Optiver - Quantitative Researcher

Paolo Prenassi

January 2022

1 Cover Letter

My professional path is not a traditional one. Pushed by a strong curiosity and a desire for new challenges, I had the opportunity to experiment both academic and company environments, acquiring diverse mindsets, and becoming aware of how research is done in different fields.

I have always been proud of my competitive attitude which led me to several achievements in the field of Mathematics (winning a national silver medal at the Mathematical Olympiads) and during my university path (obtaining one of the most prestigious Italian scholarships for university students at SGSS). This desire to get involved gave me the opportunity to really get in touch with other top students, broadening my vision with enriching experiences and relationships, and feeling part of a stimulating group with the desire to do breakthrough research.

Attracted by the recent researches and advancements in Machine Learning, I moved from a Bachelor in Physics to a Master in Data Science, where I strengthened my Statistics and Computer Science knowledge while building the foundations of my future ML expertise. During my Master's thesis I developed a Recommender System for an Italian e-Learning platform. Starting from the analysis of the educational contents and online data obtained through web scraping, I was able to forecast the user's knowledge and behavior and recommend the best personalized learning contents. Whereas, after my graduation, I worked as an associate intern at J.P. Morgan in an AI team. In both these experiences I learned how to produce cleaner and more maintainable code for development and production, implementing first-hand the notions learned during my Master.

I decided that 2021 would be a sabbatical year for me and I spent it enhancing my knowledge about financial markets as well as machine learning cutting-edge techniques. I do believe this knowledge will be very helpful to reach my long term goals. I also improved my professionally-aimed programming skills through online courses and practice, and I realized how passionate I am about dealing with real world problems rather than pursuing an academic career.

My background and skill sets have prepared me to be a successful contributing team member in this type of environment and I do believe my detail-oriented thinking and flexible approach to problem solving would be an excellent fit for your organization. I'm eager to apply what I've learned in this new role at Optiver. Thank you for your consideration.

Yours sincerely,
Paolo Prenassi

2 Puzzle

2.1 Results

I found are the following results:

Q1 $\frac{9}{2} = 4.5$

Q2 $+\infty$

Q3 14

In the following sections I will describe the solutions I found that lead me to these results.

2.2 General Considerations

Since the problem is invariant to dilations, we can rescale our measures by a factor 10. All the possible ants positions could now be represented by the nodes with integer coordinates in an euclidian plane.

2.3 Question 1

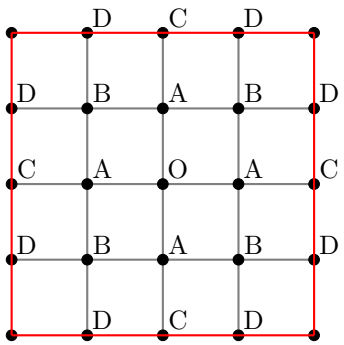

Figure 1: Grid containing all the possible ants positions before reaching the food for the first time. The points represent the nodes with integer coordinates in the cartesian plane (for example $O = (0, 0)$). Equal letters represent nodes with equal probability (at any time) of finding an ant on them.

Let's assume that the ants initial position is $(0, 0)$ and they move following the axis directions. The food is located on a square whose edges are 2 units away from the origin. Then, all the possible ants positions before reaching the food belongs to the following sets (Figure 1):

- $O = \{(0, 0)\}$
- $A = \{(1, 0), (0, 1), (-1, 0), (0, -1)\}$
- $B = \{(1, 1), (-1, 1), (-1, -1), (1, -1)\}$

All the possible positions they can reach the food for the first time belogs to the following sets (Figure 1):

- $C = \{(2, 0), (0, 2), (-2, 0), (0, -2)\}$
- $D = \{(2, 1), (1, 2), (-2, 1), (1, -2), (2, -1), (-1, 2), (-2, -1), (-1, -2)\}$

Let X_t be the position of an ant at time t. Then, the expected average time to reach the food starting from the node $X_{t_0} = x$ is:

$$k_x = \begin{cases} \mathbb{E}_x \left[T \text{ s.t. } X_T \in (C \cup D), X_t \in (O \cup A \cup B) \text{ for all } t_0 \leq t < t_0 + T \,|\, X_{t_0} = x \right], & \text{if } x \in (O \cup A \cup B) \\ 0, & \text{if } x \in (C \cup D) \end{cases} \tag{1}$$

Due to the symmetry of the problem we have:

$$k_x = \begin{cases} k_O, & \text{if } x \in O \\ k_A, & \text{if } x \in A \\ k_B, & \text{if } x \in B \\ 0, & \text{if } x \in (C \cup D) \end{cases} \tag{2}$$

for certain values k_O, k_A, k_B.
The ants move:

- from points in O to points in A with probability 1
- from points in A to points in B with probability $\frac{1}{2}$
- from points in A to points in O with probability $\frac{1}{4}$
- from points in A to points in C with probability $\frac{1}{4}$
- from points in B to points in A with probability $\frac{1}{2}$
- from points in B to points in D with probability $\frac{1}{2}$

Thus, the following system of equations holds:

$$\begin{cases} k_O = 1 + k_A \\ k_A = 1 + \frac{1}{4}k_O + \frac{1}{2}k_B \\ k_B = 1 + \frac{1}{1}k_A \end{cases} \tag{3}$$

whose solution is

$$k_O = \frac{9}{2}, k_A = \frac{7}{2}, k_B = \frac{11}{4}$$

.

Then, **the average time for the ant to get food starting from $X_0 = (0,0)$ is $\frac{9}{2}$.**

2.4 Question 2

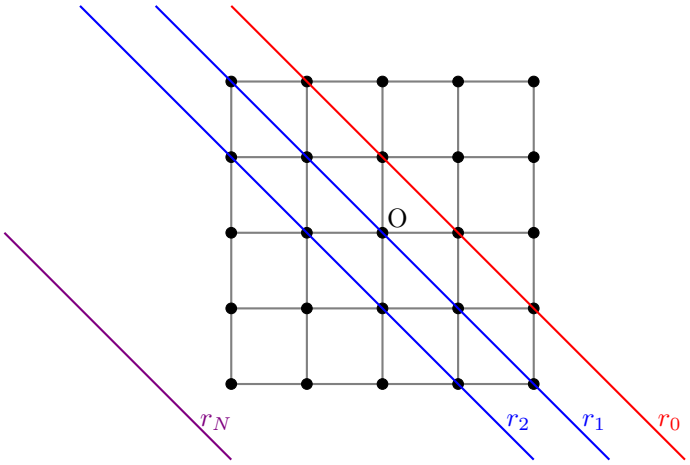

Figure 2: Grid containing possible ants positions before reaching the food for the first time. The points represent the nodes with integer coordinates in the cartesian plane (for example $O = (0,0)$). Points on the same lines r_i represent nodes with equal probability (at any time) of finding an ant on them.

The food is located on the line passing through $(1,0)$ and $(0,1)$.
Let's consider the lines r_i (Figure 2) passing through $(-i+1, 0)$ and $(0, -i+1)$ (in particular the line r_1 has slope -1 as the others). In this setup:

- the food line in represented by r_0
- all the lines are parallel. Then, the distance (in terms of ants steps) between a generic point $P_i \in r_i$ and any line r_j is a constant that depends only on i and j. In particular, $\text{dist}(P_i, r_j) = |i - j|$.
- an ant on the line r_i has probability $\frac{1}{2}$ of going to the line r_{i+1} and $\frac{1}{2}$ of going to the line r_{i-1}

Let X_t be the position of an ant at time t. Then, the expected average time to reach the food starting from the node $X_{t_0} = x$ is:

$$k_x = \begin{cases} \mathbb{E}_x \left[T \text{ s.t. } X_T \in r_0, X_t \in \bigcup_{i=1}^{\infty} r_i \text{ for all } t_0 \le t < t_0 + T \,|\, X_{t_0} = x \right], & \text{if } x \in \bigcup_{i=1}^{\infty} r_i \\ 0, & \text{if } x \in r_0 \end{cases} \tag{4}$$

Due to the symmetry of the problem we have:

$$k_x = \begin{cases} k_i, & \text{if } x \in r_i \quad \text{with } i \geq 1 \\ 0, & \text{if } x \in r_0 \end{cases} \tag{5}$$

for certain values k_i.
Using the previous considerations, we have:

$$\begin{cases} k_n = \frac{1}{2}(1 + k_{n-1}) + \frac{1}{2}(1 + k_{n+1}), & \text{if } n \geq 1 \\ k_0 = 0 \end{cases} \tag{6}$$

This is a recursive equation whose characteristic homogenous equation $\gamma^2 - 2\gamma - 1 = 0$ has root 1. The homogenous has linear solutions in the form $k_n = A + Bn$. One particular solution is $-n^2$. Then, we can write

$$\begin{cases} k_n = A + Bn - n^2, & \text{if } n \geq 1 \\ k_0 = 0 \end{cases} \tag{7}$$

or

$$\begin{cases} k_n = Bn - n^2, & \text{if } n \geq 1 \\ k_0 = 0 \end{cases} \tag{8}$$

Assuming we have another condition $k_N = \varepsilon$ for a certain $N \geq 1$, we can estiamate the B variable:

$$B = N + \frac{\varepsilon}{N}$$

This leads to the following formula:

$$\begin{cases} k_n = \left(N + \frac{\varepsilon}{N}\right) n - n^2, & \text{if } n \geq 1 \\ k_N = \varepsilon, & N \geq 1 \\ k_0 = 0 \end{cases} \tag{9}$$

The solution to the problem should be in the form of

$$k_1 = N + \frac{\varepsilon}{N} - 1 \qquad \text{with } N \geq 1 \tag{10}$$

With the information provided we cannot know or estiamte both N and ε.

However, we can re-elaborate the question and suppose that the food is located not only on r_0 but also on a certain line r_N (with everything else still the same). This implies $\varepsilon = 0$ and $k_1 = N - 1$. What we have created represent a different problem with a different solution. Nevertheless, if we consider the case $N \to \infty$, we obtain a scenario where an ant starting from r_1 has an infinite distance between its position and the food on r_N. It is as if the food on r_N does not even exist.

In this situation the solution of the new question converges to the solution of the original question. Then, the answer to the original problem correspond to the limit

$$k_1 = \lim_{N \to \infty} N - 1$$

Finally, **the average time for the ant to get food starting from O is $+\infty$.**

2.5 Question 3

2.5.1 Exact Solution

Closed boundaries imply a finite number N of available nodes before reaching the food. Each node is defined by its coordinates, but for simplicity we can assign a different integer label to each point from 1 to N.

Let's define I as the set of all those points inside the boundaries and let $(C_{i,j})_{1,..N}$ be the matrix describing the neigbours relationship among nodes inside I. In particular:

$$C_{i,j} = \begin{cases} 1, & \text{if the } i\text{-th node is conncected to the } j\text{-th one and } i, j \in I \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

A node is connceted to another if an ant can move from one to the other in one move.

According to these definitions and following the same reasoning of Eq. 3, we have:

$$k_n = \begin{cases} 1 + \sum_{i=1}^{N} \frac{1}{4} C_{i,n} k_i, & node_n \in I \\ 0, & node_n \notin I \end{cases} \tag{12}$$

or, considerig only the nodes in I:

$$k = \frac{1}{4} C k + \vec{1}_N \tag{13}$$

that leads to our final equation:

$$k = \left(\mathbb{1}_N - \frac{1}{4} C \right)^{-1} \vec{1}_N \tag{14}$$

By contraddiction, if $\left(\mathbb{1}_N - \frac{1}{4}C\right)$ is not invertible, then there exists $x \neq 0$ s.t. $\left(\mathbb{1}_N - \frac{1}{4}C\right) x = 0$, or $\left(\mathbb{1}_N - \frac{1}{4}C\right) \frac{x}{\|x\|} = 0$, that is equivalent to $x = \frac{1}{4}Cx$ with $\|x\| = 1$. This means that there exists a situtation (i.e. a probability vector x representing the probabilities over the nodes) where applying the transition matrix $\frac{1}{4}C$ dosen't change the probability distribution over the nodes. Let J be the set of nodes where $x_{j \in J} \neq 0$.

- If $J = I$, there are nodes connected to the food s.t. $x_j \neq 0$. Moreover,

$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{1}{4} C_{i,j} x_j = \sum_{\substack{j \text{ connected} \\ \text{to food}}} \left(\sum_{i=1}^{N} \frac{1}{4} C_{i,j} \right) x_j + \sum_{\substack{j \text{ not connected} \\ \text{to food}}} \left(\sum_{i=1}^{N} \frac{1}{4} C_{i,j} \right) x_j \tag{15}$$

  where

$$\begin{cases} \sum_{i=1}^{N} \frac{1}{4} C_{i,j} < 1, & j \text{ connected to food} \\ \sum_{i=1}^{N} \frac{1}{4} C_{i,j} = 1, & j \text{ not connected to food} \end{cases} \tag{16}$$

  Then

$$\sum_{i=1}^{N} x_i < \sum_{\substack{j \text{ connected} \\ \text{to food}}} x_j + \sum_{\substack{j \text{ not connected} \\ \text{to food}}} x_j = \sum_{i=1}^{N} x_i \tag{17}$$

  Absurd!

- If $J \neq I$, there exist 2 connected nodes i, j s.t. $x_i = 0$ and $x_j \neq 0$. Then

$$0 = x_i = \sum_{j=1}^{N} \frac{1}{4} C_{i,j} x_j \geq \frac{1}{4} C_{i,j} x_j > 0 \tag{18}$$

  Absurd!

In this way we have demonstrate that $\left(\mathbb{1}_N - \frac{1}{4}C\right)$ is invertible and we can calculate all the estimates of average time to find food starting from any point in I.

To solve the puzzle I have found all the nodes inside the ellipse, defined the matrix C using an auxiliary "connection function", inverted the matrix $\left(\mathbb{1}_N - \frac{1}{4}C\right)$ and calculate the result of Eq. 14 for the point $O(0,0)$. The result I got is 13.99.

Then, **the average time to get food starting from** $(0,0)$**, approximated to the nearest integer, is 14**.

2.5.2 Simulation

This approach finds the exact solution, but it requires the inverse of a matrix whose dimensions scale with the square of the number of inside nodes. With a large number of points, this method requieres too many resources, so it could be inconvinient or even impractical. For this reason I have also implemented a program to simulate a 2D random walk of the ants.

Let N be the number of initial ants starting in $(0,0)$. At any time step n a random function generates the movement (up down, lef right) per each ant and several metrics are collected and displayed. An ant that reaches the food is considered "dead".

Let's introduce some quantities:

- t = time step
- d_t = number of ants dead at time n
- D_t = number of ants dead until time n
- A_t = number of ants alive until time n
- $N = D_t + A_t$ = initial number of ants
- $\mu_t = \frac{\sum_{s=1}^{t} d_s s}{D_t}$ = average time to arrive to the food starting from $(0, 0)$ calculated on the dead ants
- $\sigma_t = \sqrt{\frac{\sum_{s=1}^{t} d_s (s-\mu_t)^2}{D_t(D_t-1)}}$ = error of the mean μ_t using the standard deviation and assuming that N is large enough
- T = last time step performed by the program (T could be set as a parameter or could represent a time threshold based on the available resources to perform the simulation)
- E_t = our best estimate at time t of the average time to get food starting from $(0, 0)$

In particular, $\lim_{\substack{N\to\infty \\ T\to\infty}} E_T$ and $\lim_{\substack{N\to\infty \\ t\to\infty}} \mu_t$ convege to the true value E_{true} that is possible to calculate as described in the previous section. Our goal is to find E_T using the simulations.

First of all we have to consider 2 scenarios:

- $D_T = N$: all the ants are dead before reaching the time threshold
- $D_T < N$: probably we have low resources with respect to the assigned problem, so there are still alive ants before reaching the time threshold

In order to simplify the formulas I will not use the subscripts unless it is necessary.

Case: $\mathbf{D_T = N}$

It is possible to define E_T as follows:

$$E_T = \mu_T = \frac{\sum_{t=1}^{T} d_t t}{N} \tag{19}$$

Moreover, we have $\sigma_T = \sqrt{\frac{\sum_{t=1}^{T} d_t (t-\mu_T)^2}{N(N-1)}}$, and, using the law of large numbers, we can approximate the distribution of the mean to a gaussian distribution $\mathcal{N}(\mu_T, \sigma_T^2)$.

The best case occurs when $[\mu_T - \alpha\sigma_T] = [\mu_T + \alpha\sigma_T]$ with α as larger as possible (where $[x]$ represents the value of x rounded to the nearest integer). In this case, the solution to the problem would be with high probability $[\mu_T]$. For example, with $\alpha = 2$ the true value belongs to the interval $[\mu_T - \alpha\sigma_T, \mu_T + \alpha\sigma_T]$ with probability of about 95%.

Therefore, the best approach is to increse the sample size N (respecting the condition $D_T = N$) to reach a confortable confidence interval $[\mu_T - \alpha\sigma_T, \mu_T + \alpha\sigma_T]$ (for example with $\alpha > 3$) that allows us to return $[\mu_T]$ as solution to the problem with high probability.

Case: $\mathbf{D_T < N}$

At each time t, we could define E_t as follows:

$$E = \frac{D\mu + A(t + x)}{N} \tag{20}$$

where x_t represents an average time that all the alive ants need to arrive to the food. In the case of highly symmetrical boundaries and for t sufficently large, we have with high probability:

- $$E < \frac{D\mu + A(t + E)}{N} \tag{21}$$

  since we expect the ants to be spread on all the nodes and that the ants in a "central position" require more time to reach the food.

- $E_t > \mu_t$ since there are still alive ants, so their time to reach the food will be for sure greater than t. If we consider the fact that at time t the alive ants need at least 1 step to reach the food, we could find a better approximation:

$$E > \frac{D\mu + A(t+1)}{N} = \mu + \frac{A}{N}(t + 1 - \mu) \tag{22}$$

  A better approximation should consider the average minumum number of steps m_A to arrive to the food (i.e. the average minumum distance to closest bounday point). This leads to the following inequality:

$$E > \mu + \frac{A}{N}(t + m_A - \mu) \tag{23}$$

Knowing that $N = D + A$ and combining the 2 inequalities, we obtain:

$$LB_t < E_t < UB_t \tag{24}$$

with

$$\begin{cases} LB_t = \mu_t + \frac{A_t}{N}(t + m_A - \mu_t) \\ UB_t = \mu_t + \frac{A_t}{D_t}t \end{cases} \tag{25}$$

This equations shows that a good estimation for E_t belongs to the interval $[LB_t, UB_t]$. Even μ_t represents a good extimation of E_t with an error smaller than the empirical value $\frac{A}{D}t$.

Notice that, in a normal condition, a large number of ants N implies a large D and a smaller $\frac{A}{D}$ for t sufficently large. If t is of the same order of magnitude of N before reaching a good bound value $\frac{A}{D}t$, then μ_t does not represent anymore a good candidate of E_t. However, this scenario occurs only if the number of initial ants is not large enough to allow the ants to reach the food before t approaching N (i.e. the boundaries are too far away from the starting point to create a significant statistical measure of μ_t based on N samples and time threshold T).

On the other hand, $E_t > \mu_t$ must hold. Using Eq. 20, we obtain $t + x_t > \mu_t$, that is satisfied with high probability for $t > \mu_t$.

Therefore, the best approach is to continue the simulation until $t = T$ or we found a proper error thereshold $\frac{A}{D}t$ that guarantees $[LB_t] = [UB_t]$ (and possibly $= [\mu_t]$). In that case, the solution to the problem should be $[UB_t]$ (and eventually $[\mu_t]$).

3 Code

The scripts I created to solve the puzzle are availabe on my GitHub page: github.com/prenassipaolo/optiver_puzzle.