstacles on the grid points around x; the **admissible set** of strategies A depends on the positon.

The **state equation** here is simple, and described by a *deterministic* function D:

$$x_{t+1} = D(x_t, a)$$

where $D(x_t, a) = x_t + (0, 1)$ if $a =$up, etc. .If we had randomness, we would require a stochastic function, given by a transition kernel describing probabilities of ending at each states given the previous position and the action taken.

The **reward** is composed of two terms; a *nonnpositive* reward of F which penalizes using pits (negative), or zero otherwise, and a positive constant reward $g > 0$ for reaching the target t. Besides, we will use a discounting factor $0 < \beta < 1$ to encourage the player to reach o.

Define $\tau = \inf\{t \in \mathbb{N} | x_t = o\}$ to be the first time that we reach o, given that we start at some point x.

Given a startegy of decisions α based on our current state x_t ($\alpha(x_t) = a$), one can define the **total reward** for that strategy given our initial position x:

$$V(x; \alpha) = \sum_{t=0}^{\tau-1} F(x_t)\beta^t + \beta^\tau g$$

where $x_0 = x$ and $x_{t+1} = D(x_t, \alpha(x_t))$. Again, $F(x_t) = c < 0$ if x_t is at a pit, and zero otherwise.

One seeks the optimal strategy α^* and the optmial value function u such that

$$u(x) = \max_{\alpha \in A} V(x; \alpha)$$

Note that this problem doesn't have a fixed horizon, as it is described using stopping times. More on that in Section 6.

This example is a typical Reinforcement Learning problem, and many efficient numerical algorithms exists to solve it. See [7] for more details.

2 Formal Problem*

More generally, one can study the following problem: Given

1. An initial position $(t_0, x) \in \mathbb{R}_+ \times \mathbb{R}^d$ and a final time $T \leq +\infty$.
2. A set of **admissible controls/strategies/policies/decisions** $\alpha \in A$.

   Usually, this means that for each $t \in [t_0, T]$, $\alpha_t \in A_t$ where A_t is the set of admissible values for the control at time t. A is the collection of controls such that at every time $t \in [t_0, T]$, the constraints imposed by A_t are satisfied. Sometimes we write $A = A_{[t_0,T]}$. We can even have A_t depending on our state, which limits the possible choices of controls in different circumstances.

Besides, it is common to impose that the controls are **non-anticipating**; the decision for the next step does not depend on the future.

3. A **state equation** of the type

$$dX_t^\alpha = \mu(t, X_t^\alpha, \alpha_t)dt + \sigma(t, X_t^\alpha, \alpha_t)dW_t$$

where μ, σ are *deterministic* functions, and W_t is a Brownian motion under some probability measure $\mathbb{P}$ generating a filtration $\mathcal{F}_t$.

The dependence of X_t^α on α has been made explicit in the superscript, but will sometimes be omitted.

4. A **value/reward/objective function** for each control $\alpha \in A$, and each position t, x

$$V(t, x; \alpha) = \mathbb{E}_{X_t=x}\left[\int_t^T F(s, X_s^\alpha, \alpha_s)ds + g(X_T^\alpha)\right]$$

where F, g are determinsitic functions.

Set

$$u(t, x) = \max_{\alpha \in A} / \min_{\alpha \in A} V(t, x; \alpha) = \max_{\alpha \in A} / \min_{\alpha \in A} \mathbb{E}_{X_t=x}\left[\int_t^T F(s, X_s^\alpha, \alpha_s)ds + g(X_T^\alpha)\right]$$

The goal is to maximize or minimze the value function over all admissible controls; find u and the optimal control α^*.

If we manage to find $\alpha_t^* = h(t, X_t)$ for some deterministic function h, then we have a **feedback law**; at each state t, x, one can find the optimal strategy α^* for the next step by taking $\alpha^* = h(t, x)$.

In the what follows, we will usually maximize the value function, unless stated otherwise. All the ideas remain the same by replacing max by min in the minimization case.

We will also keep the notations consistent with this section.

3 Dynamic programming principle*

The cornerstone idea in optimal control is the **Dynamic Programming Principle/Bellman principle** [1], referred as the DPP hereafter.

3.1 Intuition

As an example, imagine one is seeking the shortest path from a point x to a point y. Then one can take a path p starting at $x = p(0)$ and ending at some intermediate point $z = p(1)$ with distance dist(p). Then from that position, one can seek the shortest path from $z = p(1)$ to y, with distance shortestpath(z, y).

The shortest path from x to y is thus given by the best path p minimizing the sum of the two quantities;

$$shortestpath(x, y) = \min_p\{dist(x, p(1)) + shortestpath(p(1), y)\}$$

This is a DPP equation for the problem: it divides a big task (going from x to y) into two smaller tasks (going from x to z then z to y). It might be not completely obvious at first to see why this can drastically simplify the problem.

But why does this work? The shortest path from $p(1)$ to y does not depend on the path p one took to go from x to $p(1)$. It only depends on its final position $p(1)$.

A simple way of understanding the DPP using the above example is the following: Say that you found the shortest path that goes from x to y, and this path happens to go through z at some point.

[1] One could also use a slightly more general but harder to exploit principle called Pontryagin's maximum principle.

This shortest path will also give you a path from z to y, and it has to be optimal! Otherwise the path from x to y wouldn't be optimal to begin with.

We can easily see how to break the DPP if we don't have Markovian dynamics; Say that you are running from x to y, and for every intermediary points z there are two ways of going to y; a short one by going at the top of a hill, and another one by continuing on a flat but longer path.

If you started at the middle point z, then the choice would be clear; run above the hill to arrive at y.

But let's say that if we came all the way from x, we are very tired and attempting to go over the hill will actually slow us down, as we would need to take a break.

Thus the shortest path from x to y involves using an intermediary point z, and a path from z to y along the flat longer terrain, which is not the optimal path going from z to y. So the DPP doesn't work in this situation.

The reason why the DPP fails is simple; the choice of the optimal control does not depend on our current position, but rather all the past information that we have to keep track of.

It is of course easy to fix the above problem; expand our state space to include a second variable $\epsilon = 0$ or 1 describing whether we have enough energy to run above the hill. Then the state equation on the variables (position,ϵ), along with the optimal control will describe a Markov process, and we can use the DPP again to solve the problem.

It is thus important to realize that the DPP only works if the optimal control is a **feedback control**; $\alpha^*(t, X_t)$ is a function of the current state only.

If no such control is optimal for our problem, then the DPP will not yield the correct solution. In those situations, one can try to extend the state space to store 'important' past information and hope that such a feedback control exists.

3.2 The DPP*

Using the notations of Section 2, the DPP states that

$$\underbrace{u(t,x)}_{\substack{\text{best achievable reward}\\ \text{from state } (t,x)}} = \max_{\alpha \in A_{[t,t_1]}} \mathbb{E}_{X_t=x}\left[\underbrace{\int_t^{t_1} F(s, X_s^\alpha, \alpha_s) ds}_{\substack{\text{reward obtained between}\\ \text{t and } t_1 \text{ using the strategy } \alpha}} + \underbrace{u(t_1, X_{t_1}^\alpha)}_{\substack{\text{best achievable reward}\\ \text{from state } (t_1, X_{t_1}^\alpha)}} \right]$$

for any $t_1 \in [t, T]$.

A few **important remarks** should be pointed out:

Remark. 1. It is common in financial problems that we have functions F, g that also depends on the initial point t, because of discounting;

$$u(t,x) = \max_{\alpha \in A} \mathbb{E}_{X_t=x}\left[\int_t^T e^{-r(s-t)} f(s, X_s^\alpha, \alpha_s) ds + e^{-r(T-t)} g(X_T^\alpha) \right]$$

in this example, $F(s, X_s^\alpha, \alpha_s; t) = e^{-r(s-t)} f(s, X_s^\alpha, \alpha_s)$ and we replaced g by $e^{-r(T-t)} g$.

In that case, the DPP becomes

$$u(t,x) = \max_{\alpha \in A_{[t,t_1]}} \mathbb{E}_{X_t=x}\left[\int_t^{t_1} e^{-r(s-t)} f(s, X_s^\alpha, \alpha_s) ds + e^{-r(t_1-t)} u(t_1, X_{t_1}^\alpha) \right]$$

Notice that the only change stems from the fact that we had to discount the best achievable expected reward $u(t_1, X_{t_1}^\alpha)$.

2. If the state equation does not **explicitly** depends on time, and neither does the reward, then neither does u;

$$u(x) = \max_{\alpha \in A_{[0,t_1]}} \mathbb{E}_{X_0=x}\left[\int_0^{t_1} F(X_s^\alpha, \alpha_s)ds + u(X_{t_1}^\alpha)\right]$$

These types of DPP are used when we stop playing as soon as a certain condition is verifed (one can use stopping times), rather than at a fixed time in the future. More on that topic in Section 6.

To use the shortest path example again, this means that if our dynamics and reward does not depend on time (no rush hour, tiring effects due to the sun, changing traffic rules etc.), then finding the shortest path at 2am should be the same as if we started at 5pm.

3. As explained in the previous example, the DPP will only work if the optimal control α_t^* is in **feedback** form; it only depends on your current state t, X_t, and not on the whole history prior to t.

Heuristics of the DPP Why is the DPP true? Choosing a control $\alpha \in A_{[t,T]}$ can be divided into two parts; fix some intermediate time $t < t_1 < T$. One gets what to do in the interval of time $[t, t_1]$ by selecting $\alpha_1 \in A_{[t,t_1]}$, and similarly choose $\alpha_2 \in A_{[t_1,T]}$.

Let's try to give a heursitic explanation for the DPP, in the discounted version case;

Then

$$\begin{aligned}
u(t,x) &= \max_{\alpha \in A} \mathbb{E}_{X_t=x}\left[\int_t^T e^{-r(s-t)} f(s, X_s^\alpha, \alpha_s)ds + e^{-r(T-t)} g(X_T^\alpha)\right] \\
&= \max_{\alpha_1 \in A_{[t,t_1]}, \alpha_2 \in A_{[t_1,T]}} \mathbb{E}_{X_t=x}\left[\int_t^{t_1} e^{-r(s-t)} f(s, X_s^{\alpha_1}, \alpha_{1s})ds \right. \\
&\qquad \left. + e^{-r(t_1-t)}\left(\int_{t_1}^T e^{-r(s-t_1)} f(s, X_s^\alpha, \alpha_{2s})ds + e^{-r(T-t_1)} g(X_T^\alpha)\right)\right] \\
&= \max_{\alpha_1 \in A_{[t,t_1]}} \left\{\mathbb{E}_{X_t=x}\left[\int_t^{t_1} e^{-r(s-t)} f(s, X_s^{\alpha_1}, \alpha_{1s})ds\right] \right. \\
&\qquad \left. + e^{-r(t_1-t)} \max_{\alpha_2 \in A_{[t_1,T]}} \mathbb{E}_{X_t=x}\left[\mathbb{E}_{X_{t_1}^{\alpha_1}}\left[\int_{t_1}^T e^{-r(s-t_1)} f(s, X_s^\alpha, \alpha_{2s})ds + e^{-r(T-t_1)} g(X_T^\alpha)\right]\right]\right\}
\end{aligned}$$

where we used the tower property for iterated expectations in the last term, along with the Markov property to express them in terms of positions instead of filtrations.

Now for a fixed strategy $\alpha_1 \in A_{[t,t_1]}$, because we are assuming that a feedback control exists, the state equation is Markovian; the evolution of $(X_s^\alpha)_{s\in[t_1,T]}$ only depends on the starting point $X_{t_1}^{\alpha_1}$ and the strategy $\alpha_2 \in A_{[t_1,T]}$.

In other words, given the state equation of X_t, the point $X_{t_1}^{\alpha_1}$ and the strategy α_2, it is perfectly possible to simulate a path $X_s^{\alpha_2}$ for $s > t_1$, as both the strategy and the state equation do not require information prior to t_1.

Hence the second term can be expressed only using α_2 given $X_{t_1}^{\alpha_1}$ and we get that

$$\begin{aligned}
u(t,x) = \max_{\alpha_1 \in A_{[t,t_1]}} \left\{\mathbb{E}_{X_t=x}\left[\int_t^{t_1} e^{-r(s-t)} f(s, X_s^{\alpha_1}, \alpha_{1s})ds \right.\right. \\
\left.\left. + e^{-r(t_1-t)} \max_{\alpha_2 \in A_{[t_1,T]}} \mathbb{E}_{X_{t_1}^{\alpha_1}}\left[\int_{t_1}^T e^{-r(s-t_1)} f(s, X_s^{\alpha_2}, \alpha_{2s})ds + e^{-r(T-t_1)} g(X_T^{\alpha_2})\right]\right]\right\}
\end{aligned}$$

Which yields the desired formula:

$$u(t,x) = \max_{\alpha_1 \in A_{[t,t_1]}} \mathbb{E}_{X_t=x}\left[\int_t^{t_1} e^{-r(s-t)} f(s, X_s^{\alpha_1}, \alpha_{1s})ds + e^{-r(t_1-t)} u(t_1, X_{t_1}^{\alpha_1})\right]$$

3.3 The DPP in discrete time

Consider instead a discrete time version of the problem, where $n = 0, 1, ..., N$. The state space and/or control space are free to be discrete or continuous.

Given a current state S_t and a decision a_t at that state, we transition to the state S_{t+1} with a probability given by a transition Kernel $P(S_{t+1}|S_t, a_t)$; this is our discrete time **state equation** and defines a Markov process.

At each time t, we recieve a reward $Q(s, a)$ if we are in the state s and decided action a at time t. This reward migh also be random, hence one has to expand the transition Kernel to also take into account for randomness of Q; the probability of getting a reward transitioning to state s' and getting a reward q given that we are in state s and took a decision a is given by $P(s', q|s, a)$.

At the end of the day, we also potentially receive a reward $g(s)$ for ending at state s.

Given a policy $(a_t)_{t=1,...,N}$, the value function is

$$V(s; (a_t)) = \mathbb{E}_{S_0=s}\left[\sum_{t=0}^{N-1} \beta^t Q(S_t^a, a_t) + \beta^N g(S_N^a)\right]$$

Where the expectation uses the transition Kernels, and $0 < \beta < 1$ is a discount factor.

We are seeking to maximize the reward, that is the policy a^* that maximizes

$$u(s) = \max_{a \in A} \mathbb{E}_{S_0=s}[V(s; (a_t))]$$

Similarly to the continuous time version, the DPP in discrete times gives

$$u(s) = \max_{a \in A_1} \mathbb{E}_{S_0=s}[Q(s, a) + \beta u(S_1)] = \max_{a \in A_1} \sum_{s',q} [q + \beta u(s')] p(s', q|s, a)$$

In reinforcement learning for example, we try to solve such equations but we do not always know the theoretical expression of the transition kernel or the reward function; we are limited to estimations via many Monte-Carlo simulations, in which we also try to devise optimal strategies.

Many algorithms based on iterative procedures, Monte Carlo simulations and/or estimations using neural nets have recently been developped to solve these discrete DPP equations.

A good overview of this can be found in the Reinforcement Learning litterature, for example [7].

4 Hamilton-Jacobi-Bellman PDE*

We will now present the second cornerstone result in stochastic optimal control.

The DPP is a very general principle, that can be applied to a variety of problems.

Stochastic optimal control problems as defined in section 2 have more structure, and one can further exploit the DPP to get a nonlinear PDE solved by the optimal value function u.

In this section, we will on purpose be very loose on the details, as the next section (5) will provide a more rigorous justification of the heuristics we present here.

The DPP applied to stochastic optimal control problems yields in full generality:

$$u(t, x) = \max_{\alpha \in A_{[t,t_1]}} \mathbb{E}_{X_t=x}\left[\int_t^{t_1} e^{-r(s-t)} f(s, X_s^\alpha, \alpha_s) ds + e^{-r(t_1-t)} u(t_1, X_{t_1}^\alpha)\right]$$

for any intermediate time $t_1 \in [t, T]$. Choose t_1 of the type $t + dt$, for a 'small' dt. As $dt \to 0$, the choice of the control on the period $[t, t + dt]$ gets reduced to the choice of a unique vector valued decision $\alpha \in A_t$. Thus

$$
\begin{aligned}
u(t,x) &= \max_{\alpha \in A_{[t,t+dt]}} \mathbb{E}_{X_t=x}\left[\int_t^{t+dt} e^{-r(s-t)} f(s, X_s^\alpha, \alpha_s) ds + e^{-rdt} u(t+dt, X_{t+dt}^\alpha)\right] \\
&= \max_{\alpha \in A_{[t,t+dt]}} \mathbb{E}_{X_t=x}\left[f(t,X_t,\alpha_t)dt + (1-rdt)\left(u(t,X_t) + (\partial_t u(t,X_t) + \mathcal{L}^{\alpha_t} u(t,X_t))dt\right.\right. \\
&\qquad\qquad \left.\left. + \ \sigma(t,X_t,\alpha_t)\partial_x u(t,X_t) dW_t\right)\right] + o(dt) \\
&= u(t,x) + [\partial_t u(t,x) - ru(t,x)]dt + \max_{\alpha \in A_t}\left[f(t,x,\alpha_t) + \mathcal{L}^{\alpha_t} u(t,x)\right] dt + o(dt)
\end{aligned}
$$

where we used Ito's lemma on $u(t+dt, X_{t+dt}^\alpha)$, with $\mathcal{L}^\alpha$ being Ito's generator for the diffusion X_t^α:

$$
\mathcal{L}^\alpha u(t,x) = \mu(t,x,\alpha)\cdot \nabla u(t,x) + \frac{\sigma(t,x,\alpha)\sigma(t,x,\alpha)^T}{2} : \nabla^2 u(t,x)
$$

After simplifcation of u and division by dt, we are left with the **Hamilton-Jacobi-Bellman (HJB) PDE**:

$$
\partial_t u(t,x) + \underbrace{\max_{\alpha \in A_t}\left[f(t,x,\alpha) + \mathcal{L}^\alpha u(t,x)\right]}_{H(\nabla u, \nabla^2 u)} - ru(t,x) = 0
$$

or in short:

$$
\partial_t u + H(\nabla u, \nabla^2 u) - ru = 0, \quad t < T, x \in \mathbb{R}^d
$$

with the Hamiltonian H given by

$$
H(\nabla u, \nabla^2 u)(t,x) = \max_{\alpha \in A_t}\left[f(t,x,\alpha) + \mu(t,x,\alpha)\cdot \nabla u(t,x) + \frac{\sigma(t,x,\alpha)\sigma(t,x,\alpha)^T}{2} : \nabla^2 u(t,x)\right]
$$

The PDE has to be coupled with the final time condition

$$
u(T,x) = g(x)
$$

and hence it is natural to solve it backwards in time.

An optimal control following a **feedback law** is obtained by:

$$
\alpha^*(t,x) = \arg\max_{\alpha \in A_t}\left[f(t,x,\alpha) + \mu(t,x,\alpha)\cdot \nabla u(t,x) + \frac{\sigma(t,x,\alpha)\sigma(t,x,\alpha)^T}{2} : \nabla^2 u(t,x)\right]
$$

In particular, α^* will usually depend on $\nabla u, \nabla^2 u$.

Note that he HJB PDE looks very complicated, and is difficult to solve because of the non-linearity on the first and second derivatives of u.

It does however provide an explicit PDE to be solved (analytically or numerically) which yields the optimal value function u, as well as an explicit optimization problem to get the optimal feedback law α^*.

5 Verification theorem*

In this section, we will show that the value function and the control obtained from the HJB PDE are indeed optimal. Again, we will keep the notations consistent with Sections 2 and 4.

Since we do not (rigorously) know yet that $u(t,x) = \max_{\alpha \in A} V(t,x;\alpha)$ solves the HJB PDE, we will use a different letter, v, to refer to the solution of the PDE.

Theorem 5.1. *Let v be a C^2 solution of the HJB PDE:*

$$
\partial_t v(t,x) + H(\nabla v, \nabla^2 v)(t,x) - rv(t,x) = 0, \quad t < T, x \in \mathbb{R}^d
$$

$$
v(T,x) = g(x), \quad x \in \mathbb{R}^d
$$

with

$$H(\nabla v, \nabla^2 v)(t,x) = \max_{\alpha \in A_t} \left[f(t,x,\alpha) + \mu(t,x,\alpha) \cdot \nabla v(t,x) + \frac{\sigma(t,x,\alpha)\sigma(t,x,\alpha)^T}{2} : \nabla^2 v(t,x) \right]$$

Define

$$\alpha^*(t,x) = \arg\max_{\alpha \in A_t} \left[f(t,x,\alpha) + \mu(t,x,\alpha) \cdot \nabla v(t,x) + \frac{\sigma(t,x,\alpha)\sigma(t,x,\alpha)^T}{2} : \nabla^2 v(t,x) \right]$$

Then v is the optimal value function i.e.

$$v(t,x) = u(t,x) = \max_{\alpha \in A} V(t,x;\alpha)$$

and α^* is the optimal control.

Proof. First notice that by definition of α^*, v solves

$$\partial_t v(t,x) + f(t,x,\alpha^*) + \mu(t,x,\alpha^*) \cdot \nabla v(t,x) + \frac{\sigma(t,x,\alpha^*)\sigma(t,x,\alpha^*)^T}{2} : \nabla^2 v(t,x) - rv(t,x) = 0, \quad t < T, x \in \mathbb{R}^d$$

$$v(T,x) = g(x), \quad x \in \mathbb{R}^d$$

and thus by the Feynman-Kac theorem,

$$v(t,x) = \mathbb{E}_{X_t = x} \left[\int_t^T e^{-r(s-t)} f(s, X_s^{\alpha^*}, \alpha_s^*) ds + e^{-r(T-t)} g(X_T^{\alpha^*}) \right] = V(t,x;\alpha^*)$$

Thus by definition of u, one has that

$$v(t,x) \leq u(t,x)$$

Reciprocally, let $\alpha \in A$ be any fixed control in the admissible set.

Construct a process X_t^α using the state equation and the control α defined above, starting from the state (t,x).

Computing $d(e^{-rt} v(t, X_t^\alpha))$ by Ito's lemma yields:

$$\begin{aligned} d(e^{-rt} v(t, X_t^\alpha)) &= e^{-rt}(\partial_t v + \mathcal{L}^\alpha v - rv) dt + (stuff) dW_t \\ &= e^{-rt}(\partial_t v(t, X_t^\alpha) + \mathcal{L}^\alpha v(t, X_t^\alpha) + f(t, X_t^\alpha, \alpha) - rv(t, X_t^\alpha)) dt + (stuff) dW_t - e^{-rt} f(t, X_t^\alpha, \alpha) dt \end{aligned}$$

By definition of H, we have that $\mathcal{L}^\alpha v(t, X_t^\alpha) + f(t, X_t^\alpha, \alpha) \leq \mathcal{L}^{\alpha^*} v(t, X_t^\alpha) + f(t, X_t^\alpha, \alpha^*)$, and thus

$$\begin{aligned} d(e^{-rt} v(t, X_t^\alpha)) &\leq e^{-rt}(\partial_t v(t, X_t^\alpha) + \mathcal{L}^{\alpha^*} v(t, X_t^\alpha) + f(t, X_t^\alpha, \alpha^*) - rv(t, Xt^\alpha)) dt + (stuff) dW_t - e^{-rt} f(t, X_t^\alpha, \alpha) dt \\ &\leq (stuff) dW_t - e^{-rt} f(t, X_t^\alpha, \alpha) dt \end{aligned}$$

since v solves the HJB PDE.

Integrating the above between t and T, taking $\mathbb{E}_{X_t = x}$ and multiplying by e^{-rt}, one gets:

$$\mathbb{E}_{X_t = x} \underbrace{\left[e^{-r(T-t)} v(T, X_T^\alpha) \right]}_{e^{-r(T-t)} g(X_T^\alpha)} - v(t,x) \leq \mathbb{E}_{X_t = x} \left[\int_t^T e^{-r(s-t)} f(s, X_s^\alpha, \alpha_s) ds \right]$$

which is after rearranging terms:

$$V(t,x;\alpha) \leq v(t,x)$$

This being true for any admissible control α, one gets by taking $\max_{\alpha \in A}$:

$$u(t,x) \leq v(t,x)$$

We conclude that

$$u = v$$

and thus the control α^* is an optimal feedback control. ■

6 Stopping time versions

Again using the notations of Section 2, instead of studying an optimal control problem with fixed horizon (finite or infinite) $T \leq +\infty$, one can instead replace it with a stopping time

$$\tau = \min\{\tau', T\}$$

where $\tau' = \inf\{s > 0 | X_s \notin D\}$, for some domain $D \subset \mathbb{R}^d$. A technical assumption needed is that $\mathbb{E}[\tau] < +\infty$, for any control $\alpha \in A$ (τ depends on X_t, which itself depends on the control α from the state equation).

One is thus interested in

$$u(t,x) = \max_{\alpha \in A} \mathbb{E}_{X_t = x} \left[\int_t^\tau e^{-r(s-t)} f(s, X_s^\alpha, \alpha_s) ds + e^{-r(T-\tau)} g(\tau, X_\tau^\alpha) \right]$$

where $g(\tau, x)$ represents multiple different functions; If $\tau = T$, we never exited the domain and thus the function $x \mapsto g(T, x)$ represents the payoff at final time.

If we exited the domain at some $x \in \partial D$ at time τ, $\tau \mapsto g(\tau, x)$ represents the reward for exiting the domain at x.

Again, using the exact same proof as in Section 5, one can show that the optimal value function solves a similar HJB PDE, except with an additional boundary condition:

Theorem 6.1. *Let v be a C^2 solution of the HJB PDE:*

$$\partial_t v(t,x) + H(\nabla v, \nabla^2 v)(t,x) - rv(t,x) = 0, \quad t < T, x \in D$$

$$v(T,x) = g(T,x), \quad x \in D$$

$$v(t,x) = g(t,x), \quad x \in \partial D, t < T$$

with

$$H(\nabla v, \nabla^2 v)(t,x) = \max_{\alpha \in A_t} \left[f(t,x,\alpha) + \mu(t,x,\alpha) \cdot \nabla v(t,x) + \frac{\sigma(t,x,\alpha)\sigma(t,x,\alpha)^T}{2} : \nabla^2 v(t,x) \right]$$

Define

$$\alpha^*(t,x) = \arg\max_{\alpha \in A_t} \left[f(t,x,\alpha) + \mu(t,x,\alpha) \cdot \nabla v(t,x) + \frac{\sigma(t,x,\alpha)\sigma(t,x,\alpha)^T}{2} : \nabla^2 v(t,x) \right]$$

Then v is the optimal value function i.e.

$$v(t,x) = u(t,x)$$

and α^* is the optimal control.

Proof. Redo the exact same computations as in the proof of Theorem 5.1, but integrate from t to τ instead of t to T. The fact that $\mathbb{E}[\tau] < +\infty$ for all controls allows you to get rid of the term $\mathbb{E}[\int_t^\tau (...) dW_s]$. ■

Remark. The PDE still holds if we have a time independent problem; setting

$$u(x) = \max_{\alpha \in A} \mathbb{E}_{X_0 = x} \left[\int_0^\tau e^{-rs} f(X_s^\alpha, \alpha_s) ds + e^{-r\tau} g(X_\tau^\alpha) \right]$$

for $\tau = \inf\{s > 0 | X_s \notin D\}$ ($T = +\infty$).

Then u is also time independent, and is a solution of the PDE:

$$H(\nabla v, \nabla^2 v)(x) - rv(x) = 0, \quad x \in D$$

$$v(x) = g(x), \quad x \in \partial D$$

The proof is entirely anloguous to the one for the previous thereoms, and the fact that u does not depend on time comes from the translation invariance in time of the PDE.

7 Viscosity solution

The theorems in Sections 5, 6 rely on the facts that a C^2 solution to the HJB PDE exists.

This is in general not the case, as it is difficult to control the regularity of the solution because of the non-linear operator H.

Here is an easily solvable problem that yields a non C^2 solution, providing a motivation for introducing viscosity solutions.

7.1 A degenerate example

Consider a domain $D = [0,1]^2$, and a tiny ant robot starting at some point $x \in D$.

You control deterministically the velocity $v_t \in \mathbb{R}^2$ of the ant, though you have some limit on the maximal speed: $|v_t| \leq 1$.

The state equation is thus

$$dX_t^v = v_t dt, \quad \text{or} \quad \frac{dX_t^v}{dt} = v_t$$

with $X_0 = x$.

For each time dt spent inside the square D, one has a reward of $F = -1$ (or a cost of 1). Once the ants reaches the boundary of the square, the game stops and you get a reward of $g(x) = 0$.

The goal is to maximize the quantity

$$u(x) = \max_{v,|v|\leq 1} \left[\int_0^\tau (-1) ds \right] = \max_{v,|v|\leq 1} [-\tau]$$

Note that this is a time independent problem, since all rewards $F = 1, g = 0$, the drift $\mu(t,x,v) = v$ and volatility $\sigma(t,x,v) = 0$ do not explicitely depend on time. Besides, since the problem is deterministic, there is no need to take an expectation.

The goal is very simple: starting from x, the ant has to reach the boundary of the domain as fast as possible, but can't run over the speed limit 1.

What is the HJB PDE? According to Section 6, u solves the PDE:

$$\max_{v,|v|\leq 1} (-1 + v \cdot \nabla u) = 0$$

for $x \in D$, with $u = 0$ on ∂D.

The PDE can be simplified to

$$|\nabla u| = 1, \quad \text{on } D$$

and $u = 0$ on ∂D.

The optimal control is given by a constant speed 1 velocity $v^* = \nabla u / |\nabla u|$ vector, pointing to the closest boundary point.

A trivial solution to the control problem is given by $u(x) = d(x, \partial D)$ which is the distance to the boundary of the domain.

This solution isn't even C^1, as the function has a kink on the diagonals of the square.

Viscosity solutions are then introduced to fix issues of non existence of solutions that are regular enough.

7.2 Viscosity solutions

Note that the notion of vscosity solution introduced in this paragraph is not the original one, nor fully correct. For more details about viscosity solutions for HJB equations, refer to [2] for example.

Instead of considering the solution of the original HJB PDE found in Section 4, one instead solves an approximated problem: set $\epsilon > 0$, and seek the solution u^ϵ of the PDE:

$$\partial_t v + H(\nabla v, \nabla^2 v) - rv + \epsilon \Delta v = 0, \quad t < T, x \in \mathbb{R}^d$$

$$v(T,x) = g(x), \quad x \in \mathbb{R}^d$$

Note the introduction of a Laplacian Δ term. This term has a regularizing effect, and one can show under mild conditions on the sets A_t, F and g that the solution remains at least C^1, and has weak second order derivatives.

The solution u^ϵ yields a control $\alpha^{*\epsilon}$, which might not be the optimal control.

However, if we have that $u^\epsilon \to u$ in some topology, then u is called the **viscosity solution** of the HJB PDE.

An optimal control might still not exist because there is again no guarantee of regularity for u. However, $\alpha^{*\epsilon}$ will yield a sequence of 'near' optimal controls as $\epsilon \to 0$.

8 Solving the optimal investment/consumption problem

This was a long introduction to stochastic optimal control, but doesn't detail much on how to solve the problems.

In the following, we explicitely solve the optimal investment/consumption problem introduced by Merton in 1971, that we detailed in 1.1.

This is one of the few examples where one can explictely write a simple form for Hamiltonian H, solve for the optimal feedback law and the optimal value u.

Most HJB PDEs have to be approximated by discretizing the time, and solving them would require approximate dynamic programming techniques common in the litterature of Reinforcement Learning.

Using the setting of Section 1.1, and the notations of Sections 4 and 5.

Recall the state equation:

$$dX_t = [(\theta_t \mu + (1-\theta_t)r)X_t - c_t]\,dt + \theta_t \sigma X_t dW_t$$

where we chose μ, σ to be constants in this case. Assume $\mu > r$ otherwise it would always be advantageous (in expectation) to invest our money in the risk free bond.

The optimization problem is:

$$u(t_0,x) = \max_{\alpha \in A} \mathbb{E}_{X_{t_0}=x}\left[\int_{t_0}^{T} e^{-r(s-t_0)} c_s^\gamma ds + e^{-r(T-t_0)} \left(X_T^\alpha\right)^\gamma\right]$$

One can directly write the HJB PDE for the optimal value function u;

$$\partial_t u + \max_{\substack{0\le\theta\le 1 \\ c\ge 0}} \left[c^\gamma + [(\theta\mu + (1-\theta)r)x - c]\partial_x u + \frac{\theta^2\sigma^2 x^2}{2}\partial_{xx} u\right] - ru = 0$$

We will not worry too much about the regularity of the solution, since we could always resort to a viscosity solution.

Maximizing over c yields the optimal consumption strategy as a feedback law

$$c^* = \left(\frac{1}{\gamma}\partial_x u\right)^{1/(\gamma-1)}$$

assume for a moment that $\partial_x u > 0$ (the more money we start with, the more satisfaction we should get), which we should check later, so $c^* \ge 0$.

Maximizing over θ gives the optimal portfolio weights as

$$\theta^* = -\frac{(\mu - r)\partial_x u}{x\sigma^2 \partial_{xx} u}$$

assume for a moment that $\partial_{xx} u < 0$ (due to the concavity of the utility function), which we will check later. This gives that $\theta^* \ge 0$. Besides, we assumed that $\theta^* \le 1$, which might not be gauranteed. Let's proceed

with this value, give a sufficient condition for this to hold, and explain what to do more generally. Note that we would not need this upper bound $\theta \leq 1$ if we allowed borrowing money to invest in the stock.

It remains now to solve for u to completely solve the problem. After plugging in the optimal values for the controls, we get that u is the solution of:

$$\partial_t u + k_\gamma (\partial_x u)^{\frac{\gamma}{\gamma-1}} + \frac{1}{2}\frac{\gamma(\mu-r)^2}{(1-\gamma)\sigma^2}\frac{(\partial_x u)^2}{\partial_{xx} u} - ru = 0$$

with $k_\gamma = \frac{1-\gamma}{\gamma^{\gamma/(\gamma-1)}}$.

This is still quite an ugly non-linear PDE... The powers in front of the spatial derivatives of u suggests a power law dependence in its x variable. In particular, one can try a sepearation of variables method to seek a solution of the type:

$$u(t,x) = h(t)x^\gamma$$

This yields after simplifications an ODE for h:

$$h' + (1-\gamma)h^{\gamma/(\gamma-1)} + \left(\frac{1}{2}\gamma\frac{(\mu-r)^2}{(1-\gamma)\sigma^2} - r\right)h = 0$$
$$h(T) = 1$$

for some function h to determine.

This is a classical Bernoulli ODE with constant coefficients, and closed form solutions are well known and not hard to compute; one can check that the solution is given by:

$$h(t) = \left[\frac{1-\gamma}{k}\left(e^{\frac{k(T-t)}{1-\gamma}} - 1\right) + e^{\frac{k(T-t)}{1-\gamma}}\right]^{1-\gamma}$$

for $k = \left(\frac{1}{2}\gamma\frac{(\mu-r)^2}{(1-\gamma)\sigma^2} - r\right)$. Note that the quantity inside the brackets [...] is non-negative regardless of the sign of k, hence h is well defined and nonnegative.

Thus one gets the optimal value function, and more importantly the optimal controls:

$$c^*(t,x) = \frac{x}{\frac{1-\gamma}{k}\left(e^{\frac{k(T-t)}{1-\gamma}} - 1\right) + e^{\frac{k(T-t)}{1-\gamma}}}$$

which is indeed non-negative. The optimal consumption at the state (t,x) is proportional to the amount of money x one currently holds. The optimal weight is given by:

$$\theta^*(t,x) = \frac{\mu - r}{\sigma^2(1-\gamma)}$$

$\theta^* \geq 0$, but in order to guarantee that $\theta^* \leq 1$, one should assume that $\frac{\mu-r}{\sigma^2(1-\gamma)} \leq 1$. In other words, one has to assume that the *Sharpe ratio* normalized by our utility is not too large, otherwise the risky asset is a much better bet on average than investing in the riskless asset.

The optimal strategy assumes that we hold in our portfolio a constant proportion of stock and bonds. This is not surprising since the dynamics of the stock or bond do not change over time (μ, r, σ constants), and there is no need to readjust. Note that the formula above would sitll hold if one replaced μ, r, σ by *deterministic* functions of time.

What if $\frac{\mu-r}{\sigma^2(1-\gamma)} > 1$? Then the optimal weight while optimizing our Hamiltonian is given by:

$$\theta^* = \min\left(-\frac{(\mu-r)\partial_x u}{x\sigma^2\partial_{xx} u}, 1\right)$$

instead of $-\frac{(\mu-r)\partial_x u}{x\sigma^2\partial_{xx} u}$. This complicates the HJB PDE, and would require solving for u again.

Note that if one allowed borrowing cash to buy more stock ($\theta \geq 0$), one could do the exact same computations as above without worrying about the upper bound for θ^*.

References

[1] Deterministic and Stochastic Optimal Control, W. Fleming, R. Rishel, Springer

[2] Stochastic Optimization in Continuous Time, F-R. Chang, Cambridge Univ Press

[3] PDE for Finance (Lecture 5), R.V. Kohn, https://www.math.nyu.edu/faculty/kohn/pde.finance/2015/section5.pdf

[4] Arbitrage Theory in Continuous Time Finance (Chapter 19), T. Björk, Oxford Finance

[5] Stochastic Control Theory & Automated Market Making, D. Borden, slides found at http://www.columbia.edu/cu/cap/pdf-files/Borden_D_CAP_2010.pdf

[6] High-frequency trading in a limit order book, M. Avellaneda, S. Stoikov, Quantitative Finance, Vol. 8, No. 3, April 2008

[7] Reinforcement learning: An Introduction, S. Sutton, D. Barto, available at http://incompleteideas.net/book/the-book-2nd.html