A New Metric for Parental Selection in Plant Breeding

Ye Han
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Agricultural Science Commons, Agriculture Commons, Agronomy and Crop Sciences Commons, and the Operational Research Commons

Recommended Citation
Han, Ye, "A New Metric for Parental Selection in Plant Breeding" (2014). Graduate Theses and Dissertations. 13784. https://lib.dr.iastate.edu/etd/13784

This Thesis is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
A new metric for parental selection in plant breeding

by

Ye Han

A thesis submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Industrial Engineering

Program of Study Committee:
Lizhi Wang, Major Professor
William D Beavis
Sarah M Ryan

Iowa State University
Ames, Iowa
2014
Copyright © Ye Han, 2014. All rights reserved.
DEDICATION

I would like to dedicate this thesis to my parents and girlfriend without whose support I would not have been able to complete this work. I would also like to thank my friends and family for their loving guidance during the writing of this work.
# TABLE OF CONTENTS

**LIST OF TABLES** ......................................................... v

**LIST OF FIGURES** ....................................................... vi

**ACKNOWLEDGEMENTS** ................................................ vii

**ABSTRACT** ................................................................. viii

**CHAPTER 1. OVERVIEW** ................................................ 1

1.1 Introduction to Plant Breeding .................................... 1

1.2 Introduction to Markov Decision Processes .................... 3

1.3 Objective of Research .............................................. 3

1.4 Thesis Organization ................................................. 4

**CHAPTER 2. REVIEW OF LITERATURE** ............................. 5

**CHAPTER 3. PARENTAL BREEDING VALUE** ....................... 8

3.1 Introduction .......................................................... 8

3.2 Definitions and Problem Statement ............................... 8

3.2.1 Definitions ...................................................... 9

3.2.2 Problem Statement ............................................ 12

3.3 Genetic Laws and Assumptions .................................. 12

3.3.1 Genetic Laws .................................................. 12

3.3.2 Assumptions .................................................. 13

3.4 Parental Breeding Value Proposition and the Water Pipe Model .... 13

3.4.1 Parental Breeding Value for One Individual ............... 13

3.4.2 Parental Breeding Value for Two Individuals ............. 18
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 4.1</td>
<td>Probability for All Possible Outcomes from Self Pollination</td>
<td>36</td>
</tr>
<tr>
<td>Table 4.2</td>
<td>Action Space for Self Pollination</td>
<td>37</td>
</tr>
<tr>
<td>Table 4.3</td>
<td>Value for Self Crossing ($)</td>
<td>37</td>
</tr>
<tr>
<td>Table 4.4</td>
<td>Transition Probability ( p(S_j</td>
<td>S_i, a_{n2}(S_i)) ) for the Second Generation ( (\times 10^{-2}) )</td>
</tr>
<tr>
<td>Table 4.5</td>
<td>Transition Probability ( p(S_j</td>
<td>S_i, a_{n3}(S_i)) ) for the Third Generation ( (\times 10^{-2}) )</td>
</tr>
<tr>
<td>Table 4.6</td>
<td>Probability for All Possible Outcomes from Back Crossing</td>
<td>39</td>
</tr>
<tr>
<td>Table 4.7</td>
<td>Action Space for Back Crossing</td>
<td>42</td>
</tr>
<tr>
<td>Table 4.8</td>
<td>Value for Back Crossing ($)</td>
<td>42</td>
</tr>
<tr>
<td>Table 4.9</td>
<td>Transition Probability ( p(S_j</td>
<td>S_i, a_{n2}(S_i)) ) for the Second Generation ( (\times 10^{-2}) )</td>
</tr>
<tr>
<td>Table 4.10</td>
<td>Transition Probability ( p(S_j</td>
<td>S_i, a_{n3}(S_i)) ) for the Third Generation ( (\times 10^{-2}) )</td>
</tr>
<tr>
<td>Table 4.11</td>
<td>Action Space for Donor Crossing</td>
<td>45</td>
</tr>
<tr>
<td>Table 4.12</td>
<td>Value for Donor Crossing ($)</td>
<td>46</td>
</tr>
<tr>
<td>Table 4.13</td>
<td>Transition Probability ( p(S_j</td>
<td>S_i, a_{n2}(S_i)) ) for the Second Generation ( (\times 10^{-2}) )</td>
</tr>
<tr>
<td>Table 4.14</td>
<td>Transition Probability ( p(S_j</td>
<td>S_i, a_{n3}(S_i)) ) for the Third Generation ( (\times 10^{-2}) )</td>
</tr>
<tr>
<td>Table 5.1</td>
<td>Simulation Performance</td>
<td>51</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.1</td>
<td>Illustration of Plant Breeding</td>
<td>2</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>Illustration of the Water Pipe Model for $X_1$</td>
<td>16</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>Illustration of the Water Pipe Model for $X_2$</td>
<td>17</td>
</tr>
<tr>
<td>Figure 3.3</td>
<td>Illustration of the Water Pipe Model for $X^3$</td>
<td>20</td>
</tr>
<tr>
<td>Figure 3.4</td>
<td>Illustration of the Water Pipe Model for $X^4$</td>
<td>22</td>
</tr>
<tr>
<td>Figure 3.5</td>
<td>Crossing Scheme</td>
<td>23</td>
</tr>
<tr>
<td>Figure 3.6</td>
<td>Illustration of the Water Pipe Model for $X^5$</td>
<td>25</td>
</tr>
<tr>
<td>Figure 3.7</td>
<td>Illustration of the Water Pipe Model for $X^6$</td>
<td>27</td>
</tr>
<tr>
<td>Figure 4.1</td>
<td>Relation Between Number of Progeny and Success Rate</td>
<td>33</td>
</tr>
<tr>
<td>Figure 4.2</td>
<td>Transition between States in Self Crossing</td>
<td>36</td>
</tr>
<tr>
<td>Figure 4.3</td>
<td>Transition between States in Back Crossing</td>
<td>40</td>
</tr>
<tr>
<td>Figure 4.4</td>
<td>Transition between States in Donor Crossing</td>
<td>44</td>
</tr>
<tr>
<td>Figure 4.5</td>
<td>Comparison for Two Cases</td>
<td>47</td>
</tr>
<tr>
<td>Figure 5.2</td>
<td>Elite Parent</td>
<td>50</td>
</tr>
<tr>
<td>Figure 5.3</td>
<td>Donor Parent</td>
<td>50</td>
</tr>
<tr>
<td>Figure 5.4</td>
<td>$F_1$ Offspring</td>
<td>50</td>
</tr>
<tr>
<td>Figure 5.5</td>
<td>A Random Simulation</td>
<td>53</td>
</tr>
<tr>
<td>Figure 5.6</td>
<td>Simulation Result</td>
<td>54</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and writing this thesis. First and foremost, I would like to thank Dr. Lizhi Wang for his guidance, patience and support throughout this research and the writing of this thesis. His insights and words of encouragement have often inspired me and renewed my hopes for completing my graduate education. I would also like to thank my committee members for their efforts and contributions to this work: Dr. William Beavis and Dr. Sarah Ryan. I would additionally like to thank the company Syngenta AG for their support for my graduate study. In addition, I also would like to thank my friends and colleagues for their help and encouragement throughout my graduate study.
ABSTRACT

Plant breeding is the art of genetic improvement through creation and selection for novel characteristics in plants. Parental selection provides the raw materials for creating each new generation of genetic improvements. Ultimately, the probability of successfully meeting the breeding objectives depends on selection of parents for intermating. With the application of operations research, we developed a new metric, parental breeding value (PBV), based on application of conditional probability distribution to help to solve the parental selection problem to accelerate the process of plant breeding and save resources at the same time. The water pipe model is provided in the thesis to calculate PBV efficiently. Also, we discuss the potential of Markov Decision Processes to address the challenge. For small scale cases, the MDP based approach will lead to a precise result with better performance. However, when dealing with large scale problem, it will suffer calculation complexity and can hardly be applied to realistic problems. In order to fix this, we try simulations based on PBV which can overcome the limitation of the MDP approach. From the results of simulations, the PBV metric is demonstrated to shorten the plant breeding process and decrease the resources costs. We can conclude that the PBV will contribute to deriving better strategies for realistic plant breeding objectives.
CHAPTER 1. OVERVIEW

This chapter will introduce the background knowledge from genetics to operations research for the thesis.

1.1 Introduction to Plant Breeding

Since 1860s when Gregor Mendel came up with the hypothesis about passing traits from parents to offsprings, scientists have paid more and more attention to genetics.

Genetics is a discipline of biology that mainly focuses on genes and heredity. Its essence is the study about the biological information. Through the research from single cell to organisms of animal or plant, we will be able to know better about the living creatures around the world and dealing with the problems of reproducing or adapting to the environments.

Plant breeding is a subtopic of genetics. It combines genetic information with mathematical or industrial methods to improve characteristics of plants. An illustration of a whole process of plant breeding is in Figure 1.1. It describes the general work flow of a plant breeding process. In the figure, every vertical bar represents a single plant. The bars with triangular indicators represent the selected plants for crossing to produce the next generation. In a plant breeding process, we shall choose one or two or multiple plants from a population to cross in each generation. The outcome of the last generation $t$ from a plant breeding process will be more desirable compared with genotypes in the first generation.
In the process of plants’ reproducing, every parental plant will produce gametes to combine to generate the next filial generation. During the process of producing gametes, recombination will occur by chance between alleles. Allele is the unit carrying genetic information. In general, the greater distance between alleles are, the larger probability of recombination they will have. Hence, each parental plant is able to produce many distevital gametes that contain different sets of genetic information. If the parental plants are more promising, their filial generations will have higher potential to inherit their advantages. From this prospective, parental selection will play a crucial role in plant breeding.

In order to be more likely to obtain the desirable characteristics, plant breeders have designed a series of crossing strategies. Among these strategies, with the help of molecular genetic
markers to detect desirable alleles, back crossing strategy combined with self crossing is widely used. This strategy aims to cross a donor plant with elite plants. Although this strategy can preserve most of the elite background alleles, it may still waste some resources or miss valuable genetic information, which means there exists better strategies.

1.2 Introduction to Markov Decision Processes

Markov Decision Processes (MDP) is a sequential decision process through which the decisions produce a sequence of Markov chains with reward [11]. It can provide a method for modeling a dynamic stochastic problem. The outcomes of MDP are partly random and partly controlled.

An MDP model has four major components including state space, action space, transition probability and reward. At a certain time and state, if the decision maker choose an action, there will be a corresponding transition probability. This action will bring in a reward. Policy is a combination of actions and states which means that policy is made of a series of actions taken in every state. The outcome of an MDP model will be the optimal policy for each state.

Value iteration is a widely used algorithm to solve MDP problem [9]. It is also known as backward induction. The basic idea of value iteration is to find the best action for each state in each time period to maximize the value of that time period. Then, the whole series of the actions for the whole state space will form the optimal policy. An MDP model solved by value iteration for parental selection problem will be discussed in details in the following of the thesis.

1.3 Objective of Research

As mentioned before, with the assistance of molecular markers, we are able to detect and locate some specific alleles which shall be transferred into the elite plant. In real-life market, the efficiency and cost of a breeding strategy will play a significant role. The objective of our research is to design a method to select the most promising parental plants to cross to reduce the cost of resources and time (generation number).
1.4 Thesis Organization

The thesis will be organized as follows: Chapter 2 will cover the review of relevant literature about plant breeding and applications from operations research. In chapter 3, we will introduce the new concept of the parental breeding value with the water pipe model for parental selection problem in plant breeding. In Chapter 4 an MDP model for a small scale problem will be presented with analysis. In Chapter 5, we will discuss the simulation based on the PBV and the water pipe model for realistic cases. The result will be discussed as well. Chapter 6 will lead to a summary for this thesis and give some future work direction.
CHAPTER 2. REVIEW OF LITERATURE

What is the reason for similarities and differences in living creatures? The answer is genes or alleles. In every cell of biological organisms, there are chromosomes which are molecular of DNA compound [4]. As the carrier of genetic information, alleles are passed to filial generations from parental generations.

The most famous research on genetics is Mendel’s experiment on garden pea. He spent decades to observe the characteristics of peas and concluded in Mendel’s Laws. Mendel’s first and second law are as follows:

**Mendel’s First Law: Law of Segregation**
The two alleles for each trait separate (segregate) during gamete formation, and then unite at random, one from each parent, at fertilization [4].

**Mendel’s Second Law: Law of Independent Assortment**
During gamete formation, different pairs of alleles segregate independently of each other [4].

Along with the technology developing, biologists have more advanced methods to detect, select and control the transfer of alleles. Plant breeding is such a subject that focuses on crops. A whole successful process of plant breeding needs several steps. In 2003, Peleman and Jeroen Rouppe van der Voort [6] introduced the concept of ‘Breeding by Design’. With the help of marker technology and software tools, researchers are able to control inheritance of allelic variation for all alleles of economic importance. There are three major steps: mapping loci involved in all agronomically relevant traits, assessment of the allelic variation at those loci and breeding by design.

Although the frame of designing a breeding strategy seems to be simple, none of the three steps can be accomplished perfectly yet. For the first two steps, how to find the location of
specific alleles and assess their function are beyond the scope of this thesis. Once desirable functional alleles are discovered, plant breeders need to design a breeding strategy. In 2000, Hospital with Goldringer and Openshaw [3] proposed a selection method called Efficient Marker-Based Recurrent Selection for multiple quantitative trait loci. In 2004, a strategy called Gene Pyramiding was proposed. With the assistance of markers, Servin, et al [10] and other researchers designed a method to combine a series of target alleles identified in different parents into one single genotype. This method provided an algorithm that generates and compares pedigrees on the basis of the population size and total number of generation required. They claimed to find the best gene-pyramiding scheme by proving that their method is more efficient compared relative to the Hospital’s recurrent selection.

In 1999 Frisch, Bohn and Melchinger [1] compared different selection strategies using simulation models for Marker-Assisted Backcrossing of desirable alleles. In 2004, Frisch and Melchinger [2] brought in the selection theory for Marker-Assisted Backcrossing. This MAB method has been adapted by plant breeders in most situations. MAB method will predict the response to selection and give criteria for the selection of the most promising individuals for further back crossing or self crossing.

Although there is extraordinary research on the topic of plant breeding, the challenge for effective marker-assisted selection in plants is still being unsolved. In 2008, Hospital [5] summarized the existing challenges for marker-assisted selection. He not only presented some successful methods but pointed out the challenges and potential future work.

Peng, Sun and Mumm [7, 8] designed a four-step process on converting an elite hybrid for value-added traits using back crossing breeding in 2014. They finished minimizing linkage drag in single event introgression and calculated the process efficiency in event pyramiding and trait fixation. The process was based on computer simulation and proved to be efficient relative to six other strategies.

From the engineering perspective, researchers also investigate methods to improve the efficiency of plant breeding. Xu et al [12] found an optimization approach to gene stacking, which enriched the ways to consider the problem in plant breeding. It provided strategic stacking schemes to maximize the likelihood of successfully creating the target genotypes and minimize
the number of generations. It provided the Pareto optimal solutions under some assumptions, as well.

Thus we propose new ideas for marker-based selection.
CHAPTER 3. PARENTAL BREEDING VALUE

3.1 Introduction

Finding promising parental individuals will help to accomplish the goal of plant breeding quickly and efficiently. However, the existing methods of parental selection do not fully take advantage of the genetic information which means the optimal breeding strategy has not been designed. We propose two challenges for parental selection. The general parental selection challenge is defined as: Given a population, in order to gain all the desirable alleles, how to efficiently select \( k \) individuals as parental individuals to cross in every generation?

To solve the general parental problem above, a fundamental problem of assessment needs to be answered first. This problem is defined as: **How to assess the efficiency of selecting \( k \) parents?** In the thesis, we propose a new metric, the PBV based on conditional probability to solve this fundamental problem. Then, we design the water pipe model to calculate the PBV effectively. The PBV and the water pipe model will give a criterion to assess the efficiency of selected individuals.

3.2 Definitions and Problem Statement

In the later part of the thesis, indicator matrices or vectors will be used to represent individuals and their gametes. At the same time, we will use a recombination frequency vector to represent the recombination frequency between alleles. The indicator matrix for one individual is as follows:
Indicator Matrix for One Individual

\[
\begin{bmatrix}
  x_{1,1} & x_{1,2} \\
  x_{2,1} & x_{2,2} \\
  \vdots & \vdots \\
  x_{n-1,1} & x_{n-1,2} \\
  x_{n,1} & x_{n,2}
\end{bmatrix}
\]

This matrix represents the genetic information of one individual which has \(n\) loci with \(2n\) alleles. Each column of the indicator matrix represents a set of homologous chromosomes in a diploid plant. The \(i^{th}\) row of the matrix represents a pair of alleles at locus \(i\). Every element in the matrix is a binary variable indicating the desirability of the alleles in that locus.

A general definition of the indicator matrix for \(k\) individuals is given in the following part.

### 3.2.1 Definitions

**Definition 1** (Indicator Matrix). Define the indicator matrix \(x(l_1, \ldots, l_k)\) for \(k\) plants \(l_1, l_2, \ldots, l_k\) as follows:

\[
\begin{bmatrix}
  x_{1,1} & x_{1,2} & \cdots & x_{1,2k} \\
  x_{2,1} & x_{2,2} & \cdots & x_{2,2k} \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{n-1,1} & x_{n-1,2} & \cdots & x_{n-1,2k} \\
  x_{n,1} & x_{n,2} & \cdots & x_{n,2k}
\end{bmatrix}
\]

In the matrix \(x\), columns \(2i - 1\) and \(2i\) together represent the \(i^{th}\) individual among all \(k\) plants, \(\forall i \in \{1, \ldots, k\}\).

\(x_{i,j} \in \{0, 1\}\) is a random variable indicating the desirability of the allele in locus \((i, j)\).
**Definition 2** (Indicator Vector). Define the indicator vector \( g \) for a gamete as follows:

\[
\begin{bmatrix}
  g \\
  g_1 \\
  g_2 \\
  \vdots \\
  g_{n-1} \\
  g_n
\end{bmatrix}.
\]

\( g_i \in \{0, 1\} \) is a random variable indicating the desirability of the inherited allele.

**Note:** The vector \( g \) for the gamete is a random function of the indicator matrix \( x \). The distribution for \( g(x) \) will be decided by \( x \).

**Definition 3** (Recombination Frequency). The recombination frequency vector \( r \) is defined as follows:

\[
\begin{bmatrix}
  r \\
  r_1 \\
  r_2 \\
  \vdots \\
  r_{n-1}
\end{bmatrix},
\]

\[
r_i = P(g_{i+1} = x_{i+1,1}|g_i = x_{i,2})
\]

\[
= P(g_{i+1} = x_{i+1,2}|g_i = x_{i,1})
\]

\[
= 1 - P(g_{i+1} = x_{i+1,1}|g_i = x_{i,1})
\]

\[
= 1 - P(g_{i+1} = x_{i+1,2}|g_i = x_{i,2}), \forall i \in \{1, \ldots, n-1\}.
\]

With the previous definitions, we can show the essence of crossing for two individuals. Suppose plant \( x \) will cross with plant \( y \) to produce \( z \) denoted as \( x \times y \Rightarrow z \).
Plant $x$ and $y$ each will produce a set of gametes. Gamete $g^1$ and $g^2$ will combine together to form the filial plant $z$:

\[
\begin{align*}
\begin{bmatrix}
  x_{1,1} & x_{1,2} \\
  x_{2,1} & x_{2,2} \\
  \vdots & \vdots \\
  x_{n-1,1} & x_{n-1,2} \\
  x_{n,1} & x_{n,2}
\end{bmatrix}
\Rightarrow
\begin{bmatrix}
  z_{1,1} \\
  z_{2,1} \\
  \vdots \\
  z_{n-1,1} \\
  z_{n,1}
\end{bmatrix}
\end{align*}
\]

\[
\begin{align*}
\begin{bmatrix}
  y_{1,1} & y_{1,2} \\
  y_{2,1} & y_{2,2} \\
  \vdots & \vdots \\
  y_{n-1,1} & y_{n-1,2} \\
  y_{n,1} & y_{n,2}
\end{bmatrix}
\Rightarrow
\begin{bmatrix}
  z_{1,2} \\
  z_{2,2} \\
  \vdots \\
  z_{n-1,2} \\
  z_{n,2}
\end{bmatrix}
\end{align*}
\]

In process of producing gametes, recombination will occur. $z_{i,1}$ will be from either $x_{i,1}$ or $x_{i,2}$ and $z_{i,2}$ will be from either $y_{i,1}$ or $y_{i,2}$. The recombination will occur according to the recombination frequency. Separately, two gametes $g^1$ and $g^2$ will combine to produce the filial generation $z$:

\[
\begin{align*}
\begin{bmatrix}
  z_{1,1} & z_{1,2} \\
  z_{2,1} & z_{2,2} \\
  \vdots & \vdots \\
  z_{n-1,1} & z_{n-1,2} \\
  z_{n,1} & z_{n,2}
\end{bmatrix}
\Rightarrow
\begin{bmatrix}
  g^1 \oplus g^2
\end{bmatrix}
\end{align*}
\]
3.2.2 Problem Statement

According to plant breeding process with a workflow shown in Figure 1.1, we defined two problems for parental selection.

- **General problem**: Given a population of individual plants, in order to gain all the desirable alleles, how to efficiently select $k$ individuals as parents to cross in each generation?

- **Fundamental problem**: How to assess the efficiency of the $k$ selections?

The general problem is about selecting parents efficiently and the fundamental problem is about assessing the selected parents. In this thesis, we will focus on the fundamental problem to give criteria for assessment. The method to solve the general problem of parental selection problem will be the task for my PhD study.

3.3 Genetic Laws and Assumptions

3.3.1 Genetic Laws

**Axiom 1** (Mendel’s First Law: Law of Segregation). *The two alleles for each trait separate (segregate) during gamete formation, and then unite at random, one from each parent, at fertilization* [4].

**Axiom 2** (Mendel’s Second Law: Law of Independent Assortment). *During gamete formation, different pairs of alleles segregate independently of each other* [4].

**Axiom 3** (Linkage and Linkage disequilibrium). **Linkage**: Linkage is the proximity of two or more markers on a chromosome; the closer together the markers are, the lower the probability that they will be separated by recombination. **Linkage Disequilibrium**: When alleles at separate loci are associated with each other at a significantly higher frequency than would be expected by by independent assortment [4].
3.3.2 Assumptions

Assumption 1: Random Assortment for \( K = 2^k, k \geq 0 \) Individuals

\[ P(g_1 = x_{1,j}) = \frac{1}{2K}, j \in \{1, \ldots, 2K\}. \]

Assumption 2: Independency

Recombination between adjacent loci is independent of recombination among other pair of loci: 

\[ P(g_i = x_{i,j} | g_{i-1}, g_{i-2}, \ldots, g_1) = P(g_i = x_{i,j} | g_{i-1}), \forall i \in \{1, \ldots, n\}, j \in \{1, 2\}. \]

3.4 Parental Breeding Value Proposition and the Water Pipe Model

In the following section, we will give definitions of the PBV for one individual, two individuals and \( 2^k, k \geq 0 \) individuals. A \( w \) matrix is defined to calculate the PBV effectively. Finally, we will give two examples to show how the PBV and \( w \) matrix can be applied to one, two or \( 2^k, k \geq 0 \) individuals. The illustrations following the examples will explain the reason why we name our method as the water pipe model. Proofs of lemmas are in the appendix.

3.4.1 Parental Breeding Value for One Individual

**Definition 4** (Parental Breeding Value for One Individual). *The parental breeding value PB(X) for any individual X is the probability that all alleles in gamete g produced by individual X are desirable.*

\[ PB(X) = P(g(x) = 1_{n \times 1} | x = X). \]
**Definition 5** (the w Matrix for One Individual). *Define the w matrix as:*

\[
\begin{bmatrix}
    w_{1,1} & w_{1,2} \\
    w_{2,1} & w_{2,2} \\
    \vdots & \vdots \\
    w_{n-1,1} & w_{n-1,2} \\
    w_{n,1} & w_{n,2}
\end{bmatrix},
\]

\[w_{i,j} = P(g_i = x_{i,j}, x_{i,j} = 1, g_{i-1} = 1, \ldots, g_1 = 1), \forall i \in \{1, \ldots, n\}, j \in \{1, 2\}.\]

We need to calculate the conditional probability with the help of the w matrix. Before the detailed calculation, we need to bring in two lemmas.

**Lemma 1.**  
\[P(g_i = x_{i,j}, x_{i,j} = 1) = P(g_i = x_{i,j})X_{i,j}, \forall i \in \{1, \ldots, n\}, j \in \{1, 2\}.\]

**Lemma 2.**  
(a). \[w_{1,j} = (1/2)X_{1,j}, j \in \{1, 2\}.\]

(b). \[w_{i,j} = w_{i-1,j}M_iX_{i,j}, \forall i \in \{2, \ldots, n\}, j \in \{1, 2\}.\]

**Note:** For matrix A, let \(A_{i,\bullet}\) denote the \(i\)th row of \(A\), and let \(A_{\bullet,j}\) denote the \(j\)th column of \(A\).

\(M_{s,t}^i\) is the transition probability from \(w_{i-1,s}\) to \(w_{i,t}\), \(\forall s, t \in \{1, 2\}, \forall i \in \{2, \ldots, n\}\)

\[
\begin{bmatrix}
m_{1,1}^i & m_{1,2}^i \\
m_{2,1}^i & m_{2,2}^i
\end{bmatrix} = \begin{bmatrix}
1 - r_i & r_i \\
r_i & 1 - r_i
\end{bmatrix}.
\]

**Theorem 1.** \(PB(X) = P(g(x) = 1_{n \times 1}|x = X) = w_{n,1} + w_{n,2}.\)

**Proof:** Based on **Definition 4, Definition 5, Lemma 1** and **Lemma 2**, we can derive that \(PB(X) = P(g(x) = 1_{n \times 1}|x = X) = w_{n,1} + w_{n,2}.\) \(\square\)

Two examples with illustrations, Figures 3.1 and 3.2 for the parental breeding value for one individual are as follows:
Example 1.

$X^1$ is:

$$
X^1 = \begin{bmatrix}
1 & 1 \\
1 & 1 \\
1 & 1 \\
1 & 1 \\
1 & 1 \\
1 & 1
\end{bmatrix}.
$$

The recombination frequency vector $r$ is:

$$
r = \begin{bmatrix}
0.2 \\
0.1 \\
0.4 \\
0.5
\end{bmatrix}.
$$

$PB(X^1) = w_{5,1} + w_{5,2} = 1$. 
Note: The black blocks can be regarded as water containers to contain the water (probability), the blue blocks represent the pipe transferring a proportion of water (probability) to the next level. Since the model works like pipes, we name our model as the water pipe model. The diameter of the pipes represents the relative frequency of recombination from one locus in a level to the next locus.

Example 2.

$X^2$ is:

\[
X^2 = 
\begin{bmatrix}
1 & 1 \\
0 & 1 \\
1 & 0 \\
0 & 1 \\
1 & 0 \\
\end{bmatrix}
\]

The recombination frequency vector $r$ is:
\[ r \begin{bmatrix} 0.2 \\ 0.1 \\ 0.4 \\ 0.5 \end{bmatrix}. \]

\[ PB(X^2) = w_{5,1} + w_{5,2} = 0.01. \]

Figure 3.2: Illustration of the Water Pipe Model for $X_2$
3.4.2 Parental Breeding Value for Two Individuals

Definition 6. The parental breeding value \( PB(x(l_x, l_y)) \) for any two individuals \( l_x \) and \( l_y \) is the conditional probability that all alleles in the gamete \( g \) produced by the offspring \( l_z \) are desirable.

\[
PB(x(l_x, l_y)) = P(g(x) = 1_{n \times 1} | x = x(l_x, l_y)).
\]

We may want:

\[
\begin{bmatrix}
  x_{1,1} & x_{1,2} \\
  x_{2,1} & x_{2,2} \\
  \vdots & \vdots \\
  x_{n-1,1} & x_{n-1,2} \\
  x_{n,1} & x_{n,2}
\end{bmatrix}
\begin{bmatrix}
  y_{1,1} & y_{1,2} \\
  y_{2,1} & y_{2,2} \\
  \vdots & \vdots \\
  y_{n-1,1} & y_{n-1,2} \\
  y_{n,1} & y_{n,2}
\end{bmatrix}
\Rightarrow
\begin{bmatrix}
  z_{1,1} & z_{1,2} \\
  z_{2,1} & z_{2,2} \\
  \vdots & \vdots \\
  z_{n-1,1} & z_{n-1,2} \\
  z_{n,1} & z_{n,2}
\end{bmatrix}
\Rightarrow
\begin{bmatrix}
  g_1 \\
  g_2 \\
  \vdots \\
  g_{n-1} \\
  g_n
\end{bmatrix}
\]

Definition 7 (the \( w \) Matrix for Two Individuals). Define the \( w \) matrix as:

\[
\begin{bmatrix}
w_{1,1} & w_{1,2} & w_{1,3} & w_{1,4} \\
w_{2,1} & w_{2,2} & w_{2,3} & w_{2,4} \\
\vdots & \vdots & \vdots & \vdots \\
w_{n-1,1} & w_{n-1,2} & w_{n-1,3} & w_{n-1,4} \\
w_{n,1} & w_{n,2} & w_{n,3} & w_{n,4}
\end{bmatrix},
\]

\[
w_{i,j} = P(g_i = x_{i,j}, x_{i,j} = 1, g_{i-1} = 1, \ldots, g_1 = 1), \forall i \in \{1, ..., n\}, j \in \{1, ..., 4\}.
\]

Lemma 3. (a) \( w_{1,j} = (1/4)X_{1,j}, j \in \{1, ..., 4\} \).

(b) \( w_{i,j} = w_{i-1,j} M^i, \forall i \in \{2, ..., n\}, j \in \{1, ..., 4\} \).
$M^i_{s,t}$ is the transition probability from $w_{i-1,s}$ to $w_{i,t}$, $\forall s, t \in \{1, 2, 3, 4\}, \forall i \in \{2, \ldots, n\}$.

\[
M^i = \begin{bmatrix}
m_{1,1} & m_{1,2} & m_{1,3} & m_{1,4} \\
m_{2,1} & m_{2,2} & m_{2,3} & m_{2,4} \\
m_{3,1} & m_{3,2} & m_{3,3} & m_{3,4} \\
m_{4,1} & m_{4,2} & m_{4,3} & m_{4,4}
\end{bmatrix} = \begin{bmatrix}
(1 - r_i)^2 & r_i(1 - r_i) & 0.5r_i & 0.5r_i \\
 r_i(1 - r_i) & (1 - r_i)^2 & 0.5r_i & 0.5r_i \\
0.5r_i & 0.5r_i & (1 - r_i)^2 & r_i(1 - r_i) \\
0.5r_i & 0.5r_i & r_i(1 - r_i) & (1 - r_i)^2
\end{bmatrix}.
\]

**Theorem 2.** $PB(x(l_x, l_y)) = P(g(x) = 1_{n \times 1}|x = x(l_x, l_y)) = w_{n,1} + w_{n,2} + w_{n,3} + w_{n,4}$.

**Proof:** Based on Definition 6, Definition 7, Lemma 1 and Lemma 3, we can derive that $PB(x(l_x, l_y)) = P(g(x) = 1_{n \times 1}|x = x(l_x, l_y)) = w_{n,1} + w_{n,2} + w_{n,3} + w_{n,4}$. \qed

Two examples with illustrations, Figures 3.3 and 3.4 for the parental breeding value for two individuals are as follows:

**Example 3.**

The indicator matrix $X^3$ for two identical homozygous individuals is:

\[
X^3 = \begin{bmatrix}
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
\end{bmatrix}
\]

The recombination frequency vector $r$ is:
\[
\begin{bmatrix}
0.2 \\
0.1 \\
0.4 \\
0.5
\end{bmatrix}
\]

\[PB(X^3) = w_{5,1} + w_{5,2} + w_{5,3} + w_{5,4} = 1.\]

Figure 3.3: Illustration of the Water Pipe Model for \(X^3\)
Example 4.

The indicator matrix $X^4$ for two heterozygous of heterogeneous individuals is:

$$X^4 = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}.$$ 

The recombination frequency vector $r$ is:

$$r = \begin{bmatrix} 0.2 \\ 0.1 \\ 0.4 \\ 0.5 \end{bmatrix}.$$

$$PB(X^4) = w_{5,1} + w_{5,2} + w_{5,3} + w_{5,4} = 0.1139.$$
3.4.3 Parental Breeding Value for $K = 2^k, k \geq 0$ Individuals

In this section, we will discuss about $K = 2^k, k \geq 0$ individuals crossing cases. In this situation, we will need a minimum of $k + 1$ generations to obtain the desired gamete. The crossing scheme is in the following Figure 3.5.
Definition 8 (Parental Breeding Value for $K = 2^k$, $k \geq 0$ Individuals). The parental breeding value $PB(x(l_1, ..., l_K))$ for $K$ individuals $l_1, ..., l_K$, is the probability that the alleles in the gamete $g$ produced by their offspring are all desirable.

$$PB(x(l_1, ..., l_K)) = P(g_1 = 1, g_2 = 1, ..., g_n = 1 | x = x(l_1, ..., l_K)) = P(g(x) = 1_{n\times1} | x = x(l_1, ..., l_K)).$$

Definition 9 (the $w$ Matrix for $K = 2^k$ Individuals). Define the $w$ matrix as:

$$w = \begin{bmatrix}
    w_{1,1} & w_{1,2} & \cdots & w_{1,2^k} \\
    w_{2,1} & w_{2,2} & \cdots & w_{2,2^k} \\
    \vdots & \vdots & \ddots & \vdots \\
    w_{n-1,1} & w_{n-1,2} & \cdots & w_{n-1,2^k} \\
    w_{n,1} & w_{n,2} & \cdots & w_{n,2^k}
\end{bmatrix},$$

$$w_{i,j} = P(g_i = x_{i,j}, x_{i,j} = 1, g_{i-1} = 1, ..., g_1 = 1), \forall i \in \{1, ..., n\}, j \in \{1, ..., 2^k\}.$$  

Lemma 4. (a). For $K = 2^k$ individuals, $w_{1,j} = (1/2^k)X_{1,j}, j \in \{1, ..., 2^k\}.$

(b). $w_{i,j} = w_{i-1,j}M^i_{s,t} X_{i,j}, \forall i \in \{2, ..., n\}, j \in \{1, ..., 2^k\}.$

$M^i_{s,t}$ is the transition probability from $w_{i-1,s}$ to $w_{i,t}, \forall i \in \{2, ..., n\}, \forall s, t \in \{1, ..., 2^k\}.$
(a). For $2^0$ individual:

$$M^i(0) = \begin{bmatrix} 1 - r_i & r_i \\ r_i & 1 - r_i \end{bmatrix}.$$ 

(b). For $2^{k+1}, k \geq 0$ individuals:

$$M^i(k + 1) = \begin{bmatrix} (1 - r_i)M^i(k) & \left(\frac{1}{2^k}\right)r_i \mathbb{1}_{2^k \times 2^k} \\ \left(\frac{1}{2^k}\right)r_i \mathbb{1}_{2^k \times 2^k} & (1 - r_i)M^i(k) \end{bmatrix}.$$ 

**Theorem 3.** $PB(x(l_1, ..., l_K)) = P(g(x) = P(g_n = 1, ..., g_1 = 1|x = x(l_1, ..., l_K)) = \sum_{j=1}^{2^K} w_{n,j}.$

**Proof:** Based on **Definition 8, Definition 9, Lemma 1 and Lemma 4**, we can derive that

$$PB(x(l_1, ..., l_K)) = P(g(x) = P(g_n = 1, ..., g_1 = 1|x = x(l_1, ..., l_K)) = \sum_{j=1}^{2^K} w_{n,j}.$$ 

Two examples with illustrations, Figures 3.6 and 3.7 for parental breeding value for four individual are as follows:

**Example 5.**

The indicator matrix $X^5$ for four individuals is:

$$X^5 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$ 

The recombination frequency vector $r$ is:

$$r = \begin{bmatrix} 0.2 \\ 0.1 \\ 0.4 \\ 0.5 \end{bmatrix}.$$
$PB(X^5) = \sum_{i=j}^{8} w_{i,j} = 1.$

Figure 3.6: Illustration of the Water Pipe Model for $X^5$. 
Example 6.

The indicator matrix $X^6$ for four heterozygous of heterogeneous individuals is:

$$X^6 = \begin{bmatrix}
1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\
1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 \\
1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 \\
0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\
1 & 1 & 1 & 0 & 1 & 0 & 0 & 1
\end{bmatrix}.$$  

The recombination frequency vector $r$ is:

$$r = \begin{bmatrix}
0.2 \\
0.1 \\
0.4 \\
0.5
\end{bmatrix}.$$  

$$PB(X^6) = \sum_{j=1}^{8} w_{5,j} = 0.0637.$$
Figure 3.7: Illustration of the Water Pipe Model for $X^6$
CHAPTER 4. A MARKOV DECISION PROCESSES APPROACH

The parental selection problem can be regarded as a dynamic decision problem. The outcomes of this problem are under uncertainty and we need to make an optimal decision for each possible outcome. Markov Decision Processes (MDP) is such a method to deal with the dynamic decision problem. Thus, we will apply the parental breeding value to MDP and design an MDP model to solve the parental selection problem in this chapter.

4.1 Introduction

In a strategy for plant breeding, we assume that the resources cost will be relevant to the number of progeny produced in each mating and the time cost will be relevant to the generation number. Thus, a better strategy shall need fewer progeny and fewer generations.

Markov Decision Processes (MDP) is a sequential decision process for which the decisions produce a sequence of Markov chains with reward [11]. It can provide a method for modeling a dynamic stochastic problem whose outcomes are partly random and partly controlled.

An MDP model has four major components including state space, action space, transition probability and reward. At an epoch and state, if the decision maker choose a certain action, there will be a probability to transfer to another certain state. This action will bring in a reward.

In the parent selection problem for plant breeding, each possible outcome from a crossing will define a state. Besides, the transition probability will come from the recombination between each allele. The action taken in each state is deciding the number of progeny to produce. The policy will be a combination of states and actions and the optimal policy is the most valuable one for a certain state at a certain epoch.
Value iteration is a widely used method to solve MDP problem [9]. It is also known as backward induction. Its basic idea is to find the best action for each state in each time period to maximize the value in that time period. Then, the whole series of the actions will be the optimal policy. We will apply this method to solve our MDP model.

4.2 Model

This section will cover the MDP model for parental selection problem. We will provide the fundamental definitions for the MDP model. The formulations for the model will be shown in details following the definitions.

4.2.1 Definitions

Definition 10 (Decision Epoch t). Divide time into N periods, the beginning of each period is a decision epoch. \( t = \{1, 2, ..., N\} \), \( t = N + 1 \) is the end of breeding process. 
\( N \) is the number of generations we can afford.

Note: Since the time periods are finite, our model has a finite horizon.

Definition 11 (State and State Space). Each mating has \( m \) possible genotypes as outcomes. If the best genotype among all outcomes is \( i \) in current, then we are in state \( S_i \). The best genotype is the genotype with the largest parental breeding value.

The state space \( S \) contains all the possible states in current. \( S_i \in S, i \in \{1, ..., m\} \)

Definition 12 (Action and Action Space). At each epoch \( t, t \in \{1, ..., N\} \), and state \( S_i, i \in \{1, ..., m\} \) the action \( a_t(S_i) \) is to decide the number of progeny to produce for a crossing.

The action space \( A \) contains all the possible actions.

Definition 13 (Transition Probability). \( p(S_j|S_i, a_t(S_i)) \) is the transition probability from state \( S_i, i \in \{1, ..., m\} \) at epoch \( t, t \in \{1, ..., N\} \) to \( S_j \) at epoch \( t + 1 \), under action \( a_t \).

Note: \( \sum_{j=1}^{m} p(S_j|S_i, a_t(S_i)) = 1, \forall i, j \in \{1, ..., m\}, t \in \{1, ..., N\} \).
Definition 14 (Reward). $r_t(S_i, a_t(S_i))$ is the reward the decision maker will receive in state $S_i, i \in \{1, ..., m\}$ at epoch $t, t \in \{1, ..., N+1\}$ by choosing action $a_t(S_i)$.

Definition 15 (Value). $v_i(t)$ is the value for state $S_i, i \in \{1, ..., m\}$ at epoch $t, t \in \{1, ..., N+1\}$. $r_t(S_i, a_t(S_i))$ is the reward received from the action $a_t(S_i)$ taken in state $S_i$ at epoch $t$. 

$v_i(t) = r_t(S_i, a_t(S_i)) + \sum_{j=1}^{m} p(S_j|S_i, a_t(S_i))v_j(t+1), \forall i, j \in \{1, ..., m\}, t \in \{1, ..., N\}$.

Definition 16 (Discount Factor $\alpha$). The value $v(t+1)$ is worth $\alpha v(t+1)$ at epoch $t, t \in \{1, ..., N\}$.

Definition 17 (Policy). $p(t) = \begin{bmatrix} a_t(S_1) \\ a_t(S_2) \\ \vdots \\ a_t(S_m) \end{bmatrix}, t \in \{1, ..., N\}$ is the policy for epoch $t$. It describe the action shall be taken in different states at epoch $t$.

Definition 18 (Success Rate). Define the success rate as $P^n_{X,Y}$. It is the probability that at least one of the $n$ filial progeny produced by the crossing of individual $l_X$ with individual $l_Y$ contains all desirable alleles. $P^n_{X,Y} = 1 - [1 - L(l_X) L(l_Y)]^n$.

4.2.2 Formulation

We will establish our MDP model in this section. Based on value iteration, at each epoch, we will try to find the optimal policy to gain as much value as possible.

We assume that we need to finish the plant breeding process within $N+1$ generations, then the formulation for the model is:

$$v_i(t) = \max_{a_t(S_i)} \left[ r_t(S_i, a_t(S_i)) + \sum_{j=1}^{m} \alpha p(S_j|S_i, a_t(S_i))v_j(t+1) \right]$$

$t = \{1, 2, ..., N\}$

$i = \{1, 2, ..., m\}$

$j = \{1, 2, ..., m\}$
4.3 Case Study

Parameters:

We set individual $l_A$ as a "donor" and individual $l_B$ as the elite individual as follows:

$$
\begin{bmatrix}
0 & 0 \\
1 & 1 \\
0 & 0
\end{bmatrix}
\quad
\begin{bmatrix}
1 & 1 \\
0 & 0 \\
1 & 1
\end{bmatrix}
$$

The recombination frequency is $r$ as follows:

$$
\begin{bmatrix}
0.05 \\
0.1
\end{bmatrix}
$$

We assume that the value of an offspring with all alleles being desirable is $10^6$, and the discount factor is 0.99, a cost for producing one progeny is 7$ in every generation.

Objective:

We would like to transfer the desirable allele in donor $l_A$ to elite individual $l_B$.

Cases:

- **Case 1.** The plant breeding process shall be finished in 2 generations. Otherwise, it fails.

- **Case 2.** The plant breeding process shall be finished in 3 generations. Otherwise, it fails.

Back crossing strategy and self crossing strategy are two possible and common breeding methods.

Definition 19 (Back Crossing). *Cross the given individual $l_X$ with the elite individual $l_E$.*
**Definition 20** (Self Crossing). *Cross the given individual* $l_X$ *with itself, which is also known as self pollination.*

For any case, the crossing for the first generation is fixed. We have to cross $l_A$ with the elite individual $l_B$ to get individual $l_{F_1}$ which is as follows:

$$
\begin{pmatrix}
1 & 0 \\
0 & 1 \\
1 & 0
\end{pmatrix}
$$

As the definition of parental breeding value, $PB(l_{F_1}) = 0.5 \times (0.2 \times 0.1) = 0.01$.

**4.3.1 Case 1: Two Generations**

The crossing scheme for two generations case is as follows:

$$
l_A \times l_B \\
\downarrow \\
l_{F_1} \times l_{F_1} \\
\downarrow \\
l_{F_2}
$$

For the two generations case, if we want to have an individual $l_{F_2}$ with all alleles being desirable, the only crossing strategy for individual $l_{F_1}$ is self pollination. The action is to decide the number of progeny we need to produce for self pollination.

With the given parameters $l_A$, $l_B$ and recombination frequency $r$, the relation between the number of progeny to produce in the second generation and the success rate is illustrated in Figure 4.1. From the Figure 4.1, we can find the relation between number of progeny and success rate obeys an exponential trend. After calculation, in two generations case with self crossing option, $l_{F_1}$ needs to produce at lease 2321 progeny to achieve a 0.997 success rate. In order to achieve a high success rate, this case calls for a large number of progeny to produce, which will increase the resources cost significantly.
Figure 4.1: Relation Between Number of Progeny and Success Rate
4.3.2 Case 2: Three Generations

In this case, we will discuss the performance of the MDP model. Based on the result of two generations case, we set the largest progeny number to produce for each generation is 3000. The first step is the same as two generations case. The crossing scheme is as follows:

\[
\begin{align*}
    l_A & \times l_B \\
    \downarrow & \\
    l_{F_1} & \times l_x = ? \\
    \downarrow & \\
    l_{F_2} & \times l_{F_2} \\
    \downarrow & \\
    l_{F_3} & 
\end{align*}
\]

From the second generation, we have more choices to achieve the objective. We can select the donor \(l_A\), the elite individual \(l_B\) or \(l_{F_1}\) itself as individual \(l_x\) to cross. These are the three options for this parent selection problem. For each circumstance, we will apply our MDP model to draw the optimal policy.

**Choice 1: \(l_x = l_{F_1}\)**

In such situation, we choose \(l_{F_1}\) to do self crossing. There will be 14 possible outcomes. Each outcome \(l_i, i \in \{1, ..., 14\}\) will have its own parental breeding value \(PB(l_i), i \in \{1, ..., 14\}\). Based on the rank of the parental breeding value from high to low, 14 outcomes are shown as follows:

\[
\begin{bmatrix}
    1 & 0 \\
    0 & 1 \\
    1 & 0
\end{bmatrix}
\times
\begin{bmatrix}
    1 & 0 \\
    0 & 1 \\
    1 & 0
\end{bmatrix}
\Rightarrow
\]
\[
\begin{align*}
\begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} & \quad PB(l_1) = 1 \\
\begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} & \quad PB(l_2) = 0.5 \\
\begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} & \quad PB(l_3) = 0.5 \\
\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} & \quad PB(l_4) = 0.5 \\
\begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} & \quad PB(l_5) = 0.475
\end{align*}
\]

\[
\begin{align*}
\begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} & \quad PB(l_6) = 0.45 \\
\begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} & \quad PB(l_7) = 0.43 \\
\begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} & \quad PB(l_8) = 0.4275 \\
\begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} & \quad PB(l_9) = 0.07 \\
\begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} & \quad PB(l_{10}) = 0.05
\end{align*}
\]

\[
\begin{align*}
\begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} & \quad PB(l_{11}) = 0.0475 \\
\begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} & \quad PB(l_{12}) = 0.025 \\
\begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} & \quad PB(l_{13}) = 0.0225 \\
\begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} & \quad PB(l_{14} \text{ or failure}) = 0.0025
\end{align*}
\]

The probability of \( l_{F1} \) crossing with itself to produce each of these 14 outcomes is in Table 4.1.

**Note:** Failure means that the maximum parental breeding value among all filial progeny is 0.

Based on these 14 outcomes, we can derive our state space \( \mathbb{S} = \{S_1, S_2, ..., S_{14}\} \):

\( S_i \): The best progeny we have in current is \( l_i, i \in \{1, ..., 13\} \).

\( S_{14} \): The best progeny we have in current is \( l_{14} \) or failure.
Table 4.1: Probability for All Possible Outcomes from Self Pollination

<table>
<thead>
<tr>
<th>(P(l_1))</th>
<th>(P(l_2))</th>
<th>(P(l_3))</th>
<th>(P(l_4))</th>
<th>(P(l_5))</th>
<th>(P(l_6))</th>
<th>(P(l_7))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1.25 \times 10^{-5})</td>
<td>(1.125 \times 10^{-4})</td>
<td>0.0021</td>
<td>(2.375 \times 10^{-4})</td>
<td>(1.125 \times 10^{-4})</td>
<td>(2.375 \times 10^{-4})</td>
<td>0.0021</td>
</tr>
<tr>
<td>(P(l_8))</td>
<td>(P(l_9))</td>
<td>(P(l_{10}))</td>
<td>(P(l_{11}))</td>
<td>(P(l_{12}))</td>
<td>(P(l_{13}))</td>
<td>(P(l_{14})) or failure</td>
</tr>
<tr>
<td>(1.25 \times 10^{-5})</td>
<td>0.0021</td>
<td>0.0192</td>
<td>0.0010</td>
<td>0.0406</td>
<td>0.0045</td>
<td>0.9275</td>
</tr>
</tbody>
</table>

**Note:** The ‘best progeny’ means the progeny with the largest parental breeding value.

The transition between every state in the second and third generation is in Figure 4.2.

![Figure 4.2: Transition between States in Self Crossing](image)

In the first generation, we cross \(l_A\) and \(l_B\) to produce \(a_{n_1}\) progeny, we will reach state \(S_{14}\) for sure. At the beginning of the second generation, suppose we take action to produce \(n_2\) progeny in state \(S_{14}\), we will reach state \(S_i, i \in \{1, \ldots, 14\}\) in the next generation with probability:

\[
P(S_1|S_{14}, a_{n_2}(S_{14})) = 1 - (1 - P(l_1))^{a_{n_2}(S_{14})}. \tag{4.2}
\]

\[
P(S_i|S_{14}, a_{n_2}(S_{14})) = 1 - \left[ 1 - \sum_{s=1}^{i} P(l_s) \right]^{a_{n_2}(S_{14})} - \sum_{s=1}^{i-1} P(S_s|S_{14}, a_{n_2}(S_{14})). \tag{4.3}
\]

\[i \in \{2, \ldots, 14\}\]

At the beginning of the third generation, suppose we are at state \(S_i, i \in \{2, \ldots, 14\}\), the best progeny we have is \(l_i\). In this situation, our planting strategy will be self crossing. We only need to decide \(n_3\), the number of progeny to produce and will have the following probability to reach \(S_1\) or \(S_{14}\):
\[ P(S_1|S_i, a_{n_3}(S_i)) = 1 - [1 - PB(l_i)^2]^{a_{n_3}(S_i)}, i \in \{2, \ldots, 14\}. \] (4.4)

\[ P(S_{14}|S_i, a_{n_3}(S_i)) = 1 - P(S_1|S_i, a_{n_3}(S_i)), i \in \{2, \ldots, 14\}. \] (4.5)

Combine equation 4.2, equation 4.3, equation 4.4, equation 4.5 with the MDP model 4.1, we will draw the following conclusion for the given parameters:

**Action Space:**

The actions (number of progeny to produce) for each state in the second and third generation shall be taken are in the Table 4.2:

<table>
<thead>
<tr>
<th>Action</th>
<th>State</th>
<th>( a_{n_2}(S_i) )</th>
<th>( a_{n_3}(S_i) )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0 0 726</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0 37 37 37 41 46 50 51 1332 2345 2553 3000 3000 0</td>
<td></td>
</tr>
</tbody>
</table>

**Value:**

The value for each state in the second and third generation are shown in Table 4.3.

<table>
<thead>
<tr>
<th>Value</th>
<th>State</th>
<th>( v(1) )</th>
<th>( v(2) )</th>
<th>( v(3) )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>99000</td>
<td>100000</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>98961 97925 97079 96903 81727 75229 -0.813</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Transition Probability for the Second Generation:**
The following Table 4.4 contains the transition probability from state $S_{14}$ in the first generation to the state $S_i$, $i \in \{1, \ldots, 14\}$ in the second generation when we take the action $a_{n2}(S_i)$ above.

Table 4.4: Transition Probability $p(S_j|S_i, a_{n2}(S_i))$ for the Second Generation ($\times 10^{-2}$)

<table>
<thead>
<tr>
<th>State</th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
<th>$S_5$</th>
<th>$S_6$</th>
<th>$S_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{14}$</td>
<td>0.9</td>
<td>7.77</td>
<td>72.01</td>
<td>3.07</td>
<td>1.28</td>
<td>2.38</td>
<td>9.94</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>State</th>
<th>$S_8$</th>
<th>$S_9$</th>
<th>$S_{10}$</th>
<th>$S_{11}$</th>
<th>$S_{12}$</th>
<th>$S_{13}$</th>
<th>$S_{14}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{14}$</td>
<td>0.02</td>
<td>2.08</td>
<td>0.55</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Transition Probability for the Third Generation:**

The following Table 4.5 contains the transition probability from state $S_i$, $i \in \{1, \ldots, 14\}$ in the second generation to the states $S_1$ and $S_{14}$ in the third generation when we take the action $a_{n2}(S_i)$ above.

Table 4.5: Transition Probability $p(S_j|S_i, a_{n3}(S_i))$ for the Third Generation ($\times 10^{-2}$)

<table>
<thead>
<tr>
<th>State</th>
<th>State</th>
<th>$S_1$</th>
<th>$S_{14}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td></td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>$S_2$</td>
<td></td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>$S_3$</td>
<td></td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>$S_4$</td>
<td></td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>$S_5$</td>
<td></td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>$S_6$</td>
<td></td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>$S_7$</td>
<td></td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>$S_8$</td>
<td></td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>$S_9$</td>
<td></td>
<td>99.86</td>
<td>0.14</td>
</tr>
<tr>
<td>$S_{10}$</td>
<td>99.72</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>$S_{11}$</td>
<td>99.69</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>$S_{12}$</td>
<td>84.67</td>
<td>15.33</td>
<td></td>
</tr>
<tr>
<td>$S_{13}$</td>
<td>78.11</td>
<td>21.89</td>
<td></td>
</tr>
<tr>
<td>$S_{14}$</td>
<td>0</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

**Conclusion:**

According the results above, the expected number of progeny we need in total is about 804
with the success rate = 0.9999. It is very efficient compared with the two generations case.

**Choice 2: \( l_x = l_B \)**

In such situation, we choose the elite individual \( l_B \) to do back crossing. There will be 4 possible outcomes. Each outcome \( l_i \) will have its own parental breeding value \( PB(l_i) \). Based on the rank of the parental breeding value from high to low, these 4 outcomes are shown as follows:

\[
\begin{bmatrix}
1 & 0 \\
0 & 1 \\
1 & 0 \\
\end{bmatrix}
\begin{bmatrix}
1 & 1 \\
0 & 0 \\
1 & 1 \\
\end{bmatrix}
\Rightarrow
\begin{bmatrix}
1 & 1 \\
1 & 0 \\
1 & 1 \\
\end{bmatrix}
PB(l_3) = 0.5
\begin{bmatrix}
1 & 0 \\
1 & 0 \\
0 & 1 \\
\end{bmatrix}
PB(l_{10}) = 0.05
\begin{bmatrix}
0 & 1 \\
1 & 0 \\
1 & 1 \\
\end{bmatrix}
PB(l_{12}) = 0.025
\begin{bmatrix}
0 & 1 \\
1 & 0 \\
0 & 1 \\
\end{bmatrix}
PB(l_{14}) = 0.0025
\]

The probability of \( l_{F_1} \) doing back crossing with the elite individual to produce each of these 4 outcomes is in Table 4.6.

<table>
<thead>
<tr>
<th>( l_3 )</th>
<th>( l_{10} )</th>
<th>( l_{12} )</th>
<th>( l_{14} ) or failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>\begin{bmatrix} 1 &amp; 1 \ 1 &amp; 0 \ 1 &amp; 1 \end{bmatrix}</td>
<td>\begin{bmatrix} 1 &amp; 1 \ 1 &amp; 0 \ 0 &amp; 1 \end{bmatrix}</td>
<td>\begin{bmatrix} 0 &amp; 1 \ 1 &amp; 0 \ 1 &amp; 1 \end{bmatrix}</td>
<td>\begin{bmatrix} 0 &amp; 1 \ 1 &amp; 0 \ 0 &amp; 1 \end{bmatrix}</td>
</tr>
</tbody>
</table>

Table 4.6: Probability for All Possible Outcomes from Back Crossing

<table>
<thead>
<tr>
<th>( P(l_3) )</th>
<th>( P(l_{10}) )</th>
<th>( P(l_{12}) )</th>
<th>( P(l_{14}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0021</td>
<td>0.0192</td>
<td>0.0406</td>
<td>0.9380</td>
</tr>
</tbody>
</table>

Based on these 4 outcomes, we can derive our state space \( S = \{S_3, S_{10}, S_{12}, S_{14}\} \):
$S_i$: The best progeny we have in current is $l_i, i \in \{3, 10, 12\}$.

$S_{14}$: The best progeny we have in current is $l_{14}$ or failure.

The relation between every state in the second and third generation is in Figure 4.3.

Figure 4.3: Transition between States in Back Crossing
In the first generation, we cross $l_A$ and $l_B$ to produce $a_{n_1}$ progeny, we will reach state $S_{14}$ for sure. At the beginning of second generation, suppose we take action to produce $n_2$ progeny in state $S_{14}$, we will reach state $S_i$, $i \in \{3, 10, 12, 14\}$ in the next generation with probability:

$$
P(S_3|S_{14}, a_{n_2}(S_{14})) = 1 - (1 - P(l_3))^{a_{n_2}(S_{14})}.
$$

$$
P(S_i|S_{14}, a_{n_2}(S_{14})) = 1 - \left[1 - \sum_{s=1}^{i} P(l_s)\right]^{a_{n_2}(S_{14})} - \sum_{s=1}^{i-1} P(S_s|S_{14}, a_{n_2}(S_{14})).
$$

$$
i \in \{10, 12, 14\}
$$

At the beginning of the third generation, suppose we are at state $S_i$, $i \in \{3, 10, 12, 14\}$, the best progeny we have is $l_i$. In this situation, our planting strategy will be self crossing for sure. We only need to decide the number of progeny to produce $n_3$ and will have the following probability to reach $S_1$ or $S_{14}$:

$$
P(S_1|S_i, a_{n_3}(S_i)) = 1 - \left[1 - PB(l_i)^2\right]^{a_{n_3}(S_i)}, i \in \{3, 10, 12, 14\}.
$$

$$
P(S_{14}|S_i, a_{n_3}(S_i)) = 1 - P(S_1|S_i, a_{n_3}(S_i)), i \in \{3, 10, 12, 14\}.
$$

Combine equation 4.6, equation 4.7, equation 4.8, equation 4.9 with equation 4.1, we will draw the following conclusion for the given parameters:

**Action Space:**

The actions shall be taken for each state in the second and third generation are in Table 4.7.

**Value:**

The value for each state in the all three generations is shown in Table 4.8.

**Transition Probability for the Second Generation:**

The following Table 4.9 and Table 4.10 are the transition probability for states in the second and third generation when we take the actions above.
Table 4.7: Action Space for Back Crossing

<table>
<thead>
<tr>
<th>Action</th>
<th>State</th>
<th>$S_{14}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_{n_2}(S_i)$</td>
<td>1247</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Action</th>
<th>State</th>
<th>$S_3$</th>
<th>$S_{10}$</th>
<th>$S_{12}$</th>
<th>$S_{14}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_{n_3}(S_i)$</td>
<td>37</td>
<td>2345</td>
<td>3000</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.8: Value for Back Crossing ($)

<table>
<thead>
<tr>
<th>Value</th>
<th>State</th>
<th>$S_{14}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v(1)$</td>
<td></td>
<td>95973</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Value</th>
<th>State</th>
<th>$S_3$</th>
<th>$S_{10}$</th>
<th>$S_{12}$</th>
<th>$S_{14}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v(2)$</td>
<td>98972</td>
<td>97079</td>
<td>81727</td>
<td>-0.813</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Value</th>
<th>State</th>
<th>$S_1$</th>
<th>$S_{14}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v(3)$</td>
<td>100000</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

**Transition Probability for the Third Generation:**

The following Table 4.10 contains the transition probability from state $S_i$, $i \in \{3, 10, 12, 14\}$ in the second generation to the states $S_1$ and $S_{14}$ in the third generation when we take the actions above.

**Conclusion:**

According the results above, the expected number of progeny we need in total is about 1444 with the success rate being 0.9998. To achieve the same success rate, back crossing strategy needs less progeny to produce compared with finishing the process in two generations. However, this strategy is less efficient than self crossing strategy in our case study.

Table 4.9: Transition Probability $p(S_j|S_i, a_{n_2}(S_i))$ for the Second Generation ($\times 10^{-2}$)

<table>
<thead>
<tr>
<th>State</th>
<th>$S_3$</th>
<th>$S_{10}$</th>
<th>$S_{12}$</th>
<th>$S_{14}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{14}$</td>
<td>93.06</td>
<td>0.0694</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 4.10: Transition Probability $p(S_j|S_i, a_{n3}(S_i))$ for the Third Generation ($\times 10^{-2}$)

<table>
<thead>
<tr>
<th>State</th>
<th>$S_1$</th>
<th>$S_{14}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_3$</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>$S_{10}$</td>
<td>99.72</td>
<td>0.28</td>
</tr>
<tr>
<td>$S_{12}$</td>
<td>84.67</td>
<td>15.33</td>
</tr>
<tr>
<td>$S_{14}$</td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>

**Choice 3: $l_x = l_A$**

We can also discuss a strategy which is back crossing to the donor parent. In such a strategy, we choose the donor $l_A$ to do crossing with. There will be only 2 possible outcomes, either success or failure:

$$
\begin{bmatrix}
1 & 0 \\
0 & 1 \\
1 & 0
\end{bmatrix}
\times
\begin{bmatrix}
0 & 0 \\
1 & 1 \\
0 & 0
\end{bmatrix}
\Rightarrow

\begin{bmatrix}
1 & 0 \\
1 & 1 \\
1 & 0
\end{bmatrix}
$$

$PB(l_7) = 0.43$  $PB(l_{14}) = 0.0025$

The probability of $l_{F1}$ crossing with donor to produce $l_7$ is 0.0021. The probability of $l_{F1}$ crossing with donor to produce $l_{14}$ or failure is 0.9979.

Based on these 2 outcomes, we can derive our state space $\mathbb{S} = \{S_7, S_{14}\}$.
$S_7$: The best progeny we have is $l_7$.

$S_{14}$: The best progeny we have is $l_{14}$ or failure.

The transition between every state in the second and third generation is in Figure 4.4.

Figure 4.4: Transition between States in Donor Crossing
In the first generation, we cross \( l_A \) and \( l_B \) to produce \( a_{n_1} \) progeny, we will reach state \( S_{14} \).

At the beginning of second generation, suppose we take action to produce \( n_2 \) progeny in state \( S_{14} \), we will reach state \( S_7 \) or \( S_{14} \) with probability:

\[
P(S_7|S_{14},a_{n_2}(S_{14})) = 1 - \left[ 1 - P(l_7)^{a_{n_2}(S_{14})} \right]. \quad (4.10)
\]

\[
P(S_{14}|S_{14},a_{n_2}(S_{14})) = 1 - P(S_7|S_{14},a_{n_2}(S_{14})). \quad (4.11)
\]

At the beginning of the third generation, suppose we are at state \( S_i, i \in \{7,14\} \), the best progeny we have is \( l_i \). In this situation, our planting strategy will be self crossing for sure. We only need to decide the number of progeny to produce \( n_3 \) and will have the following probability to reach \( S_1 \) or \( S_{14} \):

\[
P(S_1|S_i,a_{n_3}(S_i)) = 1 - \left[ 1 - PB(l_i)^2 \right]^{a_{n_3}(S_i)}, i \in \{7,14\}. \quad (4.12)
\]

\[
P(S_{14}|S_i,a_{n_3}(S_i)) = 1 - P(S_1|S_i,a_{n_3}(S_i)), i \in \{7,14\}. \quad (4.13)
\]

Combine equation 4.10, equation 4.11, equation 4.13 with equation 4.1, we will draw the following conclusion for the given parameters:

**Action Space:**

The actions shall be taken for each state in the second and third generation are in the Table 4.11.

<table>
<thead>
<tr>
<th>Actions</th>
<th>State</th>
<th>( S_{14} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_{n_2}(S_i) )</td>
<td></td>
<td>2338</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Actions</th>
<th>State</th>
<th>( S_7 )</th>
<th>( S_{14} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_{n_3}(S_i) )</td>
<td></td>
<td>50</td>
<td>0</td>
</tr>
</tbody>
</table>
Value:

The value for each state in the second and third generation is the following Table 4.12.

Table 4.12: Value for Donor Crossing ($) 

<table>
<thead>
<tr>
<th>Value</th>
<th>State</th>
<th>$S_{14}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v(1)$</td>
<td></td>
<td>94104</td>
</tr>
<tr>
<td>$v(2)$</td>
<td>$S_7$</td>
<td>98961</td>
</tr>
<tr>
<td></td>
<td>$S_{14}$</td>
<td>-0.813</td>
</tr>
<tr>
<td>$v(3)$</td>
<td>$S_1$</td>
<td>100000</td>
</tr>
<tr>
<td></td>
<td>$S_{14}$</td>
<td>0</td>
</tr>
</tbody>
</table>

Transition Probability for the Second Generation:

The following Table 4.13 and Table 4.14 are the transition probability for states in the second and third generation when we take the actions above.

Table 4.13: Transition Probability $p(S_j|S_i, a_{n_2}(S_i))$ for the Second Generation ($\times 10^{-2}$) 

<table>
<thead>
<tr>
<th>States</th>
<th>States</th>
<th>$S_7$</th>
<th>$S_{14}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{14}$</td>
<td></td>
<td>99.33</td>
<td>0.67</td>
</tr>
</tbody>
</table>

Transition Probability for the Third Generation:

The following Table 4.14 contains the transition probability from states $(S_i, i \in \{1, 14\})$ in the second generation to the states $(S_1$ and $S_{14} )$ in the third generation when we take the actions above.

Table 4.14: Transition Probability $p(S_j|S_i, a_{n_3}(S_i))$ for the Third Generation ($\times 10^{-2}$) 

<table>
<thead>
<tr>
<th>States</th>
<th>States</th>
<th>$S_1$</th>
<th>$S_{14}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_7$</td>
<td></td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>$S_{14}$</td>
<td></td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>
Conclusion:

From the results above, we will find that this is an even worse strategy which will take 2388 progeny to produce in total to achieve a 0.9932 success rate. What’s more, it is worse than the two generations case.

4.4 Analysis

![Figure 4.5: Comparison for Two Cases](image)

Figure 4.5: Comparison for Two Cases
We can compare the results from different cases and choices in the Figure 4.5 above to get the following conclusion:

**Advantage:**

In general, to achieve the same success rate, three generations case will need less progeny to produce compared with two generations case. In addition, in the MDP model for three generations case, self crossing is more efficient compared with back crossing and donor crossing.

**Limitation:**

In actual breeding systems, thousands of alleles can be arrayed, which will increase state space enormously. In such situations, considering the uncertainty of each state is difficult to make the MDP model be inefficient.

If we can find some better methods to category and redefine the state space, we can find better ways to apply the MDP model to plant parental selection problem. This will be part of the future work.
CHAPTER 5. SIMULATIONS

5.1 Introduction

To overcome the limitation of the MDP model, we design the parental breeding value based simulation which simulates the performance of the parental breeding value. We will compare two different crossing strategies to show the improvement of efficiency based on the parental breeding value.

• **Strategy 1. Parental Breeding Value For Crossing Decisions:** In every generation, choose two individuals with the largest parental breeding value from the population to cross.

• **Strategy 2. Parental Breeding Value In Back Crossing Strategy:** In every generation, choose an individual with the largest parental breeding value from the population to do back crossing with the elite individual. When we find one individual with an entire column being desirable, do self crossing.

5.2 Simulation Parameters

We assume the donor plant contains 50 alleles with 3 of the 50 are desirable for improving an elite individual. The first desirable allele is at locus 7, the second desirable allele is at locus 10 and the third desirable allele is at locus 35. We randomly generate a recombination frequency vector. The first step is to cross the donor with the elite individual to produce $F_1$ generation. The number of progeny to produce in each crossing is simulated from 50 to 500 with a step size of 50. We generated 100 simulations for two crossing strategies.
The gray blocks in the Figure 5.2, 5.3 and 5.4 represent the desirable alleles and the black blocks represent the undesirable alleles. The illustration of the elite plant is in Figure 5.2. The illustration of the donor is in Figure 5.3 and an illustration of their $F_1$ progeny as offspring is in Figure 5.4.
A random simulation with 100 \( F_1 \) progeny plants per generation is shown in Figure 5.5.

We would like to get the relation between average generation number needed to achieve the genotype with all alleles being desirable, different crossing strategies and the number of progeny to produce in every crossing.

5.3 Results

From the result of the simulation by MATLAB, we conclude the Figure 5.6. The Figure 5.6 shows the relation between number of progeny to produce in each crossing and average generation number for two crossing strategies.

From the simulation, we are able to derive the Table 5.1 to show the performance of the parental breeding value based simulation (\( n_p \): progeny number per generation, \( n_g(S_1) \): average generation number for strategy 1, \( n_g(S_2) \): average generation number for strategy 2, \( \Delta \): difference of generation number between two strategies, \( \Delta/n_g(S_2) \): improvement of strategy 1 compared with strategy 2):

<table>
<thead>
<tr>
<th>( n_p )</th>
<th>50</th>
<th>100</th>
<th>150</th>
<th>200</th>
<th>250</th>
<th>300</th>
<th>350</th>
<th>400</th>
<th>450</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n_g(S_1) )</td>
<td>7.29</td>
<td>6.72</td>
<td>6.51</td>
<td>6.17</td>
<td>6.07</td>
<td>6.02</td>
<td>5.98</td>
<td>5.98</td>
<td>5.98</td>
<td>5.93</td>
</tr>
<tr>
<td>( n_g(S_2) )</td>
<td>9.36</td>
<td>7.81</td>
<td>6.94</td>
<td>6.75</td>
<td>6.68</td>
<td>6.61</td>
<td>6.45</td>
<td>6.35</td>
<td>6.27</td>
<td>6.32</td>
</tr>
<tr>
<td>( \Delta )</td>
<td>2.07</td>
<td>1.09</td>
<td>0.43</td>
<td>0.58</td>
<td>0.61</td>
<td>0.59</td>
<td>0.47</td>
<td>0.37</td>
<td>0.38</td>
<td>0.39</td>
</tr>
<tr>
<td>( \Delta/n_g(S_2) )</td>
<td>0.22</td>
<td>0.14</td>
<td>0.06</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0.07</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
</tr>
</tbody>
</table>
**Analysis:**

- The resource cost associated with the parental breeding value based crossing strategy is less than parental breeding value based back crossing. Given the same number of generation, the parental breeding value based crossing strategy needs fewer progeny to produce each generation of crossing.

- The time cost of the parental breeding value based crossing strategy is less than the parental breeding value based back crossing. Given the same number of progeny to produce, the parental breeding value based crossing strategy needs fewer generations to produce the desired breeding individual.

We can get the conclusion that parental breeding value based crossing strategy is more efficient compared with parental breeding value based back crossing strategy. It costs fewer progeny and less time (resources and time) to accomplish the objective of getting a progeny with all alleles being desirable.
Figure 5.5: A Random Simulation
Figure 5.6: Simulation Result
CHAPTER 6. SUMMARY AND DISCUSSION

6.1 Conclusion

This thesis is aiming at solving the parent selection problem in plant breeding. It introduced the history and development of genetics and plant breeding at the first beginning. Then the thesis discussed about the existing methods to design an efficient strategy for plant breeding. There are many classic and extraordinary approaches come up by plant breeders and scientists. At the same time, some methods from operations research prospective broaden the horizon of plant breeding. Taking advantage of operations research, finding an optimal strategy for plant breeding is possible.

Then, the thesis brought in a new concept of parental breeding value and the water pipe model for calculation. It mainly focused on calculating the probability to produce a gamete with all alleles being desirable. This method can effectively assess the efficiency of selected parental individuals. An MDP model was introduced in the following section. The model treats each possible outcome from a crossing as a state and work out the optimal policy for a simple parental selection problem. However, due to the complexity of genotype and limitation of computer solver, it is difficult to apply this approach to realistic strategy design.

To test the performance of the parental breeding value for realistic problems, we designed parental breeding value based crossing strategy. Using MATLAB and commercial PC for simulations, the strategy is proved to be more efficient compared with back crossing strategy in terms of resources and time.

In conclusion, the thesis utilizes the methods from operations research to find a new direction to solving the parental selection problem. Based on the performance of the MDP approach and simulation, the parental breeding value is believed to be able to make contribution to finding
better plant breeding strategies.

6.2 Future Work

The future work of this thesis will be mainly on improving the water pipe model for calculating the parental breeding value and designing new strategies for complex crossing problem like parallel crossing. An integer-programming problem will be involved to find the maximum parental breeding value for a large pool of candidate individuals. Besides, we will modify our parental breeding value for more general cases. It will be able to calculate the parental breeding value for any \( k \) individuals. With the parental breeding value for any \( k \) individuals, we can deal with some complex problems like transferring multiple desirable alleles from multiple donor individuals to one elite individual with parallel crossing. For the MDP model, new states for large-scale problem will be designed to make the model work efficiently.

Together with this new parental breeding value, we hope to find the optimal crossing strategy for any type of parent selection problem. This future work will be part of my PhD study.
APPENDIX A. PARENTAL BREEDING VALUE PROPOSITION
PROOF FOR ONE INDIVIDUAL

Proof: $P(g_i = x_{i,j} = 1) = P(g_i = x_{i,j}, x_{i,j} = 1) = P(g_i = x_{i,j}) P(x_{i,j} = 1) = P(g_i = x_{i,j}) X_{i,j}$. \(\square\)

Proof: (a).

\[
w_{1,j} \\
= P(g_1 = x_{1,j} = 1) \\
= P(g_1 = x_{1,j}) P(x_{1,j} = 1), j \in \{1, 2\}. \\
P(g_1 = x_{1,j}) = 1/2, j \in \{1, 2\}. \\
P(x_{1,j} = 1) = X_{1,j} \\
\Rightarrow w_{1,j} = (1/2) X_{1,j}, j \in \{1, 2\}.
\]
(b).

\[ w_{i,1} \]

\[ = P(g_i = x_{i,1} = 1, g_{i-1} = x_{i-1,1} = 1, g_{i-2} = 1...g_1 = 1) \]

\[ = P(g_i = x_{i,1} = 1, g_{i-1} = x_{i-1,1} = 1, g_{i-2} = 1...g_1 = 1) \]

\[ + P(g_i = x_{i,1} = 1, g_{i-1} = x_{i-1,2} = 1, g_{i-2} = 1...g_1 = 1) \]

\[ = P(g_i = x_{i,1} = 1|g_{i-1} = x_{i-1,1} = 1)P(g_{i-1} = x_{i-1,1} = 1, g_{i-2} = 1...g_1 = 1) \]

\[ + P(g_i = x_{i,1} = 1|g_{i-1} = x_{i-1,2} = 1)P(g_{i-1} = x_{i-1,2} = 1, g_{i-2} = 1...g_1 = 1) \]

\[ = P(x_{i,1} = 1)[P(g_i = x_{i,1}|g_{i-1} = x_{i-1,1} = 1)w_{i-1,1} \]

\[ + P(g_i = x_{i,1}|g_{i-1} = x_{i-1,2} = 1)w_{i-1,2}] \]

\[ = [(1 - r_{i-1})w_{i-1,1} + r_{i-1}w_{i-1,2}]X_{i,1} \]

\[ = w_{i-1,1}M^i_{1}X_{i,1}. \]

The proof for \( w_{i,2} \) is following the similar procedures. \( \square \)
APPENDIX B. PARENTAL BREEDING VALUE PROPOSITION
PROOF FOR TWO OR MULTIPLE INDIVIDUAL

B.1 Parental Breeding Value Proposition for Two Individuals

Proof: (a).

\[ w_{1,j} = P(g_1 = x_{1,j} = 1) \]
\[ = P(g_1 = x_{1,j})P(x_{1,j} = 1), j \in \{1, ..., 4\}. \]
\[ P(g_1 = x_{1,j}) = 1/4, j \in \{1, ..., 4\}. \]
\[ P(x_{1,j} = 1) = X_{1,j} \]
\[ \Rightarrow w_{1,j} = (1/4)X_{1,j}, j \in \{1, ..., 4\}. \]
(b).

\[ w_{i,1} \]

\[ = P(g_i = Z_{i,1} = x_{i,1} = 1, g_{i-1} = 1, \ldots g_1 = 1) \]

\[ = P(g_i = Z_{i,1} = x_{i,1} = 1, g_{i-1} = Z_{i-1,1} = x_{i-1,1} = 1, g_{i-2} = 1 \ldots g_1 = 1) + P(g_i = Z_{i,1} = x_{i,1} = 1, g_{i-1} = Z_{i-1,1} = x_{i-1,2} = 1, g_{i-2} = 1 \ldots g_1 = 1) + P(g_i = Z_{i,1} = x_{i,1} = 1, g_{i-1} = Z_{i-1,2} = y_{i-1,1} = 1, g_{i-2} = 1 \ldots g_1 = 1) + P(g_i = Z_{i,1} = x_{i,1} = 1, g_{i-1} = Z_{i-1,2} = y_{i-1,2} = 1, g_{i-2} = 1 \ldots g_1 = 1) \]

\[ = P(g_i = Z_{i,1} = x_{i,1} = 1|g_{i-1} = Z_{i-1,1} = x_{i-1,1} = 1)P(g_{i-1} = Z_{i-1,1} = x_{i-1,1} = 1, g_{i-2} = 1 \ldots g_1 = 1) + P(g_i = Z_{i,1} = x_{i,1} = 1|g_{i-1} = Z_{i-1,1} = x_{i-1,2} = 1)P(g_{i-1} = Z_{i-1,1} = x_{i-1,2} = 1, g_{i-2} = 1 \ldots g_1 = 1) + P(g_i = Z_{i,1} = x_{i,1} = 1|g_{i-1} = Z_{i-1,2} = y_{i-1,1} = 1)P(g_{i-1} = Z_{i-1,2} = y_{i-1,1} = 1, g_{i-2} = 1 \ldots g_1 = 1) + P(g_i = Z_{i,1} = x_{i,1} = 1|g_{i-1} = Z_{i-1,2} = y_{i-1,2} = 1)P(g_{i-1} = Z_{i-1,2} = y_{i-1,2} = 1, g_{i-2} = 1 \ldots g_1 = 1) \]

\[ = P(x_{i,1} = 1)[P(g_i = Z_{i,1} = x_{i,1}|g_{i-1} = Z_{i-1,1} = x_{i-1,1} = 1)w_{i-1,1} + P(g_i = Z_{i,1} = x_{i,1}|g_{i-1} = Z_{i-1,1} = x_{i-1,2} = 1)w_{i-1,2} + P(g_i = Z_{i,1} = x_{i,1}|g_{i-1} = Z_{i-1,1} = y_{i-1,1} = 1)w_{i-1,3} + P(g_i = Z_{i,1} = x_{i,1}|g_{i-1} = Z_{i-1,1} = y_{i-1,2} = 1)w_{i-1,4}] \]

\[ = X_{i,1} \left[ (1 - r_{i-1})^2 w_{i-1,1} + r_{i-1}(1 - r_{i-1})w_{i-1,2} + 0.5r_{i-1}(w_{i-1,3} + w_{i-1,4}) \right] \]

\[ = w_{i-1}\cdot M_{i,1} X_{i,1}. \]

The proof for \( w_{i,2}, w_{i,3}, w_{i,4} \) is following the similar procedures. \( \square \)
B.2 Parental Breeding Value Proposition for $K = 2^k$ Individuals

Proof: (a).

\[
w_{1,j} = P(g_1 = x_{1,j} = 1) = P(g_1 = x_{1,j})P(x_{1,j} = 1), j \in \{1, \ldots, 2K\}.
\]

\[
P(g_1 = x_{1,j}) = 1/2^K, j \in \{1, \ldots, 2K\}.
\]

\[
P(x_{1,j} = 1) = X_{1,j}
\]

\[
\Rightarrow w_{1,j} = (1/2^K)X_{1,j}, j \in \{1, \ldots, 2K\}.
\]

(b). By mathematical induction, we are able to prove this part.

The proof for $w_{i,j}, j \in \{2, \ldots 2K\}$ is following the similar procedures. \qed


