AI and Games

Todd Davies

March 22, 2016

Overview

The aim of this semester is to understand the following topics.

What are Stackleberg (Leader-Follower) games? What constitutes a solution to a Stackleberg game? How can learning and optimisation be used to learn good solutions? Applications to price setting and marketing will be discussed.

What is reinforcement learning and how is it applied to on-line learning? What are the important mechanisms of on-line learning? How can reinforcement learning be applied to games situations?

Attribution

These notes are based off of both the course notes (¡course notes¿). Thanks to ¡names¿ for such a good course! If you find any errors, then I'd love to hear about them!

Contribution

Pull requests are very welcome: `https://github.com/Todd-Davies/third-year-notes`

Contents

In Semester One, of this course we covered zero and non-zero sum games (Nash games) and search algorithms to find equlibriums (minimax and alpha beta pruning). In this semester, we are no longer assuming that each player has the same 'role' and that the games are zero-sum. We are letting there be an infinite number of positions for each player to take and not assuming that players have perfect information.

We're going to be looking at Stackelberg games; an example of which could be retail pricing. Different firms might offer different prices on products, they can choose a price from an infinate set (the real numbers) and some firms might have private information that others might not[1].

An industry can be controlled or dominated by a one, two or multiple entities:

Monopoly: When an industry consists of a single firm.

Duopoly: When an industry consists of two firms.

Oligopoly: When an industry consists of a few firms (and when each decision one takes impacts on the other's profits).

In duopoly/oligopoly situations, the roles of the players are different. If one firm chooses to change its prices first, then the others have to decide what to do in light of the change; and may not know the full economic and/or political motivations behind the initial move.

In a two player Stackelberg game, one player selects his strategy first, and then the other responds. The first player is called the leader (P_L) and the second is called the follower (P_F). This contrasts with Nash games, where all of the players select their strategies simultaneously.

Both the leader and follower have payoff functions[2], which they want to maximise:

$$J_L(U_L, U_F)$$
$$J_F(U_L, U_F)$$

A strategy space is the set of all possible strategies for a single player. The leader's strategy space is U_L and the follower's is U_F. They can include finite (discrete) or infinite (continuous) strategies. To play the game, the leader must first choose some strategy $u_L \in U_L$ and then the follower must choose some strategy $u_F \in U_F$.

In a Stackelberg game, when the leader has announced u_L, the follower selects a strategy using their reaction function $R(u_L) \in U_F$ such that:

$$J_F(u_L, R(u_L)) = MAX_{u_F \in U_F} J_F(u_L, u_F)$$

This looks complicated, but what it says is that the payoff for the follower is given the parameters u_L and u_F. The follower obviously wants to maximise their payoff, and so tries to find the strategy $u_F \in U_F$ that will maximise their payoff. This isn't always easy, since enumerating every possible strategy and knowing what will happen if any given strategy is deployed is sometimes impossible.

In order to solve a Stackelberg game, we need to solve the following problem:

- What is the follower's reaction function $R(u_L)$?
- When we've found that, what leader strategy u_L maximises the leader's payoff function:

$$J_L(u_L^*, R(u_L^*)) = MAX_{u_L \in U_L} J_L(u_L, u_F)$$

If the follower has a reaction function $R(u_L)$ and there exists a leader strategy $u_L^* \in U_L$ and the response strategy is $u_F^* = R(U_L^*) \in U_F$ such that $J_L(u_L^*, u_F^*) = MAX_{u_L \in U_L} J_L(u_L, R(u_L))$, then (u_L^*, u_F^*) is called a **Stackelberg strategy/equilibrium**; i.e. both sides are paying an optimal strategy.

[1] Typical examples of unknown information include payoff functions or unit cost.

[2] A payoff function indicates the benefit for that player given the chosen strategies of all the players.

We always assume that the follower is rational and tries to find the best reaction strategy. Sometimes this is untrue, for example if taking a non-optimal strategy results in a rival player having a big loss.

For a player to be a leader, he needs to act first. There is no requirement that they are a leader in an economic or political sense. The only requirement to be a follower is to react to the leader's strategy[3]. As mentioned before, Stackelberg games can be continuous or discrete, depending on if the strategy space is finite or infinite.

Since the only difference between a Nash game and a Stackelberg game is whether moves are played simultaneously or in order, should we prefer Nash or Stackelberg games? If the player has the opportunity to be the leader, then Stackelberg games should be preferred, since he will always be better off than in a Nash game in this instance. Here's a mini-proof:

Let u_1 and u_2 be the strategies for player's one and two, and let $J_1(u_1, u_2)$, $J_2(u_1, u_2)$ be their respective payoff functions. If there is a Stackelberg strategy and a Nash strategy, then:

$$J_1(u_1^{Stackelberg}, u_2^{Stackelberg}) \geq J_1(u_1^{Nash}, u_2^{Nash})$$

Note that sometimes the follower can be better off in a Stackelberg game, but the leader would still be better off playing the Stackelberg game than playing a Nash game. Other times, the player could win as the leader or follower, but win better as the follower.

For the follower to play the game, he needs to know nothing about the leader or his strategy space. However, for the leader to play, he needs to know the follower's payoff function and strategy space in order to work out the action that will give the max payoff. If such information is not available, then the leader must guess or learn it (based off previous experience or data).

1 Solving Stackelberg game problems

We can solve with two sequential maximisation problems, each for a single player:

1. For each of the leader's strategies $u_L \in U_L$, solve:

   $$max_{u_F \in U_F}(J_F(u_L, u_F)$$

   The solution for this is the follower's reaction function $R(u_L)$.

2. Find the best strategy for the leader by solving:

   $$max_{u_L \in U_L}(J_L[u_L, R(u_L)])$$

   If u_L^* is the strategy found in the above maximisation, and $u_L^* = R(u_L^*)$, then (u_L^*, u_F^*) is a Stackelberg strategy.

So, in order to find the Stackelberg strategy, we need to solve two maximisation problems. This is A-level maths, but we can recap it in the next bit.

1.1 How to find the maximum point of a function over a continuous space

We want to find a point $x^* \in X$ that maximises $f(x)$ over X, such that $\forall x \in X, f(x^*) \geq f(x)$. Such a point is called the global maximum point of $f(x)$ on X. A local maximum point y^* is the maximum value of $f(x)$ within the region U, such that $\forall x \in U, f(y^*) \geq f(x)$.

The standard method for trying to find a maximum of a function is by using derivatives. A derivative is usually notated by a dash after the function. Finding the derivative of a derivative is called the second order derivative function, and has two dashes.

[3]Note that this does not mean the follower is in a weaker position.

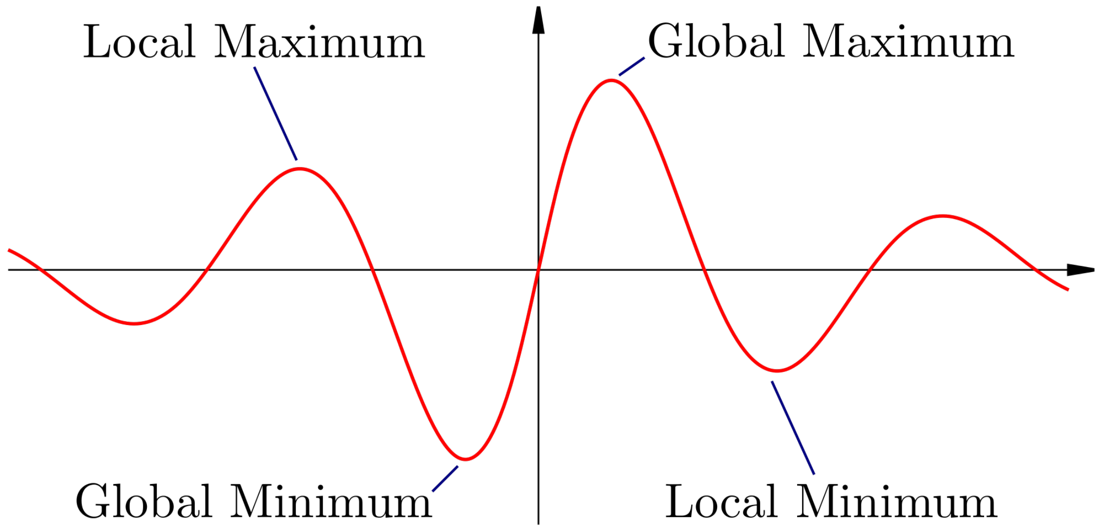

Figure 1: Wikipedia's pictorial explanation of minima and maxima.

A first order derivative represents the gradient of the line drawn by the function at a specific point. If $f'(x) > 0$, then the line slopes upwards, and if $f'(x) < 0$ then it's sloping down. Obviously if it is equal to zero, then the line is horizontal.

1.1.1 Calculating a derivative

The most simple rule is:

$$f(x) = x^n \therefore f'(x) = nx^{n-1}$$

This means if the function is constant, it disappears:

$$f(x) = C \therefore f'(x) = 0$$

Here are three simple examples:

$$f(x) = x^2 \therefore f'(x) = 2x$$

$$f(x) = x^3 \therefore f'(x) = 3x^2$$

If a function is composed of other functions that are added together (of the form $f(x) = f_1(x) + \cdots + f_n(x)$ then:

$$f'(x) = f_1'(x) + \cdots + f_n'(x)$$

Just like:

$$f(x) = -2x^2 + 4x + 5 \therefore f'(x) = -4x + 4$$

Finally, if it's composed of functions that are multiplied together, then $f'(x) = f_1'(x)f_2(x) + f_1(x)f_2'(x)$:

$$f(x) = x^2 \times x \therefore f'(x) = [x^2]' \times x + x^2 \times [x]' = 2x \times x + x^2 \times 1 = 2x^2 + x^2 = 3x^2$$

1.1.2 Finding maxima

Once you've found the derivative of your function, and then found the points where the derivatives are 0 (and the gradient of the line is therefore zero), you need to determine whether that point is a minima or maxima. To do this, we differentiate again to get the second order derivative. If the value of $f''(x) \geq 0$ then the gradient is increasing, therefore it's a minima. Because of this, we want points where the second order derivative gives a negative value, indicating that the point is a maxima.

To find the largest maxima, we simply find the one that is largest value of $f(x)$.

2 Solving Stackelberg game problems

We know that there are two steps to solve a Stackleberg game; first of all, you must solve the maximisation problem $\max u_F \in U_F J_F(u_L, u_F)$, giving the reaction function $R(u_L)$. Then you must find the best leader strategy by solving $\max_{u_L \in U_L} J_L(u_L, R(u_L))$ such that (u_L, u_F) is the Stackelberg strategy.

The strategy space (i.e. the different prices that the leader and follower can set) is $U_L = [c_L, +\infty], U_F = [c_F, +\infty]$, where c_L, c_F are the cost of each unit for the leader and follower respectively.

The payoff function (i.e. how much profit the leader and follower make) is defined by:

$$J_L(u_L, u_F) = (u_L - c_L) \times S_L(u_L, u_F)$$

$$J_F(u_L, u_F) = (u_F - c_F) \times S_F(u_L, u_F)$$

This is essentially the profit per unit (sale price minus cost) multiplied by the number of units sold in the sale. The sale function is given to you in exam questions and is a quadratic function over u_L and u_F.

We want to find the values of u_L and u_F. To do this, we first differentiate J_F., then we differentiate J_L and sub in the value for u_F we found in the first stage so we can get answers. Here is an example:

$$J_L(u_L, u_F) = (u_L - c_L) \times S_L(u_L, u_F)$$

$$J_F(u_L, u_F) = (u_F - c_F) \times S_F(u_L, u_F)$$

$$c_F = 1 = c_F$$

$$S_L = 5 - 2u_L + u_F$$

$$S_F = 6 + u_L - 2u_F$$

To get the reaction function, we differentiate J_F:

$$
\begin{aligned}
0 &= \frac{d}{du_F} J_F(u_L, u_F) \\
&= \frac{d}{du_F}(u_F - 1)(6 + u_L - 2u_F) \\
&= \frac{d}{du_F} 6u_F + u_F u_L - 2u_F^2 - 6 - u_L + 2u_F \\
&= u_L - 4u_F + 8
\end{aligned}
$$

If we differentiate again, we see that this is a maxima $(-4 > 0)$.

We can rearrange it so that: $R(u_L) = \frac{u_L}{4} + 2$.

Now we can sub in the value for u_F into J_L:

$$
\begin{aligned}
J_L(u_L, R(u_L)) &= (u_L - 1)(502u_L + R(u_L)) \\
&= \frac{1}{4}(35u_L - 7u_L^2 - 28)
\end{aligned}
$$

And differentiate:

$$
\frac{d}{du_L} J_L(u_L, R(u_L)) = \frac{1}{4}(35 - 14u_L) = 0
$$

If we rearrange that, we get $u_L = \frac{35}{14}$, therefore $u_F = 2.625$.

This means the profits are:

$$
J_L(\frac{35}{14}, 2.625) = (\frac{35}{14} - 1)(5 - 2\frac{35}{14} + 2.625) = 3.9375
$$

$$
J_F(\frac{35}{14}, 2.625) = (2.625 - 1)(6 + \frac{35}{14} - 2(2.625)) = 5.28125
$$

3 Stackelberg games with imperfect information

It is very rare that both players in a two person Stackelberg game will have perfect information; in any case, it is in their interest to hide information from one-another. In a game of imperfect information, the follower is not affected; their job is the same (respond to whatever the leader does).

However, this is a huge change for the leader. Without knowing the strategy space for the follower, or the follower's reaction function, they can't work out a good strategy. We need to come up with a technique to let the leader 'guess' a good strategy based of what we *do* know about the game.

There are two obvious solutions, and one machine learning solution to this. We're interested in the ML solution, but the obvious solutions are:

- Choose to be the follower. Then we don't have to decide on the first strategy, and can just choose a reaction. However, this isn't always possible (sometimes you're forced to move first), and in some games, the follower may always win.
- Find the best strategy for the worst scenario; that is to say we should find the best, worst strategy:

  $$u_L = argmax_{u_L \in U_L}(min_{u_F \in U_F} J_L(u_L, u_F))$$

  Though this gives a lower bound on the payoff, it is also far too conservative. Imagine if you always planned for the worst and never considered (never mind hoped) for the best!

However, the best solution is to attempt to learn what the follower will do based on what has happened in the past. Many two person Stackelberg games are played repeatedly (e.g. setting oil prices every day), and therefore there is a lot of information available to learn from. All we need to learn is the follower's reaction function; i.e. what value they will choose for u_F given any chosen u_L.

The approach we're going to use to do this is called linear regression. This involves finding a linear function that gives similar values to what the follower has done in the past. You could also use a polynomial function or a neural network (neural networks are called the *universal approximation*, since they can mimic any function).

So, if $y = R(x)$ be the unknown function, and imagine we have a set of T datapoints where we know the input x and the value y. We want to make a new function $\hat{R}(x)$ that will approximate $R(x)$.

We define $\epsilon(x) = R(x) - \hat{R}(x)$, and try to make ϵ as small as possible for all values of x. Unfortunately, since we don't know R, we can't work out ϵ. However, we can work out the value of ϵ for all the datapoints we do have:

$$\sum_{t=1}^{T} \epsilon^2[x(t)] = \sum_{t=1}^{T} \Big(y(t) - \hat{R}(x(t))\Big)^2$$

Trying to make this value as small as possible is our goal. The value is called the least square criterion.

Designing a function is problem dependent, but we know our reaction function will be linear:

$$R(x) = a + bx$$

To make $\hat{R}(x)$ imitate $R(x)$, we must find values for a and b that minimise the squared criterion we had before:

$$min_{a,b} \sum_{t=1}^{T} (y(t) - (a + bx(t))))^2$$

Instead of solving the minimising problem, we can instead solve the following maximisation problem:

$$max_{a,b} - \sum_{t=1}^{T} (y(t) - (a + bx(t))))^2$$

Of course, we can use partial differentiation to do this (just like in `COMP36212`), where you eventually get:

$$a = \frac{\sum_{t=1}^{T} x^2(t) \sum_{t=1}^{T} y(t) - \sum_{t=1}^{T} x(t) \sum_{t=1}^{T} x(t)y(t)}{T \sum_{t=1}^{T} x^2(t) - (\sum_{t}^{T} = 1x(t))^2}$$

$$b = \frac{T \sum_{t=1}^{T} x(t)y(t) - \sum_{t=1}^{T} x(t) \sum_{t=1}^{T} y(t)}{T \sum_{t=1}^{T} x^2(t) - (\sum_{t}^{T} = 1x(t))^2}$$

4 Multivariable regression

Now we know how to solve a linear regression problem with one variable, but often, there isn't just one variable. If there are three fuels in a petrol station (diesel, unleaded, super unleaded), then there will be three variables, and we want to learn strategies for all of them.

Here, the leader & follower strategies will be:

$$u_L = (u_1^L, u_2^L, u_3^L)$$

$$u_R = (u_1^R, u_2^R, u_3^R)$$

The follower's reaction function is:

$$\hat{R}(u_L) = \hat{A} + \hat{B}u_L \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \dots \\ \hat{a}_f \end{bmatrix} + \begin{bmatrix} \hat{b}_{1,1} & \hat{b}_{1,2} & \dots & \hat{b}_{2,n_L} \\ \hat{b}_{2,2} & \hat{b}_{2,2} & \dots & \hat{b}_{2,n_L} \\ \dots & \dots & \dots & \dots \\ \hat{b}_{n_f,1} & \hat{b}_{n_f,2} & \dots & \hat{b}_{n_f,n_L} \end{bmatrix} \begin{bmatrix} \hat{u}_1^L \\ \hat{u}_2^L \\ \dots \\ \hat{u}_{n_L}^L \end{bmatrix}$$

Given the leader's strategy (u_L), the follower will react using:

$$\hat{R}_i(u_L) = \hat{a}_i + \sum_{j=1}^{n_L} \hat{b}_{i,j} u_j^L$$

In essence, we're trying to find the mathematical relationship between an output variable and several input variables. All of them take continuous values (if the dependent is discrete, then it's a classification problem). Such problems are widely used in prediction and forecasting.

The idea is that we want to find values for θ that we can sub into $\hat{R}$, such that:

$$\begin{aligned} \theta^* &= arg\ min_\theta \sum_{t=1}^{T} \epsilon^2(t) \\ &= arg\ min_\theta \sum_{t=1}^{T} (R[X(t)] - \hat{R}[X(t), \theta])^2 \\ &= arg\ min_\theta \sum_{t=1}^{T} (y(t) - \hat{R}[X(t), \theta])^2 \end{aligned}$$

Since $\hat{R}(X, \theta) = (1, x_1 \dots, x_n) \begin{pmatrix} a_0 \\ a_1 \\ \dots \\ a_n \end{pmatrix} = a_0 + a_1 x_1 + \dots + a_n x_n$, we just need to find a solution to:

$$\theta^* = arg\ min_\theta \sum_{t=1}^{T} (y(t) - [a_0 + a_1 x_1(t) + \dots + a_n x_n(t)])^2$$

Where $\theta = (a_0, a_1, \dots, a_n)$.

5 Learining in games

There are two types of Stackelberg strategy generators; online and offline ones. In offline learning, the reaction function is regarded as a linear regression problem, and is solved by applying the least square method to the historical data. All the learning is done before the playing starts.

In online learining, we aim to carry on learning even while the game is being played. The reaction strategy from the follower is updated as we play the game. We need to work out how to incorperate the follower's moves into our strategy.

The reason why online methods are preferred, is that the environment in which the game is played is always changing; costs may change, share prices may change etc. The follower's reaction function will change over time to reflect this.

There are two commonly used methods:

Moving Window Approach:
Here, we only consider n historical points, and as new data comes in, we discard old data. To make it better, we can add a parameter λ to make old datapoints mean less than new ones. The impact of the nth most recent datapoint is $n \times \lambda^{n-1}$, and $\lambda \approx 0.95 - 0.99$.

Recursive Least Square Approach: