EVOLUTIONARY DYNAMICS AND ITS REAL-WORLD IMPLICATIONS

YOUZE ZHENG

Abstract. The paper explores three stages of evolution – replication, selection, and mutation – using differential equations and probability models. The relationships between the population size and time are revealed by solving the differential equations. The probability models give the probability for one type of population to dominate the whole. The dynamics of these biological processes are visualized upon the discussion of simplices and the simplicial complex. At the end of this paper, two linear regression methods are introduced to help compute the numerical value of fitness, one of the critical components of the models discussed throughout the paper.

1. Introduction

We can split the topic "Evolutionary Dynamics" into two parts – evolution and dynamics. What is evolution? How are dynamics involved in the processes of evolution? I asked myself these questions, and I believe those are questions people ask when they first hear the topic. *Evolutionary Dynamics* is a broad topic that studies how biological creatures evolve over time. The evolution could be due to linguistic changes, cultural changes, etc. As far as biologists are concerned, evolution mainly involves three processes: replication, selection, and mutation; where replication refers to the process individuals reproduce, selection refers to the case that one individual is chosen to evolve towards one particular direction, and mutation means the change of type of individual as a result of the change in things like DNA sequence. The dynamics come after evolution; it reveals the interaction among different population types as each evolves. Therefore, it is important for us to discuss evolution first so that we can dive deeper into the dynamics part. For the sake of simplicity, most of our discussions are limited to two types of individuals.

Though we can describe the three processes of evolution in words as we did above, mathematics enables us to express each of them explicitly. From a deterministic view of the topic, we can use continuous mathematics, the equations, to show the relationship between population size and time. In the simplest case, we can use the exponential growth model to show how one individual could reproduce repeatedly. The rate at which an individual grows depends on the numerical value of fitness, the reproductive success, of the population that individual belongs. By solving the differential equation of the growth, we could show how the size of the population can approach infinity as time goes to infinity. Additionally, since some individuals would die while some reproduce, we can compare the death rate with reproductive success, thus exploring the destination of the population as time passes long enough. However, it is not possible for a type of individual to reproduce endlessly if the reproductive success is greater than the death rate because one given environment would have limited nutrients to feed and finite resources for individuals to rely upon. Therefore,

Date: July 10, 2022.

the maximum carrying capacity – the maximum number of individuals that one environment could contain – is introduced to our discussion, and we are able to express the relationship between the size of the population and time using the logistic model. Similarly, we are able to use differential equations to express the process of selection and mutation by exploring the factors that affect the rate of change in population size.

More than showing the evolution with mathematical equations, we can also use simplices to visualize the evolution. We can use a 1-simplex line segment to show the change of population structure between two types of the population; a 2-simplex filled triangle can do the same for three types of individuals; and a 3-simplex tetrahedron could be used to represent the case for four types of individuals. Suppose we are discussing a more complicated case that involves more than four types of population. In that case, we can use simplical complex to show the relationship among all different types of population. The discussion of simplices helps us visualize the evolutionary changes and make the problem more manageable. It would sometimes be difficult to use equations to express the evolution of many different types of individuals. The discussion of evolution using simplices enables us to show that population structure could evolve over time due to factors that affect the evolution of each type of individual involved, so the dynamics underlying evolution could be represented using simplices.

Besides the deterministic view of evolution, we can also view the *Evolutionary Dynamics* discretely, using probability. Based on the Moran Process that tells the principle of evolutionary selection, we can use two probability models to compute the probability for one type of individual to dominate the whole population. The first one uses a transition matrix to represent the probability for one individual to be chosen for reproduction, elimination, or not. We can interestingly find an arithmetic sequence to represent the probability for an individual to dominate and conclude the probability for one individual to dominate at any starting position. In the second model, variables are frequently used to simplify the expressions. We can use the variables introduced to represent the probability for one type of individual to dominate. The probability could be expressed by the fitness of that population if we take one step further to consider fitness in our second model.

Finally, since fitness is often mentioned in the models throughout the paper, I describe two different computational methods to compute the fitness of one specific population using linear regression – one is the gradient descent algorithm commonly used in machine learning, and the other is the Ordinary Least Square (OLS) technique that gives a closed-form solution of the best-fit line for the given dataset. The advantages and disadvantages of each are evaluated at the end of the method description. Also, a comparison between the two is made to clarify the differences.

2. Replication

Among three biological processes – replication, selection, and mutation – replication is the most important. It is the basis for the latter two. Therefore, introducing the mathematical representations of replication can lay a foundation for understanding the mathematical models in later sections of this paper.

2.1. **Replication without Constrains.**

The simplest case of replication is unbounded replication, meaning the population could grow forever. Suppose there is an environment for one population to grow, which means there is no disease that would cause death, no lack of food that would cause the lack of nutrients, and, ideally, every individual is capable

of reproducing. Assume that time is measured in days; the population would grow at a rate r in such a perfect environment, which means the population would go through the reproduction process r times each day. Let $x(t)$ denote the size of population at time t. We can use the differential equation below to represent the rate of reproduction for the whole population [Now06, Chapter 2.1]:

$$\frac{dx}{dt} = rx. \tag{2.1.1}$$

By equation (2.1.1), the following computation could be carried out to find the original exponential growth equation for the given population.

Firstly, divide both sides of the equation by x and multiply each side by dt.

$$\frac{dx}{x} = r\,dt.$$

Then, take the integral of both sides and solve for the equation.

$$\int \frac{1}{x}\,dx = \int r\,dt$$
$$\log(x) = rt + C.$$

Exponentiating both sides of the equation, the general solution for the differential equation is obtained.

$$x(t) = e^{rt+C}. \tag{2.1.2}$$

where C is an arbitrary constant.

Before replication taking place, the original population size could be denoted by x_0 at time $t = 0$. Therefore, the size of population whose rate of reproduction is r is solved for to be following equation:

$$x(t) = x_0 e^{rt}. \tag{2.1.3}$$

To better illustrate this model, let us consider how bacterial cells divide [Now06]. For a bacterial cell that lives perfectly with all nutrients needed for growth, cell division would take place as few as every 20 minutes [Jet15]. Therefore, in every 20 minute, one bacterial cell would divide into 2 daughter cells, the next generation of the bacterial cell family. Therefore, the cell division could take place at 72 daily ($24 * 60/20 = 72$). The total number of cells after one day span could be expressed as $x_0 e^{216}$, given that the initial number of parent cells is x_0.

2.2. **Bounded Replication.** The unbounded reproduction is very idealized and it would not actually take place in reality. In real life, some individuals reproduce yet some die. Suppose the natural death rate for a particular population is d, we would have the following differential equation to represent the rate of change of population size [Now06, Chapter 2.1]:

$$\frac{dx}{dt} = (r - d)x. \tag{2.2.1}$$

Following a similar computation to the derivation of equation that represent the population size at given time t, the solution to the differential equation is

$$x(t) = x_0 e^{(r-d)t}. \tag{2.2.2}$$

In this case, we can use the reproductive ratio, r/d, to show if the size of the population would expand or contract, in which r means growth rate and d represents death rate. The

ratio shows how many offspring each individual is expected to have, and the death rate d reveals that the average lifespan for an individual within the population is $1/d$. It is clear that if $r/d < 1$, then the growth rate r is less than the death rate d. Therefore, more proportion of the given population would die than the proportion of the population that would reproduce. Over time, the whole population would die out, given that nothing has changed. Conversely, if r is greater than d, more individuals are reproduced than those who die. As a result, the total population would expand over time.

However, does $r > d$ mean the population would grow indefinitely, as the model explained in 2.1 that grows exponentially? Ideally, it will, if there are infinitely much food and no predators and diseases. Yet, in reality, the indefinite growth seems unreasonable. The population will not grow forever and cannot do that either. In real life, the population has finite resources to rely upon. If the population size grows large enough, there will surely be competition for resources occur. Consequently, there exists a maximum number of population, technically called maximum carrying capacity, that the population size cannot grow further. Suppose the carrying capacity for one specific population is K, with the initial population size x_0, the number of individuals for that population at time t could be represented by the logistic equation as following [Now06, Chapter 2.1]:

$$\frac{dx}{dt} = rx(\frac{1-x}{K}) \tag{2.2.3}$$

where r represents the rate of reproduction and population size is denoted by x.

As the population size x increases, the rate of change of population dx/dt decreases and eventually becomes 0 when $x = K$, which explains why the population size would not expand further as the number of individuals hits carrying capacity K. We could use the following computation to find the relationship between population size x and time t.

$$\frac{dx}{dt} = rx(\frac{1-x}{K}) = -rx(\frac{x-1}{K})$$

Multiply each side of the equation by dt, we have

$$dx = -(\frac{rx(x-1)}{K})\, dt.$$

Dividing both sides by $-\frac{x(x-1)}{K}$, we have

$$-\frac{K\, dx}{x(x-1)} = r\, dt.$$

Taking the integral on both sides, we then have

$$\int -\frac{K\, dx}{x(x-1)} = \int r\, dt.$$

By observation, we can see $1/x(1-x) = 1/(x-1) - 1/x$, so we can compute the next steps using integration by partial fractions

$$\begin{aligned} -K \int (\frac{1}{x-1} - \frac{1}{x})\, dx &= \int r\, dt \\ K(\log(x) - \log(x-1) &= rt + C \\ K \log(\frac{x}{x-1}) &= rt + C. \end{aligned}$$

Exponentiating both sides of the equation above and changing the ordering of the equation, we can get the general solution of $x(t)$

$$x(t) = \frac{e^{rt+C/K}}{e^{rt+C/K}+1}. \tag{2.2.4}$$

where C is an arbitrary constant.

In our case, since $x(t) = x_0$ at $t = 0$ and $x(t) = K$ when $t \to \infty$, it could be concluded that

$$x(t) = \frac{Kx_0e^{rt}}{K + x_0(e^{rt}-1)}. \tag{2.2.5}$$

3. Selection

Beyond single population reproduction, selection would take place when more than one type of population exists in one given environment. Due to their different adaptations to the natural environment, different types of individuals would have varied reproduction and death rates. Consequently, some would eventually die out, while some would flourish, depending on their survival abilities.

3.1. **Selection with unlimited reproduction.**

To simplify the case, we can first discuss the scenario in which natural death does not happen. Suppose we have only two types of individuals competing for resources, A and B. The reproduction rate now could represent the fitness, the quantitative representation of individual reproductive success, of each population, as the natural death rate is 0 under our assumption. Denote the size of population A by $x(t)$, with fitness a, and the size of population B by $y(t)$, with fitness b. The following two differential equations would hold:

$$\begin{aligned} \frac{dx}{dt} &= ax \\ \frac{dy}{dt} &= bx. \end{aligned} \tag{3.1.1}$$

As showed previously, the solution to the first differential equation is $x(t) = x_0e^{at}$ and the solution to the second one is $y(t) = y_0e^{bt}$. Let us now consider the two equations together.

Let $p(t) = x(t)/y(t)$, which, similar to the reproductive ratio, reveals the relative population size between population A and B at any given time t. Differentiate both sides of the equation, let we can get

$$\frac{dp}{dt} = \frac{x'y - xy'}{y^2} = (a-b)p. \tag{3.1.2}$$

Using the same computation we carried out in 2.1, given that $p_0 = x_0/y_0$ at $t = 0$, we can obtain the following result:

$$p(t) = p_0 e^{(a-b)t}. \tag{3.1.3}$$

The solution reveals that if $a > b$, then p tends to grow to infinity, which means A would out-compete B if $t \to \infty$. Conversely, if $a < b$ is the case, then population B would out-compete population A, and therefore p converges to zero as $t \to \infty$.

3.2. **Selection with limited reproduction.** As suggested by the logistic model in 2.2, each type of individual would hit a maximum carrying capacity that the population size could not grow beyond. Therefore, the whole environment would have a maximum carrying capacity as well. We could have the average fitness between two populations denoted by ϕ. For a better illustration, we can use $x(t)$ now to represent the proportion of population A in the environment and $y(t)$ to denote the proportion of population B that makes up the total population in the environment. By our definition to $x(t)$ and $y(t)$, we can conclude that $x + y = 1$ and $\phi = ax + by$ because they represent the proportions that range from 0 to 1. In this case, we have the following equations:

$$\begin{aligned} \frac{dx}{dt} &= x(a - \phi) \\ \frac{dy}{dt} &= x(b - \phi). \end{aligned} \tag{3.2.1}$$

By replacing y with $1 - x$, we have

$$\frac{dx}{dt} = x(1 - x)(a - b). \tag{3.2.2}$$

We would have some critical observations without the need to compute the solution to the differential equation.

- If $a > b$, then the rate of change of population A, dx/dt, is always greater or equal to 0. The reason is that the proportion of population A to the total population of $A + B$, denoted by $x(t)$, ranges between 0 and 1. Any positive term of $(a - b)$ would result in a dx/dt that is greater or equal to 0. As a result, the selection would favor population A, meaning that the size of population A would grow larger and larger and eventually dominate the whole ecosystem.
- If $a < b$, then the opposite situation would happen. Following a similar analysis, we would conclude that dx/dt is always less than or equal to zero, implying that dy/dt is always greater or equal to zero. As a result, population B would eventually dominate the ecosystem, and population A would die out at the end.
- If $a = b$, then there exists an equilibrium point in between such that the numerical values of $x(t)$ and $y(t)$ remain the same as time progresses. Both population A and B are reproducing at the same rate, so their relative proportion would not change over time.

Figure 1. 1-simplex line segment, $\sigma =< A, B >$

We can use the above 1-simplex, a line segment, to visualize the situation with two types of population [Now06, Chapter 2.2]. Any point in the line would reveal the relative proportion of each population to the whole group. Suppose the initial population distribution is at point C. It is clear to observe that the abundance of population B exceeds the abundance of population A at time $t = 0$. If population A reproduces at a faster rate throughout the time, meaning $dx/dt > dy/dt$, then at some time, the population distribution would be represented by C_1 and eventually by the endpoint A. Conversely, if $dx/dt < dy/dt$, then the population distribution would at some point be represented by point C_2, and ultimately endpoint B. If $dx/dt = dy/dt$, then the population distribution would remain at point C on the 1-simplex. It should be clarified that dx/dt and dy/dt can both be greater than zero if the maximum carrying capacity is not met. Otherwise, one would be positive, the other would be negative, or both are zero.

4. Population structure with multiple types of individuals

We have shown the methods to represent the reproductive dynamics with at most two types of different individuals. What if it is the case that there are more than two types of the population? Suppose x_i represents the relative proportion of population i in the given environment, and f_i represents the fitness of that population. Then for all $i = 1, 2, ..., n$, the average fitness could be computed by the following equation:

$$\phi = \sum_{i=1}^{n} x_i f_i. \tag{4.0.1}$$

Similar to the differential equations in 3.2, the rate of change of population i for all $i = 1, 2, ..., n$ in thic case would be

$$\frac{dx_i}{dt} = x_i(f_i - \phi). \tag{4.0.2}$$

Since there exists a maximum carrying capacity, as discussed before, the total size of the population would remain the same over time. Therefore, $\sum_1^n x_i = 1$ and $\sum_1^n dx_i/dt = 0$. If the relative proportion of type i individuals increases, then it means $x_i > 0$, implying that the fitness of population i is greater than the average fitness of the whole population of different types.

Graphically, the situation of the reproductive dynamics among three types of population could be demonstrated by a 2-simplex filled triangle, each vertex represent one type of population.

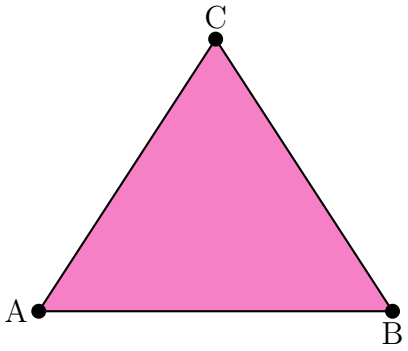

Figure 2. 2-simplex filled triangle, $\sigma =< A, B, C >$

The graph has three components: vertices A, B, and C; edges AB, AC, BC, and interior filled region bounded by the edges. Suppose there are three types of individuals A, B, and C. We could use a diagram to show three general scenarios.

(1) If the population distribution is represented by the vertex A, then it means as $t \to \infty$, both population B and C would die out, and only population A remains. Similarly, vertices B and C represent the situations in that only B exists at the end and only C survives eventually, respectively. Therefore, if the population distribution is represented by one vertex, one population dominates the ecosystem.
(2) If the edges represent the population distribution, then it means one type of population eventually dies out, and the other two remain in the ecosystem. The relative proportion of the remaining two types of the population could be represented by the points on edge, as explained before, which is determined by their fitness relative to each other. For example, if the edge, or a line segment, AB is used to describe the population distribution, then type C eventually loses the competition, maybe due to its relatively lower fitness compared with the other two. Type A and type C individuals coexists in the ecosystem to share the given resources.
(3) If any point in the interior region is used to represent the population distribution, then three types of the population would coexist at different proportions, according to their fitness.

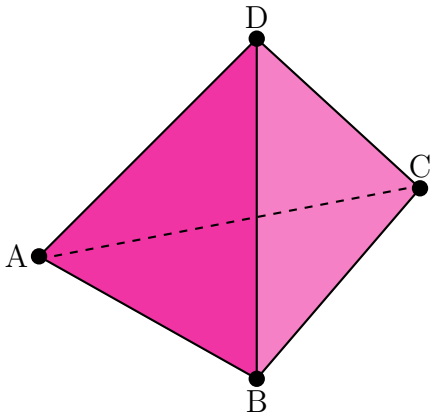

Figure 3. 3-simplex tetrahedron, $\sigma =< A, B, C, D >$

Now, let us consider the case when there are four different types of population. Suppose type D is introduced to the environment that type A, B, and C exited before. The reproductive dynamics could be represented by a tetrahedron, 3-simplex, shown above, combining three 2-simplex triangles. The cases of vertices and edges have been explained in the 2-simplex part. The face of a tetrahedron, or the 2-simplex contained in it, shows something different than it does in the three-type case. There are two additional representations involved:

(1) Each triangular face of the tetrahedron represents the coexistence of three types of individuals, with one other type dying out as $t \to \infty$. For example, in face ABD, type C individuals do not exist and type A, B, and C individuals each represents a part of the total population.
(2) Any point inside the 3-simplex tetrahedron $\sigma =< A, B, C, D >$ represents the case when all 4 types of the population share the same environment over time. Again, the proportion of each type depends on the relative fitness to each other.

We have explained how 1-simplex line segment, 2-simplex filled triangle, and 3-simplex tetrahedron could be used to show the reproductive dynamics among different types of individuals. We have to mention that 0-simplex, a point, is also a basic representation of the single type population environment, although there will not be as much dynamic as for the case of multiple types.

After introducing simplices, we can use the simplicial complex, a set of composed points, line segments, triangles, and their n-dimensional counterparts [Mat03], to represent an environment that involves the interaction among even more types of individuals. I would not explain deeply into the concept but rather give a brief taste of what it is about, as it would be very complicated to explain and understand. Here is the graph of one arbitrary simplicial complex, yet containing much information to be analyzed.

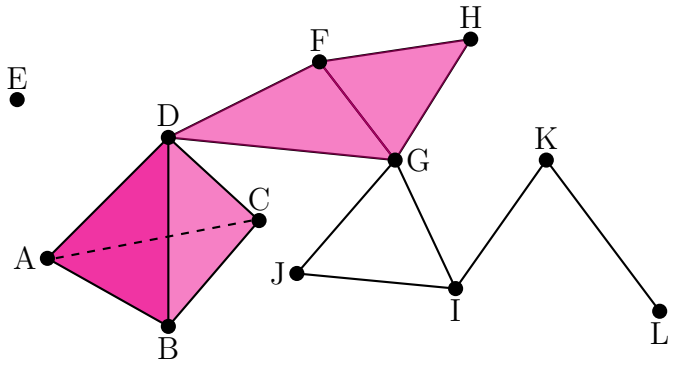

Figure 4. Simplicial Complex $\mathcal{K}$

As before, each vertex, or node, in Figure 4 represents one type of population. As interpreted before, we could have reproductive dynamics within each simplex in the simplicial complex $\mathcal{K}$.

What is more complicated is that the dynamics within one simplex would affect the dynamics in the other simplices that are connected with it. Two simplices are connected because there are one, two, or more types of individuals that live both with some types of individuals in one simplex and some in the other simplex. Therefore, that common population type

becomes the medium of the dynamic between two or more simplices. Such connections are important components of how we explain the reproductive dynamics among multiple types of individuals using the simplicial complex diagram.

For example, the 0-simplex $\sigma_0 =< E >$ has no connection with any other simplices in Figure 4. Therefore, this type of individual would not live with any types of individuals from A to L, but on its own. On the other hand, the 3-simplex $\sigma_1 =< A, B, C, D >$ is connected with 2-simplex $\sigma_2 =< D, F, G >$. Therefore, the dynamics among population A, B, C, and$D$ would affect the population structure. If the dynamics within σ_1 move towards triangle ABC; or edge AB, BC, or AC, then the population proportion of type D individuals decreases. Equivalently, the fitness of population type D diminishes. If at one point, the fitness of the type D population reduces low enough, smaller than both type F's and type G's, then the reproductive dynamics within σ_2 would also approach edge FG or vertices F and G.

For simplices σ_2 and $\sigma_3 =< F, H, G >$, they share two common types of individuals, F and G. Therefore, the affect in the change of dynamics in σ_2 to σ_3 is determined by both type F and type G individuals.

One interesting case is found among three 1-simplices $\sigma_4 =< G, J >$, $\sigma_5 =< G, I >$, and $\sigma_6 =< J, I >$. Each is connected with the other two, but they do not form a 2-simplex. One explanation is that only two of them could live in the same environment, and each type of individual could only adapt to two of the three existing environments. For example, suppose type J individuals could live in an environment 1 and 2. Type G population could live with type J in environment 1 but never in environment 2, but type I individuals could live with type J in environment 2 but not I. Besides the two environments, there exists an environment 3 that could accommodate both types G and I. Although three types of individuals could not live in one environment and interact directly, there will be indirect interaction among the three due to the interaction between two of them.

The case of 0-simplex analysis would follow the same logic for σ_2 and σ_3. The shared population type's change in fitness in one environment would eventually affect the dynamic in the other dynamic.

5. Mutation

The last most important biological process is the mutation, which involves the unexpected change in the DNA sequence that results in a change in population type. If the mutation is considered, the reproductive dynamic will change slightly, as there would be some individuals within one population changing to other types and some other types changing to that type.

5.1. **Mutation with two population types.**

Lastly, we can discuss the models of mutation. If there are only two types of individuals, A and B, then the case is straightforward: some type A individuals would change to type B, while some type B individuals would join type A due to random mutation. To simplify our discussion, population A and B are assumed to have equal fitness, which means any reproduction among two population types will not change the population structure in that given environment. Suppose the mutation rate for type A population is u_A, and the mutation rate for population B is u_B. Given the

average fitness of two population types is ϕ, we have the following two differential equations [Now06, Chapter 2.3]:

$$\begin{aligned} \frac{dx}{dt} &= x(1-u_A) + yu_B - \phi x \\ \frac{dy}{dt} &= y(1-u_B) + xu_A - \phi y. \end{aligned} \tag{5.1.1}$$

where x and y both represent the proportion of each population type among the total population.

By the definition of variables x and y, and our assumption that two population types have the same fitness, we could have $x + y = 1$ and $\phi = 1$. Therefore, dx/dt could be further evaluated to

$$\frac{dx}{dt} = u_B - x(u_A + u_B). \tag{5.1.2}$$

We can follow the computation below to solve for the differential equation.

$$\begin{aligned} \frac{dx}{dt} &= u_B - x(u_A + u_B) \\ x(t) &= e^{-\int (u_A+u_B)\,dt}\Big(\int u_B\, e^{\int (u_A+u_B)\,dt}\,dt + C\Big) \\ &= e^{-(u_A+u_B)t}\Big(\frac{u_B}{u_A+u_B}\, e^{(u_A+u_B)t} + C\Big) \\ &= \frac{u_B}{u_A+u_B} + \frac{C}{e^{(u_A+u_B)t}} \end{aligned}$$

where C is an arbitrary constant.

Therefore, we have

$$\lim_{t\to\infty} x(t) = \lim_{t\to\infty}\Big(\frac{u_2}{u_1+u_2} + \frac{C}{e^{(u_1+u_2)t}}\Big) = \frac{u_2}{u_1+u_2}. \tag{5.1.3}$$

As time passes long enough, the population size A converges to $u_2/(u_1 + u_2)$, given that A has the same fitness as B.

5.2. **Mutation with more than two population types.** It is often the case that multiple types of individuals are living in the same environment. Therefore, the direction of mutation could go anywhere within the given environment. We can use the notation q_{ij} to represent the rate of mutation for which one type i individual would mutate to type j. Therefore, we can write the mutation dynamics as follows:

$$\frac{dx}{dt} = \sum_{j=1}^{n} x_j\, q_{ji} - \phi x_i \tag{5.2.1}$$

for all $i = 1, 2, ..., n$.

6. Stochastic view of evolution

In the deterministic process of evolution, each variable changes according to the differential equations we have explained. However, in the stochastic view, evolution takes place due to random fluctuation, which means there are probabilities for reproduction, death, and mutation. So we would only have discrete integer numbers of individuals rather than continuous numbers that deterministic models could calculate. For the purpose of the paper, I will only discuss the stochastic process for reproduction and death and leave mutation, as the probability of mutation would be complicated to discuss clearly and require much knowledge in Biology.

6.1. **Moran Process.**

To discuss the stochastic view, we have to introduce Moran Process first and foremost, which is a simple stochastic model. Named after the Australian geneticist Patrick Moran, the Moran Process tells that at each time step, a random individual is chosen for reproduction, and a random individual is chosen for elimination [Now06, Chapter 6.1]. The reproduced individual replaces the eliminated individual, so the total population remains unchanged. Since the population size is unchanged, the process would end with one type of individual dominating the whole population.

Suppose there are only two types of individuals, type A and type B, with a total population size of N. If there are i individuals for type A population, then there are $N-i$ type B individuals for all $i = 1, 2, ..., N$. Therefore, the probability that a type A individual is randomly chosen is i/N. The probability of one type B individual being randomly chosen is $(N-i)/N$. As such, we have four different probabilities of randomly choosing two individuals within the population.

(1) One type A individual is chosen for reproduction, and another type A individual is chosen for elimination. In this case, the number of type A individuals will not change because $i - 1 + 1 = i$ and the probability for this is $1/N * 1/N = 1/N^2$.
(2) One type B individual is chosen for reproduction, and another type B individual is chosen for elimination. Same as in the previous case, the number of type B individuals will not change, and the probability for this case is $(N-i)/N*(N-i)/N = (N-i)^2/N^2$.
(3) One type A individual is chosen for reproduction, and one type B individual is chosen for elimination. The number of type A individuals becomes $i+1$ and the number of type B individuals becomes $(N-i-1)/N$. The probability for this to happen is $i/N * (N-i)/N = i(N-i)/N^2$.
(4) One type B individual is chosen for reproduction, and one type A individual is chosen for elimination. Unlike the previous case, the number of type A individuals becomes $i-1$. The probability for this case is the same as before, $i(N-i)/N^2$, as the probabilities of choosing one individual to reproduce and die are the same.

6.2. **Birth-Death Process Model I.**

The probabilities of change of population structure could be represented by a N by N transition matrix. Suppose we have a transition matrix $P = [p_{ij}]$, which represent the probability for an individual to move from position i to j, we

would have the following expressions to represent every entries:

$$
\begin{aligned}
p_{i,i-1} &= \frac{i(N-i)}{N^2} \\
p_{i,i+1} &= \frac{i(N-i)}{N^2} \\
p_{i,i} &= 1 - p_{i,i-1} - p_{i,i+1} = \frac{(N-i)^2}{N^2} \\
p_{0,0} &= p_{N,N} = 1 \\
p_{0,i} &= p_{N,i} = 0
\end{aligned}
\tag{6.2.1}
$$

For such a matrix, the sum of each row is one. The reason why $p_{0,0} = p_{N,N} = 1$ is because once i, the number of type A population becomes 0, the number of type A population cannot increase. Similarly, when $i = N$, the environment is dominated by the type A population. Therefore, the type B population could never come back again. So once $i = 1$ or $i = N$ is reached, the population structure will no longer change and stay forever. This also explains why $p_{0,i} = p_{N,i} = 0$, as once one type of population dies out, they cannot dominate the whole population, not even coming back. The states of $i = 0$ and $i = N$ are called "absorbing states," meaning that one type of population absorbs the other. The states when $i = 1, 2, ..., (N-1)$ are called transient states. The population structure will remain at transient stages for a limited time and eventually reach one of the absorbing states, either all type A individuals or all type B individuals.

Now that we know the probability of every move in one time step, how can we show the probability that the population size A starts from i and ends with N? That is a more specific question we want to ask because we can tell the exact probability for population A to dominate the whole population, starting from point i.

Here is the way we compute it. Let x_i denotes the probability for the number of population A to change from i to N. Then the probability for population B to dominate is $1 - x_i$. We would have the following general equations

$$
\begin{aligned}
x_0 &= 0 \\
x_N &= 1 \\
x_i &= p_{i,i-1}x_{i-1} + p_{i,i}x_i + p_{i,i+1}x_{i+1}
\end{aligned}
\tag{6.2.2}
$$

for all $i = 1, 2, ..., (N-1)$.

Again, it is impossible for type A individuals to flourish and dominate the whole population if the starting point i is zero. The starting point of $i = N$ is a guaranteed nomination for type A individual, given that there is no mutation taking place. For any integer values for $i = 1, 2, ..., (N-1)$, the probability of type A population eventually dominate the whole population would be the sum of 1)the probability of moving from i to $i-1$ multiplied by the probability for type A to dominate starting at $i-1$, x_{i-1}; 2) the probability of remaining at the same position i multiplied by x_i; 3) the probability of moving from i to $i+1$ multiplied by x_{i+1}.

Since $p_{i,i-1} = p_{i,i+1}$, as explained before, $p_{i,i} = 1 - p_{i,i-1} - p_{i,i+1} = 1 - 2p_{i,i+1}$. We can further reduce the expression of x_i as below:

$$\begin{aligned} x_i &= p_{i,i-1}x_{i-1} + p_{i,i}x_x + p_{i,i+1}x_{i+1} \\ x_i &= p_{i,i-1}x_{i-1} + (1 - 2p_{i,i+1}) + p_{i,i+1}x_{i+1} \\ 2p_{i,i+1}x_i &= p_{i,i-1}x_{i-1} + p_{i,i+1}x_{i+1} \end{aligned}$$

for all $i = 0, 1, ..., N$.

Dividing both sides by $p_{i,i+1}$, we would get

$$2x_i = x_{i+1} + x_{i-1} \tag{6.2.3}$$

By observation, we can conclude that x_i follows an arithmetic sequence, with common difference $d = (x_N - x_0)/N - 0 = 1/N$. Therefore, we can easily conclude that

$$x_i = \frac{i}{N} \tag{6.2.4}$$

for all $i = 0, 1, ..., N$.

6.3. **Birth-Death Process Model II.** Besides using i/N to represent the possibilities that type A population would dominate the whole population starting at size i, we have another method to do the computation – using variables.

We have the exact same three moves as the previous representation for type A population: from i to $i-1$, remain at i, and from i to $i+1$. Denote α_i the probability that the number of type A population would increase by 1, from i to $i+1$, β_i the probability that the number of type A individuals changes from i to $i-1$. So the probability that the number of type A individuals remains the same is $1 - \alpha_i - \beta_i$. Similar to (6.2.2), we would have the following three equations:

$$\begin{aligned} x_0 &= 0 \\ x_N &= 1 \\ x_i &= \beta_i x_{i-1} + (1 - \alpha_i - \beta_i)x_i + \alpha_i x_{i+1} \end{aligned} \tag{6.3.1}$$

for all $i = 1, 2, ..., (N-1)$.

Let $y_i = x_i - x_{i-1}$ for all $i = 1, 2, ..., N$, which is the difference in probability for population A to dominate the whole, between two adjacent integer starting numbers. We would have $x_i = y_i$ for all $i = 0, 1, ..., N$. It is interesting that $\sum_{i=1}^{N} y_i = x_1 - x_0 + x_2 - x_1 + ... + x_N - x_{N-1} = x_N - x_0 = 1 - 0 = 1$. Furthermore, let us denote $\gamma_i = \beta_i/\alpha_i$ for the purpose of our model [Now06, Chapter 6.2]. Combining these expressions and (6.3.1), we could get $y_{i+1} = \gamma_i y_i$. The computation is as following:

$$\begin{aligned} x_i &= \beta_i x_{i-1} + (1 - \alpha_i - \beta_i)x_i + \alpha_i x_{i+1} \\ (\alpha_i + \beta_i)x_i &= \beta_i x_{i-1} + \alpha_i x_{i+1} \\ \beta_i(x_i - x_{i-1}) &= \alpha_i(x_{i+1} - x_i) \\ \beta_i y_i &= \alpha_i y_{i+1}. \end{aligned}$$

Dividing two sides of the equation by α_i, we would get

$$y_{i+1} = \frac{\beta_i}{\alpha_i} y_i = \gamma_i y_i = \prod_{k=1}^{i} \gamma_k x_1 \tag{6.3.2}$$

for all $i = 1, 2, ..., (N-1)$.

In case $i = 1$, we have $y_2 = \gamma_1 y_1 = \gamma_1 x_1$. From this equation, we can also get $y_1 = x_1$ by dividing each side of $\gamma_1 y_1 = \gamma_1 x_1$ by γ_1. For $i = 2$, we can get $y_3 = \gamma_2 y_2 = \gamma_2 \gamma_1 x_1$. The next terms of y_i for $i = 3, 4, ..., N$ could be computed as well. By inspection, every term of y_i includes x_1. By summing all terms of y_i, we could compute x_1 as following:

$$\begin{aligned} y_1 + y_2 + y_3 + ... + y_N &= \sum_{i=1}^{N} y_i \\ x_1 + x_1\gamma_1 + x_1\gamma_1\gamma_2 + ... + \gamma_{N-1}...\gamma_2\gamma_1 &= 1 \\ x_1(1 + \sum_{j=1}^{N-1} \prod_{k=1}^{j} \gamma_k) &= 1. \end{aligned}$$

Therefore, by dividing both sides by $1 + \sum_{j=1}^{N-1} \prod_{k=1}^{j} \gamma_k$, we get

$$x_1 = \frac{1}{1 + \sum_{j=1}^{N-1} \prod_{k=1}^{j} \gamma_k}. \tag{6.3.3}$$

Since $y_i = x_i - x_{i-1}$, as defined it is, we can follow the following recursion to express x_i in terms of what we have already had.

$$\begin{aligned} y_i &= x_i - x_{i-1} \\ x_i &= y_i + x_{i-1} \\ &= x_1\gamma_{i-1}...\gamma_2\gamma_1 + y_{i-1} + x_{i-2}. \end{aligned}$$

By computing the recursion over and over, we would eventually be able to express x_i by the following expression, we could show the representation of x_i using variables.

$$x_i = x_1(1 + \sum_{j=1}^{i-1} \prod_{k=1}^{j} \gamma_k). \tag{6.3.4}$$

Together with (6.3.3), we obtain

$$x_i = \frac{1 + \sum_{j=1}^{i-1} \prod_{k=1}^{j} \gamma_k}{1 + \sum_{j=1}^{N-1} \prod_{k=1}^{j} \gamma_k}. \tag{6.3.5}$$

So far, we have shown another computation method to find the probability for one type of individual to dominate the whole population, denoted by x_i.

6.4. **Take fitness into account.** [Now06, Chapter 6.3] Now, let us take one step further to include fitness in our discussion for the model. Still, we only have two types of individuals. Our discussion is based on the model described in 6.3. For the simplicity of comparison and calculation, let us assume that population A has fitness r and population B has fitness 1. Following the explanation in 3.1, population A would out-compete population B if $r > 1$. Conversely, if $r < 1$, then population would die out as $t \to \infty$.

Suppose the number of type A individuals is still represented by i and the number of type B individuals is therefore represented by $N - i$, given that the total number of population is N. After considering the fitness of each population type, the probability for a type A

individual to be chosen for reproduction is expressed as $ri/(ri+N-i)$. The probability for a type B individual to be chosen to reproduce is $(N-i)/(ri+N-i)$. The probability for any type A individual to be eliminated is $1/N$, and for any type B individual to be chosen for elimination is $(N-i)/N$. Therefore, similar to (6.2.1), we can express each entry of the transition matrix $P=[pij]$ as follows:

$$\begin{aligned} p_{i,i-1} &= \frac{N-i}{ri+N-i}\,\frac{i}{N} \\ p_{i,i+1} &= \frac{ri}{ri+N-i}\,\frac{N-i}{N} \\ p_{i,i} &= 1-p_{i,i-1}-p_{i,i+1} \end{aligned} \tag{6.4.1}$$

for all $i=1,2,...,(N-1)$.

To have population type A move from i to $i-1$, one type A individual has to be chosen for elimination. At the same time, there needs to have one type B individual reproduced. Therefore, the result of $p_{i,i-1}$ is the product of the probability for one type A individual to be eliminated and the probability for one type B individual to be reproduced.

Similarly, entries $p_{i,i+1}$ should be the product of the probability of one type A individual being chosen for reproduction and one type B individual being chosen for elimination.

Since there are only three possible moves for the size of the type A population – reduce by 1, increase by 1, and size remain the same – the probability that the total number of type A individuals stays unchanged would be $1-p_{i,i-1}-p_{i,i+1}$.

Let us denote $\gamma_i = p_{i,i-1}/p_{i,i+1}$, similar to $\gamma_i=\beta_i/\alpha_i$ used in 6.3. Following our representations for $p_{i,i-1}$ and $p_{i,i+1}$, we can easily get

$$\gamma_i = \frac{1}{r} \tag{6.4.2}$$

for all $i=1,2,...,(N-1)$.

Since the value of γ remains constant, we could conclude that $\prod_{k=1}^{j}\gamma_k$ follows a geometric sequence. Therefore, we have

$$\begin{aligned} x_i &= \frac{1+\sum_{j=1}^{i-1}\prod_{k=1}^{j}\gamma_k}{1+\sum_{j=1}^{N-1}\prod_{k=1}^{j}\gamma_k} \\ &= \frac{1+\gamma_1+\gamma_1\gamma_2+...+\gamma_1\gamma_2...\gamma_{i-1}}{1+\gamma_1+\gamma_1\gamma_2+...+\gamma_1\gamma_2...\gamma_{N-1}} \\ &= \frac{1+\frac{1}{r}+...+\frac{1}{r^{i-1}}}{1+\frac{1}{r}+...+\frac{1}{r^{N-1}}} \\ &= \frac{1+\frac{1}{r}(\frac{1-\frac{1}{r^{i-1}}}{1-\frac{1}{r}})}{1+\frac{1}{r}(\frac{1-\frac{1}{r^{N-1}}}{1-\frac{1}{r}})} \\ &= \frac{r-\frac{1}{r^{i-1}}}{r-\frac{1}{r^{N-1}}}. \end{aligned}$$

Dividing both the nominator and denominator by r, we can get the simplified form of x_i

$$x_i = \frac{1 - \frac{1}{r^i}}{1 - \frac{1}{r^N}} \tag{6.4.3}$$

for all $i = 1, 2, ..., N$.

7. Real-World Implications with the models

In previous sections, we have discussed three main biological processes – reproduction, selection, and mutation – and their related models to explain *Evolution Dynamics*. Fitness is often the keyword we discuss throughout the discussion and is one of our models' key components, especially when discussing more complicated cases. From evolutionary biologists' perspectives, fitness is equivalent to reproductive success, revealing how a population adapts to the surrounding environment it lives in. How can we compute the fitness, often denoted by r, in real life, so we can put in the numerical numbers and compute the results of the models instead of giving the general solutions? In the following subsections, we will introduce two methods to compute fitness using linear regression, each having unique advantages over the other.

7.1. **Gradient descent algorithm [Ng12].** The first method is gradient descent, an optimization algorithm commonly used to train machine learning models. By continuously adjusting the parameters and computing the cost function, the algorithm could eventually identify a best-fit model that helps explain our training dataset and suggest new output, or target, variables that are not contained in the training dataset. With the trained function model, we can use the input variables to estimate the output. Mathematically, we want to find the following training function that best explains current data and predicts implicit outputs that are not contained by the given dataset.

$$f_{w,b} = wx^{(i)} + b \tag{7.1.1}$$

where w is the weight of $x^{(i)}$, $x^{(i)}$ is the i th training input, and b is a constant of the function when $x = 0$, all of which are parameters to be trained.

We need to have $\hat{y^{(i)}} = f_{w,b}$ such that $\hat{y^{(i)}}$ is close to $y^{(i)}$ for all training dataset $(x^{(i)}, y^{(i)})$. To compute such a $\hat{y^{(i)}}$, let us first compute the cost function for the given training dataset, which tells the numerical error between the predicted outputs and the acutual outputs. The cost function is defined as

$$J_{(w,b)} = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y^{(i)}} - y^{(i)})^2 \tag{7.1.2}$$

where w and b are parameters and m is the total number of the traning data $(x^{(i)}, y^{(i)})$.

The purpose of dividing $\sum_{i=1}^{m} (\hat{y^{(i)}} - y^{(i)})^2$ by $2m$ instead of m is to make data calculation neater. After calculating the partial derivative of the cost function in the next steps, $1/2$ will get canceled out by the power 2, which we will show in the latter part of the paper.

We can use a $3D$ plot to visualize the relationship among $J_{(w,b)} = \hat{y^{(i)}}$, w, and b. The actual plot may vary, but here is an example from Andrew Ng's newest machine learning course on Coursera.

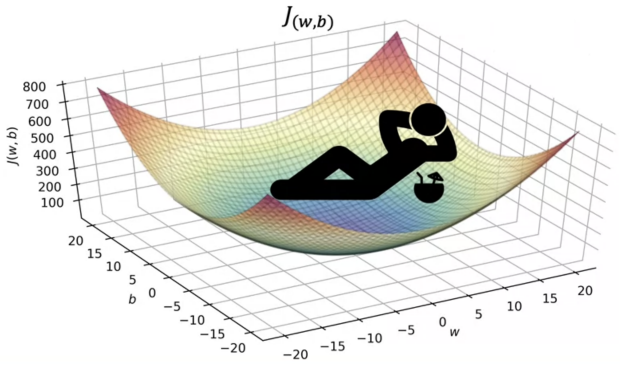

Figure 5. $3D$ plot for $J_{(w,b)}$, w, and b [Ng22]

No matter at which point we start, depending on the values of w and b, and the corresponding $J_{(w,b)}$, our goal is to find the values of w and b such that the numerical value of $J_{(w,b)}$ is a local minimum. To do that, we will use the following two equations, which are main parts of the gradient descent algorithm:

$$\begin{aligned} w &= w - \alpha \frac{\partial}{\partial w} J_{(w,b)} \\ b &= b - \alpha \frac{\partial}{\partial b} J_{(w,b)} \end{aligned} \tag{7.1.3}$$

where α is called the learning rate, which we will explain later, and w, b are parameters for the training model and the cost function.

The partial derivative of the cost function $J_{(w,b)}$, with respect to w can be computed as following:

$$\begin{aligned} \frac{\partial}{\partial w} J_{(w,b)} &= \frac{\partial}{\partial w} \frac{1}{2m} \sum_{i=1}^{m} (\hat{y^{(i)}} - y^{(i)})^2 \\ &= \frac{\partial}{\partial w} \frac{1}{2m} \sum_{i=1}^{m} (wx^{(i)} + b - y^{(i)})^2 \\ &= \frac{1}{2m} \sum_{i=1}^{m} (wx^{(i)} + b - y^{(i)}) 2x^{(i)} \\ &= \frac{1}{m} \sum_{i=1}^{m} (wx^{(i)} + b - y^{(i)}) x^{(i)} \\ &= \frac{1}{m} \sum_{i=1}^{m} (f_{w,b}(x^{(i)}) - y^{(i)}) x^{(i)}. \end{aligned}$$

Similarly, we can compute the result of partial derivative of the cost function $J_{(w,b)}$, with respect to b

$$\begin{aligned}\frac{\partial}{\partial b} J_{(w,b)} &= \frac{\partial}{\partial b} \frac{1}{2m} \sum_{i=1}^{m} (\hat{y^{(i)}} - y^{(i)})^2 \\ &= \frac{1}{m} \sum_{i=1}^{m} (wx^{(i)} + b - y^{(i)}) \\ &= \frac{1}{m} \sum_{i=1}^{m} (f_{w,b}(x^{(i)}) - y^{(i)}).\end{aligned}$$

Each of the above results explains why 7.1.2 is expressed how it is to clean the computation.

By repeating the algorithm until w and b both converge, we can identify the values of w and b such that the corresponding $J_{(w,b)}$ is a local minimum.

Let us go back and explain the learning rate α. Well, the learning rate regulates the step length of adjusting parameters. It is important to choose a sound α. If α is too big, the parameters will never converge. If α is too small, our parameter will update very slowly, and it may take long for parameters to converge. Though the learning rate varies for different models, a sound starting learning rate could be 0.1. We can adjust the learning rate according to the time taken for parameters to converge.

To compute the fitness of one specific population, we would have $y^{(i)}$ as fitness and $x^{(i)}$ as an input feature that affects the fitness. However, many factors could together determine the fitness of one population type — for example, intelligence, size, strength, etc. Also, we need the actual training data that we can rely upon to train our model. As a result, the model discussed above does not allow us to compute fitness in real life. The data could be collected by scientific methods such as natural observation and laboratory experiments. Though it may take much time to get the dataset, the advancement in technology allows for a faster way of data collection. For the model problem, we can use vectorization to put all different input features x together and compute the weight w for each of them, which is represented by the equation below:

$$f_{\vec{w},b}(\vec{x}) = \vec{w} * \vec{x} + b \tag{7.1.4}$$

where $\vec{w} = \begin{bmatrix} w_1 & w_2 & w_3 & ... & w_n \end{bmatrix}$, $\vec{x} = \begin{bmatrix} x_1 & x_2 & x_3 & ... & x_n \end{bmatrix}$, b is the parameter defined as before, and n be the number of factors. The gradient descent algorithm becomes as following

$$\begin{aligned} w_1 &= w_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)}) x_1^{(i)} \\ w_2 &= w_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)}) x_2^{(i)} \\ &\vdots \\ w_n &= w_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)}) x_n^{(i)} \\ b &= b - \alpha \frac{1}{m} \sum_{i=1}^{m} (f_{w,b}(x^{(i)}) - y^{(i)}) \end{aligned} \tag{7.1.5}$$

Again, we need to choose a suitable learning rate α to ensure that the parameters eventually converge while the time taken is reasonable. Upon the convergence of each parameter, we can get our training model that best explains the relationship between factors that affect the fitness of a population and the resulting fitness and predicts the fitness that is not contained in our dataset.

7.2. **Ordinary Least Squares technique [LRW78].** Indeed, the gradient descent algorithm is a powerful method that we can use to find the training model that best fits the given dataset. We can use the algorithm to find the approximate outcomes of the model with different inputs. The choice of learning rate is a key to the accuracy of the training model, as it will determine whether the parameters will converge and the time for convergence. However, it is hard for us to find a reasonable learning rate quickly – the computation process may take a very long time to finish if the learning rate is too low, or the parameters may diverge if the learning rate is too large. Although the gradient descent algorithm can be used broadly for any optimization problems, the Ordinary Least Squares (OLS) method is more specific. It will directly produce a closed-form solution, which is, in most cases, more accurate than the gradient descent algorithm. Suppose the linear combination of input could express the output variable features $\boldsymbol{X}$ plus a residual e; the following equation is our target model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{7.2.1}$$

where $\boldsymbol{X}$ is a $n \times r$ matrix representing n input features, $\boldsymbol{y}$ is a $n \times 1$ column matrix representing n estimated outcomes, and $\hat{\boldsymbol{\beta}}$ is a $r \times 1$ column matrix representing the weight of each input feature.

The OSL technique can help us compute the vector $\hat{\boldsymbol{\beta}}$ that represents the weight of each input feature $\boldsymbol{X}$ by the following equation:

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{7.2.2}$$

which implied the dimensions of each matrix mentioned above by the definition of matrix multiplication and transposition.

Proof. Let us construct a linear equation representing the relationship between $\boldsymbol{y}$ and $\boldsymbol{X}$:

$$\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\epsilon}. \tag{7.2.3}$$

Therefore, we have

$$\boldsymbol{\epsilon} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

where $\boldsymbol{\epsilon}$ is the prediction error that tells the difference between the predicted output value and the actual output value.

Since the sum of squared errors is equal to the product of $\boldsymbol{\epsilon}$ and its transpose [Sta20], we then have the sum of error expressed as following computation

$$\begin{aligned}
\boldsymbol{\epsilon}^{\boldsymbol{T}}\boldsymbol{\epsilon} &= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\
&= (\boldsymbol{y}^T - (\boldsymbol{X}\hat{\boldsymbol{\beta}})^T)(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\
&= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T(\boldsymbol{X}\hat{\boldsymbol{\beta}}) - (\boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{y} + (\boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{X}\hat{\boldsymbol{\beta}} \\
&= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T(\boldsymbol{X}\hat{\boldsymbol{\beta}}) - (\boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{y} + \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}.
\end{aligned}$$

Because $(\boldsymbol{y}^T(\boldsymbol{X}\hat{\boldsymbol{\beta}}))^T = (\boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{y}$, we can conclude that $\boldsymbol{y}^T(\boldsymbol{X}\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{y}$. Therefore, we can simplify the equation:

$$\boldsymbol{\epsilon}^{\boldsymbol{T}}\boldsymbol{\epsilon} = \boldsymbol{y}^T\boldsymbol{y} - 2(\boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{y} + \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\hat{\boldsymbol{\beta}}\boldsymbol{X}. \tag{7.2.4}$$

To find the minimum sum of errors, we can take the partial derivative of $\boldsymbol{\epsilon}^{\boldsymbol{T}}\boldsymbol{\epsilon}$ with respect of $\hat{\boldsymbol{\beta}}$, and set it to zero

$$\frac{\partial \boldsymbol{\epsilon}^{\boldsymbol{T}}\boldsymbol{\epsilon}}{\partial \hat{\boldsymbol{\beta}}} = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = 0. \tag{7.2.5}$$

We can further simplify the equation to

$$\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{y}. \tag{7.2.6}$$

By left multiplying $(\boldsymbol{X}^T\boldsymbol{X}^{-1})$ to both sides, we can get

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{7.2.7}$$

which is the expression we wanted. ■

Although the OSL technique helps us identify the closed-form solution of linear regression, it is very sensitive to outliers. The presence of outliers will significantly affect our model. Therefore, doing an outlier check before all the math work is very important to ensure the model's accuracy. For the time-efficiency of calculation, it is better to filter important features and avoid redundant ones. And most importantly, only linear relationships could be suggested by the OLS technique, whereas we can find more complicated models with the gradient descent algorithm.

Acknowledgements

I would like to thank Simon Rubinstein-Salzedo for inviting knowledgeable guest speakers to give inspirational mathematics talks and teaching me essential skills to write expository math papers and do professional math presentations. I enjoyed working on the interesting research topic "*Evolutionary Dynamics*" that Simon suggested; his advice has brought more depth to my paper.

I would like to thank Alex DeWeese, my TA for the Euler Circle course, for guiding me through the entire research and problem-solving process, addressing my questions about writing the paper, and patiently answering my questions about machine learning and mathematics in general. It is exciting to constantly learn new things through discussing my research project with Alex. Without his support, I would not have the knowledge and ability to explain the Ordinary Least Squares (OLS) technique I described in my paper.

I would also like to thank Kevin Shi, Thomas Catalan, Varun Rao, and Yuxuan Chen, who were also in Alex's group, for listening to my progress and sharing their exciting research topics with me during discussion sessions.

References

[Jet15] James H Jett. How long does it take a cell to divide? *Cytometry Part A*, 87(5):383–384, 2015. URL: `https://onlinelibrary.wiley.com/doi/pdf/10.1002/cyto.a.22665#:~:text=Some%20bacterial%20cells%20can%20divide,in%20a%20culture%20versus%20time`. Last visited on 2022/06/19.

[LRW78] Tze Leung Lai, Herbert Robbins, and Ching Zong Wei. Strong consistency of least squares estimates in multiple regression. *Proceedings of the national academy of sciences*, 75(7):3034–3036, 1978. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC392707/pdf/pnas00019-0030.pdf. Last visited on 2022/07/03.

[Mat03] Charles Matthews. Simplicial complex, June 2003. URL: https://en.wikipedia.org/wiki/Simplicial_complex#:~:text=In%20mathematics%2C%20a%20simplicial%20complex,in%20modern%20simplicial%20homotopy%20theory. Last visited on 2022/06/27.

[Ng12] Andrew Ng. Stanford University CS 229, Lecture Notes: Machine Learning, 2012. URL: https://cs229.stanford.edu/lectures-spring2022/main_notes.pdf. Last visited on 2022/07/03.

[Ng22] Andrew Ng. Coursera "Supervised Machine Learning: Regression and Classification", Lecture Slide: Week 1 Visualizing the cost function, 2022. URL: https://www.coursera.org/learn/machine-learning. Last visited on 2022/07/01.

[Now06] Martin A. Nowak. *Evolutionary Dynamics: Exploring the equations of life*. Harvard University Press, 2006.

[Sta20] Dustin Stansbury. Derivation: Ordinary least squares solution and the normal equations, Jul 2020. URL: https://dustinstansbury.github.io/theclevermachine/derivation-normal-equations. Last visited on 2022/07/03.