3.1 - Nature of the System

Here, I make several abstractions to the structure of the environment to make the problem tractable as a game. First, reduce the scale of the environment to a 4 x 4 grid, with a fixed A location at 1, 1 and a fixed B location at 4,4. Agents can start randomly at any location on the grid other than the location of A and B. The grid thus is as follows initially:

A			
			B

Then consider two players, P1 and P2, who control type 1 and type 2 agents respectively. P1 receives utility of 50 from successfully collecting the package from A, and 50 from handing it over to P2. P2 receives utility 50 from successfully collecting the package from P1, and a further 50 for delivering it to B, at which point the game ends. These values have been selected rather than the 500 and 1000 values used in the DQN implementation to highlight an important effect which occurs at lower payoff thresholds (see section 3.4). P1 and P2 have reduced utility equal to the square of the total Manhattan distance their agents move over the course of the game. Trying to solve this spatially is unwieldy, so we shall reduce the scope of the game.

3.2 - Simplifying the System

Let us make some important observations about the nature of the grid. First, players will always move the agent *closest* to A to collect the package and then deliver it, to minimise the cost of movements. Thus, we are only interested in tracking one agent per player. Second, observe that, from A to B, we can model every square in terms of its distance from B (or from A) as follows:

A 6	5	4	3
5	4	3	2
4	3	2	1
3	2	1	**B 0**

So 1,4 and 4,1 are effectively identical in terms of player choices. As a result, further discussion will refer to those squares as *Type 3* squares, squares such as 3,4 and 4,3 as *Type 1* squares, and so forth. Further note the following:

a) P1 will receive 50 utility from sending any agent to A, so their best strategy will always involve sending an agent to A to collect the package
b) P2 will receive utility from both sending an agent to collect the package from P1, and deliver it to B, so their best strategy must involve a successful pickup and delivery.

As a result, the main strategic choices from both P1 and P2 pertain to *where* on the board that package handoff (ie. exchange of parcel between P1 and P2) should happen - as a result, rather than being a more convoluted spatial movement game, we can model this scenario as a game of cost-sharing, where P1 and P2 must collectively settle on how the costs of package delivery should be split between their respective agent. Now P1 and P2 will both move the agents closest to their destinations. Each will be able to choose d_1, d_2 from the set {0, 1, 2, 3, 4, 5, 6}, representing the type of square they will choose to meet at.

3.3 - Meet-in-the-Middle Game

We can now produce functions to model the utility gained from both P1 and P2. In the case of a successful handoff for P1:

$U_1 = 100 - (s + 6 - d_1)^2$

where s is the Manhattan distance moved by P1's from the start to A, and d_1, the type of square at which P1 meets. Note that P1 prefers to move to Type 6 squares the most (representing not needing to move after collecting the parcel at A). Then for P2:

$U_2 = 100 - (t + d_2)^2$

where t is the distance from P2's agent to the location where it meets P1, while they will then travel d_2 to the dropoff point. Note formally that the payoff of 100 is only obtained when $d_1 = 6 - d_2$, a successful dropoff. If there is no successful handoff, U_1 instead = $50 - (s + 6 - d_1)^2$, while $U_2 = - (t + d_2)^2$, a grim result indeed!

Although the values of *s* and *t* are generated via random processes, we can treat them using the expectation of a random variable with relatively consistent behaviour by calculating the expected distance between the *closest* agent and A for *s*. This will simply be the mean minimum distance of the two agents, which can be calculated by taking the mean of minima of each possible combination of starting locations, and we find that s = 2.3.

The expected value of *t* is variable, depending on the actual target location, but has only four distinct values, representing handoffs at A or B, at Type 1 or 5 squares, Type 2 or 4 squares, and Type 3 squares. At A or B, *t* = 2.3, by the same logic as *s*. Similarly, at Type 1 or Type 5, *t* =

1.3, at Type 2 or 4 $t = 0.77$, and for Type 3 $t = 0.6$. These values allow us to compute a payoff matrix which has the following characteristics:

Meet-in-the-Middle Game		*P2*						
		$d_2 = 0$	$d_2 = 1$	$d_2 = 2$	$d_2 = 3$	$d_2 = 4$	$d_2 = 5$	$d_2 = 6$
P1	$d_1 = 0$	31, 95	-19, -5	-19, -8	-19, -13	-19, -23	-19, -40	-19, -69
	$d_1 = 1$	-3, -5	**47, 95**	-3, -8	-3, -13	-3, -23	-3, -40	-3, -69
	$d_1 = 2$	10, -5	10, -5	**60, 92**	10, -13	10, -23	10, -40	10, -69
	$d_1 = 3$	22, -5	22, -5	22, -8	**72, 87**	22, -23	22, -40	22, -69
	$d_1 = 4$	32, -5	32, -5	32, -8	32, -13	**82, 77**	32, -40	32, -69
	$d_1 = 5$	39, -5	39, -5	39, -8	39, -13	39, -23	**89, 60**	39, -69
	$d_1 = 6$	45, -5	45, -5	45, -8	45, -13	45, -23	45, -40	**95, 31**

3.4 - Solutions

Nash Equilibria are indicated in bold. Note that, with the exception of $\{d_1 = 0, d_2 = 0\}$ (which represents the players meeting at B), all meeting squares are Nash Equilibria. This is because there is no incentive for players to deviate from an existing meeting, as they would lose the substantial benefit from a successful handoff. The only exception here arises due to the squared cost of movement; in the expectation, P1 would rather remain at $d_1 = 0$ and miss the handoff, than continuing all the way to B and incurring a hefty movement penalty. This illustrates an important principle for the Deep Learning practitioner interested in building a DQN to solve this problem - rewards in the DQN must be set suitably high to properly encourage collaborative behaviour.

There are two socially optimal solutions; meeting at a Type 3 square or a Type 4 square. The represents a true meeting in the middle, while the second represents a meeting slightly closer to A than to B. Both of these require that P1 and P2 share the cost as efficiently as possible between their agents, even though they have opposed preferences as to how that cost be shared. In this respect, the game is more alike the Battle of the Sexes than it is unalike. In general, we should expect agent behaviour to converge on any of these Nash Equilibria.