

To the Graduate Council:

I am submitting herewith a thesis written by Mahendra Duwal Shrestha entitled “Analysis and Simulation Of A Simple Evolutionary System.” I have examined the final paper copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Computer Science.

Michael D. Vose, Major Professor

We have read this thesis
and recommend its acceptance:

Michael D. Vose

Hairong Qi

Judy D. Day

Accepted for the Council:

Dixie Thompson

Vice Provost and Dean of the Graduate School

To the Graduate Council:

I am submitting herewith a thesis written by Mahendra Duwal Shrestha entitled “Analysis and Simulation Of A Simple Evolutionary System.” I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Computer Science.

Michael D. Vose, Major Professor

We have read this thesis
and recommend its acceptance:

Michael D. Vose

Hairong Qi

Judy D. Day

Accepted for the Council:

Dixie Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Analysis and Simulation Of A Simple Evolutionary System

A Thesis Presented for
The Master of Science
Degree

The University of Tennessee, Knoxville

Mahendra Duwal Shrestha

August 2016

© by Mahendra Duwal Shrestha, 2016
All Rights Reserved.

dedication...

Acknowledgements

I would like to thank...

Some quotation...

Abstract

Abstract text goes here...

Contents

1	Introduction	1
1.1	Notation	1
1.2	Background	2
1.3	Random Heuristic Search	6
1.4	Problem Statement	10
2	Extending A Genetic Algorithm Model To The Diploid Case	13
2.1	Model	14
2.2	Reduction	15
2.3	Specialization	17
2.3.1	Mutation	17
2.3.2	Crossover	18
2.3.3	Mixing Matrix	19
2.4	Walsh Transform	20
2.4.1	Fast Walsh Transform	20
2.4.2	Walsh Transform Adaptation	21
2.5	Distance	23
2.6	Simplification	24
2.7	Convergence	25
2.8	Summary	28

3	Evolutionary Limits	30
3.1	Limits	30
3.2	Computation of Mutation and Crossover Distribution	32
3.3	Initial Population	34
3.4	Oscillation	36
3.5	Violation	47
3.5.1	Violation in Mutation Distribution	47
3.5.2	Violation in Crossover Distribution	76
3.6	Summary	104
4	Conclusion	105
	Bibliography	107
	Vita	111

List of Tables

3.1	Expected single step distance d for population size N	46
3.2	Experimental distance measured for oscillation: N is population size, ℓ is genome length and average distance between finite and infinite population is tabulated in the last three columns. $\{4096, 40960, 81920\}$	46
3.3	Experimental distance measured for violation in μ for haploids: ℓ is genome length, ϵ is error introduced to μ for violation, and average distance between finite and infinite populations is tabulated in last three columns.	62
3.4	Experimental distance measured for violation in μ for diploids: ℓ is genome length, ϵ is error introduced to μ for violation, and average distance between finite and infinite populations is tabulated in last three columns.	75
3.5	Experimental distance measured for violation in χ for haploids: ℓ is genome length, ϵ is error introduced to χ for violation, and average distance between finite and infinite populations is tabulated in last three columns.	90
3.6	Experimental distance measured for violation in χ: ℓ is genome length, ϵ is error introduced to χ for violation, and average distance between finite and infinite populations is tabulated in last three columns.	103

List of Figures

1.1	Finite GA	3
1.2	Population points	8
2.1	Convergence of finite population behaviour: d is distance between finite population \mathbf{f}^n and infinite population \mathbf{q}^n at generation n , population size N , for genome length ℓ (bits).	27
2.2	Regression parameters: multi-plot of slope m and intercept b for generation $n \in \{1, 2, 4, 8, 16, 32, 64, 128\}$	28
3.1	Initial population computation	34
3.2	Infinite and finite haploid population oscillation behavior for genome length $\ell = 8$ (bits): In left column, d is distance of finite population of size n or infinite population to limits for g generations. In right column, d is distance of finite population to infinite population for g generations and d_{avg} is average of distance from 1 to 50 generations.	38
3.3	Infinite and finite diploid population oscillation behavior for genome length $\ell = 8$ (bits): In left column, d is distance of finite population of size n or infinite population to limits for g generations. In right column, d is distance of finite population to infinite population for g generations and d_{avg} is average of distance from 1 to 50 generations.. . . .	39

- 3.4 **Infinite and finite haploid population oscillation behavior for genome length $\ell = 10$ (bits):** In left column, d is distance of finite population of size n or infinite population to limits for g generations. In right column, d is distance of finite population to infinite population for g generations and d_{avg} is average of distance from 1 to 50 generations.. 40
- 3.5 **Infinite and finite population oscillation behavior for genome length $\ell = 10$ (bits):** In left column, d is distance of finite population of size n or infinite population to limits for g generations. In right column, d is distance of finite population to infinite population for g generations and d_{avg} is average of distance from 1 to 50 generations.. 41
- 3.6 **Infinite and finite haploid population oscillation behavior for genome length $\ell = 12$ (bits):** In left column, d is distance of finite population of size n or infinite population to limits for g generations. In right column, d is distance of finite population to infinite population for g generations and d_{avg} is average of distance from 1 to 50 generations.. 42
- 3.7 **Infinite and finite diploid population oscillation behavior for genome length $\ell = 12$ (bits):** In left column, d is distance of finite population of size n or infinite population to limits for g generations. In right column, d is distance of finite population to infinite population for g generations and d_{avg} is average of distance from 1 to 50 generations.. 43
- 3.8 **Infinite and finite haploid population oscillation behavior for genome length $\ell = 14$ (bits):** In left column, d is distance of finite population of size n or infinite population to limits for g generations. In right column, d is distance of finite population to infinite population for g generations and d_{avg} is average of distance from 1 to 50 generations.. 44

3.9	Infinite and finite diploid population oscillation behavior for genome length $\ell = 14$ (bits): In left column, d is distance of finite population of size n or infinite population to limits for g generations. In right column, d is distance of finite population to infinite population for g generations and d_{avg} is average of distance from 1 to 50 generations..	45
3.10	Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 8$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	49
3.11	Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 8$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	50
3.12	Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 8$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	51
3.13	Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 10$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	52

3.14	Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 10$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	53
3.15	Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 10$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	54
3.16	Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 12$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	55
3.17	Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 12$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	56

3.18	Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 12$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	57
3.19	Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 14$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	58
3.20	Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 14$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	59
3.21	Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 14$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	60

3.22	Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 8$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	62
3.23	Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 8$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	63
3.24	Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 8$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	64
3.25	Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 10$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	65

3.26	Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 10$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	66
3.27	Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 10$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	67
3.28	Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 12$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	68
3.29	Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 12$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	69

3.30	Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 12$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	70
3.31	Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 14$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	71
3.32	Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 14$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	72
3.33	Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 14$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	73

3.34	Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 8$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	77
3.35	Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 8$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	78
3.36	Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 8$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	79
3.37	Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 10$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	80

3.38	Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 10$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	81
3.39	Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 10$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	82
3.40	Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 12$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	83
3.41	Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 12$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	84

3.42	Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 12$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	85
3.43	Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 14$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	86
3.44	Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 14$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	87
3.45	Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 14$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	88

3.46	Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 8$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	90
3.47	Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 8$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	91
3.48	Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 8$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	92
3.49	Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 10$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	93

3.50	Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 10$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	94
3.51	Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 10$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	95
3.52	Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 12$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	96
3.53	Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 12$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	97

3.54	Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 12$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	98
3.55	Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 14$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	99
3.56	Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 14$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	100
3.57	Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 14$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.	101

Chapter 1

Introduction

This thesis begins with notation that is used throughout this document.

1.1 Notation

Some standard mathematical notations as well as some non-standard mathematical notations are introduced here.

Angle brackets $\langle \cdots \rangle$ denote a tuple which is to be regarded as a column vector. The column vector of all 1s is denoted by $\mathbf{1}$. Transpose is indicated with superscript T . The standard vector norm is $\|x\| = \sqrt{x^T x}$. Modulus (or absolute value) is denoted by $|\cdot|$. When S is a set, $|S|$ denotes the cardinality of S .

The notation $O(f)$ denotes a function g (with similar domain and codomain as f), such that pointwise $g \leq cf$ for some constant c . The notation $\theta(f)$ is a function g such that pointwise $c_0 f \leq g \leq c_1 f$ for some constants c_0, c_1 . Curly brackets $\{\cdots\}$ are used as grouping symbols and to specify both sets and multisets. Square brackets $[\cdots]$ are, besides their standard use as specifying a closed interval of real numbers, used to denote an indicator function: if $expr$ is an expression which may be true or false, then $[expr]$ denotes 1 if $expr$ is true, and 0 otherwise; $[\cdots]$ is sometimes referred to as an *Iverson bracket*.

The supremum is the least upper bound, and is denoted by \sup . The infimum is the greatest lower bound, and is denoted by \inf .

The set of length ℓ binary strings is denoted by \mathcal{R} . It is a commutative ring under component-wise addition and multiplication modulo 2. If $x \in \mathcal{R}$, then it may be regarded as the vector $x = \langle x_0, x_1, \dots, x_{\ell-1} \rangle$. Denote the additive identity of \mathcal{R} by $\mathbf{0}$ and the multiplicative identity by $\mathbf{1}$. Let \bar{g} abbreviate $\mathbf{1} + g$. Except when explicitly indicated otherwise, operations acting on elements of \mathcal{R} are as defined in this paragraph. In particular, $g\bar{g} = \mathbf{0} = g + g$, $g^2 = g$, $g + \bar{g} = \mathbf{1}$ for all $g \in \mathcal{R}$.

1.2 Background

The genetic algorithm (GA) is inspired by nature, and seeks to evolve useful constructs. It is typically population based, and proceeds over a number of generations to evolve solutions to problems not yielding to other known methods. Basic elements of a GA are: selection according to fitness, crossover, and random mutation (see [Mitchell \(1999\)](#)). In the simplest case, population members are fixed-length binary strings. The fitness function assigns a score (fitness) to the elements (chromosomes) of the current population.

Selection: select population members in the current population for reproduction; those with higher fitness are more likely to be selected to reproduce.

Crossover: with some probability (the crossover rate), choose a random point in two parents (population members selected for reproduction) and exchanges subsequences after that point to create two offspring.

Mutation: flip bits of an individual with some small probability, the mutation rate.

Figure [1.1](#) shows the procedural flow of a basic finite population genetic algorithm.



Figure 1.1: Finite GA

A simple Holland style genetic algorithm (see [Holland \(1992\)](#)):

A simple genetic algorithm

1. Start with some initial population P containing r binary strings of length ℓ .
 2. Choose (with replacement) parents u and v from the current population P (using any selection scheme).
 - a. Cross u and v to produce children u' and v' .
 - b. Mutate u' and v' with some probability to produce u'' and v'' .
 - c. Keep, with uniform probability, one of u'' and v'' for the next generation
 3. If the next generation contains fewer than r members, repeat step 2.
 4. Replace P by the new generation formed and go to step 2.
-

Each iteration of this process produces a generation. The process is repeated until the system stops to improve or some threshold is met.

The infinite population GA models a population as a probability vector \mathbf{p} where component \mathbf{p}_j can be interpreted as the proportion of string j in the population. If \mathcal{G} is the function mapping infinite population \mathbf{p} to the next generation, $\mathcal{G}(\mathbf{p})$ is a

probability vector such that

$\mathcal{G}(\mathbf{p})_j$ = the probability that string j occurs in the next generation.

The evolution of infinite population \mathbf{p} is the sequence

$$\mathbf{p} \rightarrow \mathcal{G}(\mathbf{p}) \rightarrow \mathcal{G}(\mathcal{G}(\mathbf{p})) \rightarrow \dots$$

Several people working in the 1950s and the 1960s – like Box (1957), Friedman (1959), Bledsoe (1961), Bremermann (1962), and Reed, Toombs and Baricelli (1967) – developed evolution-inspired algorithms, but little attention or theoretical analysis was given to them (see [Mitchell \(1999\)](#)). Genetic algorithms were popularized by Holland and his colleagues in the 1960s and the 1970s. Holland introduced a population-based algorithm with crossover and mutation, and promoted his schema theorem (see [Holland \(1992\)](#)) as a theoretical means to analyze on genetic algorithm dynamics. Holland’s Schema theorem provides a lower bound for schema survival in next generation. * The schema theorem is an inequality however, and can not predict which strings are expected in the next generation. Bethke (see [Bethke \(1980\)](#)) gave equations computing the expected number of any string in the next generation. Goldberg (see [Goldberg \(1987\)](#)) used such equations to model the evolutionary trajectory of a two bit GA under crossover and proportional selection. Vose and Liepins (see [Vose and Liepins \(1991\)](#)) simplified and extended these equations by integrating mutation into the recombination of arbitrarily long binary strings. Although their model computes infinite population trajectories, given a *finite* population represented by vector \mathbf{p} (component \mathbf{p}_i is the proportion of string i in the *finite population*), the infinite population model computes the expected proportion

*A schema is a template that identifies a set of strings in the population with similarities at certain string positions; it is made up of 1s, 0s, and *s where * is the ‘don’t care’ symbol that matches either 0 or 1.

$\mathcal{G}(\mathbf{p})_i$ of string i in the next generation. This is perhaps the most direct connection between the infinite population model and a finite population GA.

Nix and Vose (see [Nix and Vose \(1992\)](#)) explored issues regarding the relationship between the finite population GA and the infinite population model. For a positive mutation rate, a finite population GA will form an ergodic Markov chain, visiting every state infinitely often in the long run. The short term trajectory followed by a finite population is related to the evolutionary path determined by the infinite population model. For large populations, the short term trajectory follows closely and with large probability, that path predicted by the infinite population model.

Vose compiled and extended previous work regarding the infinite population model in the book *Simple Genetic Algorithm: Foundations and Theory* (see [Vose \(1999\)](#)). In particular, he discussed how the Walsh transform can be applied to increase computational efficiency in calculating the infinite population model. There have been previous applications of the Walsh transform to GAs. Bethke first introduced the idea of using Walsh transforms to analyse GA fitness functions in terms of schemata (see [Bethke \(1980\)](#)). The idea was further developed in papers by Goldberg (see [Goldberg \(1989a\)](#), [Goldberg \(1989b\)](#)). However, such usage did not apply Walsh transforms to crossover, to mutation, or to any of their associated mathematical objects. In contrast, Vose and Liepins applied the Walsh transform directly to mutation and recombination, and proved that the twist M^* of the mixing matrix M is triangularized by the Walsh transform, and related eigenvalues of M^* to the stability of fixed points of \mathcal{G} (see [Vose and Liepins \(1991\)](#)).[†] In a related paper, Koehler (see [Koehler \(1994\)](#)) gives a congruence transformation defined by a lower triangular matrix that diagonalizes the mixing matrix (for 1-point crossover and mutation given by a rate) and proved a conjecture of Vose and Liepins concerning eigenvalues of M^* . Koehler, Bhattacharyya and Vose (see [Koehler et al. \(1997\)](#)) applied the Fourier

[†]The mixing matrix M has rows and columns indexed by chromosomes; entry $M_{i,j}$ is the probability that mixing parents i and j (mixing is the combined effect of crossover and mutation) will produce a child having all bits zero. The twist (M^*) of the mixing matrix M is defined by $(M^*)_{i,j} = M_{i+j,i}$.

transform in generalizing results established for binary GAs to strings over an alphabet of cardinality c (in the binary case, the Fourier transform is the Walsh transform). From a computational perspective, a major contribution of Vose and Wright (see [Vose and Wright \(1998\)](#)) was demonstrating that the mixing matrix is sparse in the Walsh basis, and the computational efficiency of computing $\mathcal{G}(\mathbf{p})$ can thereby be improved from $O(8^\ell)$ to $O(3^\ell)$ where ℓ is the chromosome length. The cost of moving from standard coordinates to the Walsh basis need not be a bottleneck; the fast Walsh transform (see [Shanks \(1969\)](#)) does that in $O(\ell 2^\ell)$ time.

1.3 Random Heuristic Search

The work presented in this thesis is based on *Random Heuristic Search (RHS)*, a general search method, defined upon the central concept of state and transition between states (see [Vose \(1999\)](#)). The simple genetic algorithm is a particular type of RHS. An instance of *RHS* is an initial collection of elements P (referred to as the initial population) chosen from some search space Ω , together with a stochastic transition rule τ , which from P will produce another collection P^* ; iterating τ produces a sequence of generations.

Let n be the cardinality of Ω , let $\mathbf{1}$ denote the column vector of all 1s. The *simplex* is the set of population descriptors:

$$\Lambda = \{x = \langle x_0, \dots, x_{n-1} \rangle : \mathbf{1}^T x = 1, x_j \geq 0\}$$

Element $\mathbf{p} \in \Lambda$ corresponds to a population; p_j = the proportion in the population of the j th element of Ω . The cardinality of each population is a constant r , called the population size. Given r , a population descriptor \mathbf{p} unambiguously determines a population.

Given current population vector \mathbf{p} , the next population vector $\tau(\mathbf{p})$ cannot be predicted with certainty because τ is stochastic; it results from r independent,

identically distributed random choices. Let $\mathcal{G} : \Lambda \rightarrow \Lambda$ be a function that maps current population vector \mathbf{p} to a vector whose i th component is the probability that the i th element of Ω is chosen. Thus, $\mathcal{G}(\mathbf{p})$ specifies the distribution from which the aggregate of r choices forms the subsequent generation. The probability that population \mathbf{q} is the next population vector given current population (vector) \mathbf{p} is (see Vose (1999))

$$\begin{aligned} Q_{\mathbf{p},\mathbf{q}} &= r! \prod \frac{(\mathcal{G}(\mathbf{p})_j)^{r\mathbf{q}_j}}{(r\mathbf{q}_j)!} \\ &= \exp\left\{-r \sum \mathbf{q}_j \log \frac{\mathbf{q}_j}{\mathcal{G}(\mathbf{p})_j} - \sum (\log \sqrt{2\pi r\mathbf{q}_j} + \frac{1}{12r\mathbf{q}_j + \theta(r\mathbf{q}_j)})\right. \\ &\quad \left.+ O(\log r)\right\} \end{aligned} \quad (1.1)$$

where summation is restricted to indices for which $\mathbf{q}_j > 0$ and θ is a function such that $0 < \theta < 1$. Each random vector in the sequence $\mathbf{p}, \tau(\mathbf{p}), \tau^2(\mathbf{p}), \dots$ depends only on the value of the preceding one, which is a special situation. The sequence forms a Markov chain with transition matrix Q . The conceptualization of RHS can be replaced by a Markov chain model which makes no reference to sampling Ω ; from current population \mathbf{p} , produce \mathbf{q} with probability $Q_{\mathbf{p},\mathbf{q}}$. The expected next generation $\mathcal{E}(\tau(\mathbf{p}))$ is $\mathcal{G}(\mathbf{p})$ (see Vose (1999)). The expression

$$\sum \mathbf{q}_j \log \frac{\mathbf{q}_j}{\mathcal{G}(\mathbf{p})_j}$$

in (1.1) is the *discrepancy* of \mathbf{q} with respect to $\mathcal{G}(\mathbf{p})$. It is a measure of how far \mathbf{q} is from the expected next population $\mathcal{G}(\mathbf{p})$. Discrepancy is nonnegative and is zero only when \mathbf{q} is the expected next population. Hence the first factor

$$\exp\left\{-r \sum \mathbf{q}_j \log \frac{\mathbf{q}_j}{\mathcal{G}(\mathbf{p})_j}\right\}$$

in (1.1) indicates the probability that \mathbf{q} is the next generation decays exponentially, with constant r , as the discrepancy between \mathbf{q} and $\mathcal{G}(\mathbf{p})$ increases. The expression

$$\sum (\log \sqrt{2\pi r \mathbf{q}_j} + \frac{1}{12r \mathbf{q}_j + \theta(r \mathbf{q}_j)})$$

measures the *dispersion* of the population vector \mathbf{q} and the second factor in (1.1)

$$\exp\{-\sum (\log \sqrt{2\pi r \mathbf{q}_j} + \frac{1}{12r \mathbf{q}_j + \theta(r \mathbf{q}_j)})\}$$

indicates the probability that \mathbf{q} is the next generation decays exponentially with increasing dispersion.

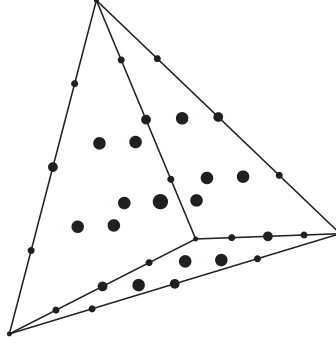


Figure 1.2: Population points

Figure 1.2 illustrates population points in a simplex for $\ell = 2, r = 4$. Finite populations are represented by dots, where smaller dots have lower dispersion and larger dots have higher dispersion. The diagram also illuminates that finite populations are constrained to occupy lattice points within Λ . As population size $r \rightarrow \infty$, the lattice points become dense in Λ , which corresponds to the fact that an infinite population can be (represented by) any point of Λ .

The variance of the next generation (with respect to the expected population) (see Vose (1999)) is

$$\mathcal{E}(\|\tau(\mathbf{p}) - \mathcal{G}(\mathbf{p})\|^2) = \frac{1 - \|\mathcal{G}(\mathbf{p})\|^2}{r} \quad (1.2)$$

It follows from Chebyshev's inequality (see [Wikipedia \(2016a\)](#)) that

$$P(\|\tau(\mathbf{p}) - \mathcal{G}(\mathbf{p})\| \geq \epsilon) \leq \frac{1 - \|\mathcal{G}(\mathbf{p})\|^2}{r\epsilon^2} \quad (1.3)$$

where P above denotes probability and $\epsilon > 0$ is arbitrary.

Let $f(r)$ be a function which grows arbitrarily slowly, such that

$$\lim_{r \rightarrow \infty} f(r) = \infty$$

and

$$\lim_{r \rightarrow \infty} f(r)/\sqrt{r} = 0.$$

If

$$\epsilon = f(r)/\sqrt{r} \quad (1.4)$$

, then (1.3) becomes

$$\lim_{r \rightarrow \infty} P(\|\tau(\mathbf{p}) - \mathcal{G}(\mathbf{p})\| \geq \epsilon) \leq \frac{(1 - \|\mathcal{G}(\mathbf{p})\|^2)}{f(r)^2} \leq 0$$

Therefore, $\tau(\mathbf{p})$ converges in probability to $\mathcal{G}(\mathbf{p})$ as the population size increases, and τ corresponds to \mathcal{G} in the infinite population case. Moreover, 1.4 suggests that the expected distance between finite and infinite population in the next generation might decrease as $1/\sqrt{r}$.

In figure 1.2, finite population points can be only at certain points, but infinite population points can be anywhere in the simplex. Theorem 3.1 in 'The Simple Genetic Algorithm: Foundations and Theory' states:

If $\mathbf{p}, \mathbf{q} \in \Lambda$ are arbitrary population vectors for population size r , and $\boldsymbol{\xi}$ denotes arbitrary element of Λ , then

$$\inf_{\mathbf{p} \neq \mathbf{q}} \|\mathbf{p} - \mathbf{q}\| = \sqrt{2}/r \quad (1.5)$$

$$\sup_{\boldsymbol{\xi}} \inf_{\mathbf{p}} \|\boldsymbol{\xi} - \mathbf{p}\| = O(1/\sqrt{r}) \quad (1.6)$$

where the constant (in the "big oh") is independent of the dimension n of Λ (see [Vose \(1999\)](#)). From ??, the distance between an infinite population $\boldsymbol{\xi}$ and finite population \mathbf{p} is $O(1/\sqrt{r})$. This suggests that the distance between $\tau(\mathbf{p})$ and $\mathcal{G}(\mathbf{p})$ might decrease as $1/\sqrt{r}$.

Let η be the random variable $\|\mathbf{q} - \mathcal{G}(\mathbf{p})\|$, and let $\phi(x) = x^2$. It follows from Jensen's Inequality (see [Wikipedia \(2016b\)](#)) that since ϕ is a convex function,

$$\phi(\mathcal{E}(\eta)) \leq \mathcal{E}(\phi(\eta))$$

Therefore,

$$\mathcal{E}(\|\mathbf{q} - \mathcal{G}(\mathbf{p})\|) = \mathcal{E}(\eta) \leq \sqrt{\mathcal{E}(\eta^2)} = \frac{\sqrt{1 - \|\mathcal{G}(\mathbf{p})\|^2}}{\sqrt{r}} \quad (1.7)$$

1.4 Problem Statement

Equation 1.7 bounds the expected rate of convergence for the single-step haploid case; the distance is inversely proportional to square root of population size. Theorem 3.1 from 'The Simple Genetic Algorithm: Foundation and Theory', and Jensen's inequality analysis of 1.2 both suggests that the distance between finite population and infinite population decreases like $1/\sqrt{r}$. We investigate the rate of decrease with experiments in Chapter 2. We begin with a simple Markov model for diploid case under influence of mutation and crossover. The model is non-overlapping, generational, infinite population model assuming random mating and no selective pressure. Through abstract development, we show that the diploid model can be

specialized by using mask based mutation and crossover operators to Vose's infinite population model which is a haploid model. Computational simplifications due to reduction of diploid model to haploid model and application of Walsh transform are exploited in experimental simulation of model. Through the experiment we demonstrate convergence of finite diploid population to infinite population behavior, and the distance between finite population and infinite population decreases like $1/\sqrt{r}$.

An instance of RHS is focused if \mathcal{G} is continuously differentiable, and for every $\mathbf{p} \in \Lambda$ the sequence

$$\mathbf{p}, \mathcal{G}(\mathbf{p}), \mathcal{G}^2(\mathbf{p}), \dots$$

converges. In this case, \mathcal{G} is also called focused, and the path determined by following at each generation what τ is expected to produce will lead to some steady state ω

$$\mathcal{G}(\omega) = \lim_{n \rightarrow \infty} \mathcal{G}(\mathbf{p}) = \omega.$$

Such points ω are called fixed points of \mathcal{G} . When specialized to a simple GA (that is explained in Chapter 2), it turns out that \mathcal{G} is focused under certain conditions. However, under other conditions the sequence $\mathbf{p}, \mathcal{G}(\mathbf{p}), \mathcal{G}^2(\mathbf{p}), \dots$ converges to a periodic orbit which oscillates between fixed points of \mathcal{G}^2 (see [Vose \(1999\)](#)). If a finite population GA follows the infinite population GA closely, and if infinite populations oscillate under certain conditions, then the finite population should also show oscillating behavior. We investigate finite population oscillation with experiments in Chapter 3. We use computation of predicted limits of infinite population and present necessary and sufficient conditions stated by Vose for population to converge in to periodic orbits. We compute mutation distribution $\boldsymbol{\mu}$ and crossover distribution $\boldsymbol{\chi}$ satisfying those conditions. Then, we study short-term behavior of finite population with respect to the limits to observe if oscillation occurs.

Consider the lattice points in the simple Λ which represent finite populations (for some fixed population size r) and let \mathbf{P}_j denote j th population represented by the j th lattice point. Let $\boldsymbol{\pi}^k$ be the probability vector having j th component equal to the

probability that k th generation is the population \mathbf{P}_j . If $\boldsymbol{\pi}^0$ is the initial distribution and Q is transition matrix, the steady state distribution $\boldsymbol{\pi}$, a row vector, is given by (see [Wikipedia \(2016c\)](#))

$$\lim_{k \rightarrow \infty} \boldsymbol{\pi}^k = \boldsymbol{\pi}^0 Q^k \quad (1.8)$$

where $\boldsymbol{\pi}^k$ can be interpreted as relative proportion of time that a GA spends in population \mathbf{P}_j . Solution to the equation 1.8 is

$$\boldsymbol{\pi} = \boldsymbol{\pi} Q. \quad (1.9)$$

where $\boldsymbol{\pi}$ is the normalized multiple of eigenvector of the transition matrix Q^T with eigenvalue 1. If GA were to oscillate between two populations \mathbf{P}_i and \mathbf{P}_j , then $\pi_i^k = 1$ (other components are 0) when k is odd, and $\pi_j^k = 1$ (other components are 0) when k is even.

If GA forms irreducible and aperiodic Markov chain, then the chain is ergodic (see [Wikipedia \(2016c\)](#)). For ergodic Markov chain, the steady state exists (it need not exist in general). If transition matrix Q is irreducible and non-negative, eigenvector associated to the Q is positive (see [Minc \(1988\)](#)). Hence, steady state distribution $\boldsymbol{\pi}$ for ergodic Markov chain converges to some vector with positive values in its component, meaning oscillation does not occur in GA if Markov chain is ergodic. In second part of Chapter 3, we further investigate if finite populations approximately oscillate or not if ergodic Markov chain is formed. To study this, we make some modifications in $\boldsymbol{\mu}$ and $\boldsymbol{\chi}$ distributions to violate the conditions necessary for population to oscillate such that GA will form ergodic Markov chain, and we observe finite population short term behavior with this violation.

Chapter 2

Extending A Genetic Algorithm Model To The Diploid Case

This chapter describes a simple Markov model for evolution under the influence of crossing over and mutation; it is a non-overlapping, generational, infinite population model under the assumption of *complete panmixia* (random mating) and no selective pressure. This chapter shows how diploid evolution equations can be represented by haploid equations and can be specialized to Vose's infinite population model, which is a haploid model.

A basic syntactic model for haploid and diploid genomes is first considered. Then the mechanics of how next generation is obtained from current generation are defined abstractly in procedural terms, which serves to motivate the equations governing evolution. Next evolution equations are developed corresponding to the procedural description defining evolution for a population of diploid genomes. Observations concerning the form and symmetry of those equations directly lead to decoupling from the diploid case a haploid model sufficient to determine evolutionary trajectories for the diploid case. Mask based mutation and crossover operators are used to specialize haploid equations to Vose's infinite haploid population model. Analytical and computational simplification resulting from specialization to Vose's infinite population

model are explained and used in experimental simulation and study of convergence of finite population short-term behavior to behavior predicted by infinite population model.

2.1 Model

A haploid genome g is defined syntactically as a length ℓ binary string. A collection of h chromosomes may be modeled by partitioning g into h segments (of arbitrary lengths ℓ_1, \dots, ℓ_h ; thus $\ell = \ell_1 + \dots + \ell_h$).

A diploid genome $\alpha = \langle \alpha_0, \alpha_1 \rangle$ is likewise defined syntactically as a pair of length ℓ binary strings. Although simple, that syntax is flexible and possesses significant modeling power by means of tailoring partitioning to application. We concentrate on the abstract level, considering the evolution of a non-overlapping, generational, infinite population model assuming panmixia and no selective pressure. We are not concerned with whether and how partitioning is defined as it is irrelevant to the development.

Following Hardy (see [Hardy \(1908\)](#)), the model q^n at generation n is a vector having for component q_α^n the prevalence of diploid α (the probability of selecting α at generation n , assuming unbiased selection).^{*} Ordered diploid $\gamma = \langle \gamma_0, \gamma_1 \rangle$ is produced for generation $n + 1$ according to following procedural description.

Assuming independent selection events:

- From parent α — selected with probability q_α^n — obtain gamete γ_0
- From parent β — selected with probability q_β^n — obtain gamete γ_1

Following Gieringer (see [Geiringer \(1944\)](#)), let the transmission function $t_\alpha(g)$ be the probability that gamete g is produced from parental genome α . It follows from the

^{*}The representation here is the conceptual equivalent of Hardy's model.

above that the equation determining the next generation q^{n+1} is

$$q_{\gamma}^{n+1} = \sum_{\alpha} q_{\alpha}^n t_{\alpha}(\gamma_0) \sum_{\beta} q_{\beta}^n t_{\beta}(\gamma_1) \quad (2.1)$$

It should be appreciated that the Mendelian (see [Mendel \(1865\)](#)) laws of segregation[†] and independent assortment[‡] need not be respected by the transmission function.

The right hand side of (2.1) is invariant under interchange of the summation variables α and β , which is equivalent to interchanging γ_0 and γ_1 . This symmetry reflects the fact that which haploid of γ is designated as γ_0 is arbitrary,

$$q_{\langle\gamma_0, \gamma_1\rangle}^{n+1} = q_{\langle\gamma_1, \gamma_0\rangle}^{n+1}$$

The model corresponding to (2.1) is low-level in the sense that it regards $\langle\gamma_0, \gamma_1\rangle$ and $\langle\gamma_1, \gamma_0\rangle$ as distinct when $\gamma_1 \neq \gamma_0$. A higher-level model based on sets is easily obtained,

$$q_{\{\gamma_0, \gamma_1\}} = \begin{cases} 2q_{\langle\gamma_0, \gamma_1\rangle} & \text{if } \gamma_0 \neq \gamma_1 \\ q_{\langle\gamma_0, \gamma_1\rangle} & \text{otherwise} \end{cases}$$

which is in agreement with Hardy (see [Hardy \(1908\)](#)).

2.2 Reduction

Evolution equation (2.1) may be reduced to the haploid case. Its right hand side is the product of two summations; denote the first by $p_{\gamma_0}^{n+1}$ and the second by $p_{\gamma_1}^{n+1}$ so that

$$q_{\langle\gamma_0, \gamma_1\rangle}^{n+1} = p_{\gamma_0}^{n+1} p_{\gamma_1}^{n+1} \quad (2.2)$$

[†] Alleles of a given locus segregate into separate gametes.

[‡] Alleles of one gene sort into gametes independently of the alleles of another gene.

where for any haploid γ_0 ,

$$p_{\gamma_0}^{n+1} = \sum_{\alpha} q_{\alpha}^n t_{\alpha}(\gamma_0) \quad (2.3)$$

It suffices to determine the evolution of the distributions p^n . Uncoupling p from q using (2.3), and equation (2.2) with superscript n — instantiate the n in (2.2) with $n - 1$ — yields the evolution equation

$$\begin{aligned} p_{\gamma_0}^{n+1} &= \sum_{\alpha_0, \alpha_1} q_{\langle \alpha_0, \alpha_1 \rangle}^n t_{\langle \alpha_0, \alpha_1 \rangle}(\gamma_0) \\ &= \sum_{\alpha_0, \alpha_1} p_{\alpha_0}^n p_{\alpha_1}^n t_{\langle \alpha_0, \alpha_1 \rangle}(\gamma_0) \end{aligned} \quad (2.4)$$

The p^n are in fact distributions; summing equation (2.2) with superscript n yields

$$1 = \sum_{\alpha} q_{\alpha}^n = \sum_{\alpha_0, \alpha_1} p_{\alpha_0}^n p_{\alpha_1}^n = \left(\sum_{\alpha_0} p_{\alpha_0}^n \right)^2$$

The weighted count of haploid g in generation n is

$$\sum_{\alpha_0, \alpha_1} q_{\langle \alpha_0, \alpha_1 \rangle}^n ([g = \alpha_0] + [g = \alpha_1]) \quad (2.5)$$

$$= \sum_{\alpha_0, \alpha_1} p_{\alpha_0}^n p_{\alpha_1}^n [g = \alpha_0] + \sum_{\alpha_0, \alpha_1} p_{\alpha_0}^n p_{\alpha_1}^n [g = \alpha_1] \quad (2.6)$$

$$= 2p_g^n \quad (2.7)$$

Hence the (normalized) prevalence of haploid g in generation n is the g th component of the distribution p^n . Moreover, (2.5) and (2.2) show (for $n > 0$) invertibility of the map

$$\pi : \mathbf{q}^n \longmapsto \mathbf{p}^n$$

Evolution equation (2.4) in matrix form is

$$p'_g = p^T M_g p \quad (2.8)$$

where current state p (generation n) and next state p' (generation $n + 1$) are column vectors, and the g th transmission matrix is

$$\left(M_g\right)_{u,v} = t_{\langle u,v \rangle}(g) \quad (2.9)$$

(vectors and matrices are indexed by haploids — length ℓ binary strings).

2.3 Specialization

This section summarizes from the development in Vose (see [Vose \(1999\)](#)). It specializes the haploid evolution equations in the previous section to a context where mask-based crossing over and mutation operators are used, leading to Vose’s infinite population model for Genetic Algorithms. Whereas in previous sections *component* referred to a component of a distribution vector q^n or p^n , in this section a component is either a probability (when when speaking of a component of a distribution vector), or a bit (when speaking of a component of a haploid).

2.3.1 Mutation

Mutation simulates low probability errors in chromosome duplication. Mutation provides a mechanism to inject new strings into the next generation. The symbol $\boldsymbol{\mu}$ denotes mutation distribution describing the probability $\boldsymbol{\mu}_i$ with which $i \in \Omega$ is selected to be a mutation mask. The result of mutating g is $g + i$ with probability $\boldsymbol{\mu}_i$. Mutating g using mutation mask i alters the bits of g in those positions the mutation mask i is 1. If g should mutate to g' with probability ρ , let

$$\boldsymbol{\mu}_{g+g'} = \rho$$

Given distribution $\boldsymbol{\mu}$, mutation is the stochastic operator sending g to g' with probability $\boldsymbol{\mu}_{g+g'}$. Abusing notation, $\boldsymbol{\mu} \in [0, 0.5)$ is regarded as a *mutation rate* which

implicitly specifies distribution $\boldsymbol{\mu}$ according to rule (see [Vose and Wright \(1998\)](#))

$$\mu_i = (\boldsymbol{\mu})^{1^T i} (1 - \boldsymbol{\mu})^{\ell - 1^T i}$$

2.3.2 Crossover

Crossover refers to crossing over (also termed recombination) between two chromosomes (strings in our case). Crossover like mutation also provides mechanism for injection of new strings into new generation population. Geiringer (see [Geiringer \(1944\)](#)) used crossover masks to generate offsprings from parent chromosomes in absence of mutation and selection. Let $\boldsymbol{\chi}_m$ be the probability distribution with which m is selected to be a crossover mask. Following Geiringer (see [Geiringer \(1944\)](#)), if crossing over u and v should produce u' and v' with probability ρ , let

$$\boldsymbol{\chi}_m = \rho$$

where m is 1 at components which u' inherits from u , and 0 at components inherited from v . It follows that

$$\begin{aligned} u' &= mu + \overline{m}v \\ v' &= mv + \overline{m}u \end{aligned}$$

Given distribution $\boldsymbol{\chi}$, crossover is the stochastic operator which sends u and v to u' and v' with probability $\boldsymbol{\chi}_m/2$.

Abusing notation, $\boldsymbol{\chi}$ can be considered as a *crossover rate* that specifies the distribution $\boldsymbol{\chi}$ given by rule (see [Vose and Wright \(1998\)](#))

$$\boldsymbol{\chi}_i = \begin{cases} \boldsymbol{\chi}^{c_i} & \text{if } i > 0. \\ 1 - \boldsymbol{\chi} + \boldsymbol{\chi}^{c_0} & \text{if } i = 0. \end{cases}$$

where $c \in \Lambda$ is referred to as *crossover type*. Classical crossover types include *1-point crossover* and *uniform crossover*. For *1-point crossover*,

$$c_i = \begin{cases} 1/(\ell - 1) & \text{if } \exists k \in (0, \ell). i = 2^k - 1. \\ 0 & \text{otherwise.} \end{cases}$$

and for uniform crossover, $c_i = 2^{-\ell}$.

2.3.3 Mixing Matrix

The combined action of mutation and crossover is referred to as *mixing*. The *mixing matrix* M is the transmission matrix corresponding to the additive identity of \mathcal{R} is

$$M = M_{\mathbf{0}}$$

Crossover and mutation are defined in a manner respecting arbitrary partitioning and arbitrary linkage to preserve the ability to endow abstract syntax with specialized semantics. Groups of loci can mutate and crossover with arbitrarily specified probabilities as discussed in above sections. For mutation distribution $\boldsymbol{\mu}$ and crossover distribution $\boldsymbol{\chi}$, then transmission function can be expressed as (see [Vose and Wright \(1998\)](#))

$$t_{\langle u, v \rangle}(g) = \sum_{i \in \mathcal{R}} \sum_{j \in \mathcal{R}} \sum_{k \in \mathcal{R}} \boldsymbol{\mu}_i \boldsymbol{\mu}_j \frac{\boldsymbol{\chi}_k + \boldsymbol{\chi}_{\bar{k}}}{2} [k(u + i) + \bar{k}(v + j) = g] \quad (2.10)$$

Here a child gamete g is produced via mutation and then crossover (which are operators that commute).

The mixing matrix M is a fundamental object, because (2.10) implies that evolution equation (2.8) can be expressed in the form

$$p'_g = (\sigma_g p)^T M (\sigma_g p) \quad (2.11)$$

where the permutation matrix σ_g is defined by component equations

$$(\sigma_g)_{u,v} = [u + v = g]$$

2.4 Walsh Transform

If $n \in \mathcal{R}$ and $t \in \mathcal{R}$, and N is cardinality of \mathcal{R} , the Walsh matrix is defined by

$$W_{n,t} = N^{-1/2}(-1)^{n^T t}$$

where $N^{-1/2}$ is normalization factor.

The matrix is symmetric, i.e.,

$$W_{n,t} = W_{t,n}$$

and it has entries satisfying

$$W_{n,t+k} = N^{1/2} W_{n,t} W_{n,k}; \quad k \in \mathcal{R}.$$

The practical importance of this symmetry is that the transform and inverse are the same mathematical operation, and *Walsh matrix* is its own inverse,

$$W = W^{-1}.$$

Given vector w and matrix A , let \hat{w} and \hat{A} denote the Walsh transform of w and A respectively. Then $\hat{w} = Ww$ and $\hat{A} = WAW$ (see [Beauchamp \(1975\)](#)).

2.4.1 Fast Walsh Transform

Computation of the discrete Walsh transform given by equation (??) might take n^2 operations (addition or subtraction) if implemented naively. An algorithm using

matrix factorization techniques is found to perform the transformation in $n \log_2 n$ operations. This algorithm is the Fast Walsh transform (FWT). Shanks (see [Shanks \(1969\)](#)) described FWT algorithm which is analogous to Cooley-Tukey algorithm (see [Cooley and Tukey \(1965\)](#)) for fast Fourier transformation. The algorithm for FWT can be translated into pseudocode as:

Algorithm 1 FWT pseudocode

```

1: procedure FWT
2:    $n = 2^d \leftarrow$  size of array  $X$  where  $d$  is positive integer
3:   for  $i = 0$  to  $d - 1$  do
4:      $m = n/2^i$ 
5:      $z = m/2$ 
6:     for  $j = 0$  to  $2^i - 1$  do
7:       for  $k = 0$  to  $z - 1$  do
8:          $t1 = m \times j + k$ 
9:          $t2 = m \times j + z + k$ 
10:         $a = X[t1]$ 
11:         $b = X[t2]$ 
12:         $X[t1] = a + b$ 
13:         $X[t2] = a - b$ 
14:      end for
15:    end for
16:  end for
17:  return  $X$ 
18: end procedure

```

2.4.2 Walsh Transform Adaptation

We adapt Walsh transform methods which have already been established for Vose's haploid model (see [Vose and Wright \(1998\)](#)) for computing evolutionary trajectories. Adaptation of Walsh transformation efficiently models infinite diploid population evolution, making feasible comparisons between finite and infinite diploid population short-term evolutionary behavior. Evolution equation (2.11), specialized to Vose's infinite population model without selection is

$$p'_g = (\sigma_g p)^T M (\sigma_g p)$$

where the permutation matrix σ_g is defined by component equations

$$(\sigma_g)_{u,v} = [u + v = g]$$

The Walsh matrix W is defined by the component equations

$$W_{u,v} = 2^{-\ell/2} (-1)^{u^T v}$$

where the subscripts u, v (which belong to \mathcal{R}) on the left hand side are interpreted on the right hand side as column vectors in \mathbb{R}^ℓ . Columns of W form the orthonormal basis — the *Walsh basis* — which simultaneously diagonalizes the σ_g . Expressed in the Walsh basis (see [Vose and Wright \(1998\)](#)), the mixing matrix takes the form

$$\widehat{M}_{u,v} = 2^{\ell-1} [uv = \mathbf{0}] \widehat{\mu}_u \widehat{\mu}_v \sum_{k \in \overline{u+v}\mathcal{R}} \chi_{k+u} + \chi_{k+v} \quad (2.12)$$

and equation (2.11) takes the form

$$\widehat{p}'_g = 2^{\ell/2} \sum_{i \in g\mathcal{R}} \widehat{p}_i \widehat{p}_{i+g} \widehat{M}_{i,i+g} \quad (2.13)$$

where $g\mathcal{R} = \{gi \mid i \in \mathcal{R}\}$ (for any $g \in \mathcal{R}$).

The mapping from generation n to generation $n + 1$, determined in natural coordinates by equation (2.8) in terms of the transmission function (2.9), and given in Walsh coordinates by equation (2.13) in terms of the mixing matrix (2.12), is Markovian; the next state p' depends only upon the current state p . Let \mathcal{M} represent the mixing transformation,

$$p' = \mathcal{M}(p) \quad (2.14)$$

and let $\mathcal{M}^n(p)$ denote the n -fold composition of \mathcal{M} with itself; thus generation $n + 1$ is described by

$$p^{n+1} = \mathcal{M}^n(p^1)$$

where $p^1 = \pi(q^1)$. We have little to say about the matrix of the Markov chain corresponding to the mixing transformation \mathcal{M} , because it is uncountable; each state is a distribution vector p describing a population. However, that is not an obstacle to computing evolutionary trajectories; (2.14) can be computed in Walsh coordinates relatively efficiently via (2.12) and (2.13).

2.5 Distance

Let vector \mathbf{f} represent a finite diploid population; component \mathbf{f}_α is the prevalence of diploid α . Let the support $S_{\mathbf{f}}$ of \mathbf{f} be the set of diploids occurring in the population represented by \mathbf{f} ,

$$S_{\mathbf{f}} = \{\alpha \mid \mathbf{f}_\alpha > 0\}$$

Let \mathbf{q} similarly represent an infinite diploid population (see section 2.1). As points in $\mathbb{R}^{2^\ell \times 2^\ell}$, the Euclidean distance between \mathbf{f} and \mathbf{q} is

$$\|\mathbf{f} - \mathbf{q}\| = \sum_{\alpha}^{\frac{1}{2}} (\mathbf{f}_\alpha - \mathbf{q}_\alpha)^2$$

Whereas a naive computation of this distance involves $2^\ell \cdot 2^\ell$ terms, leveraging equation (2.2) can significantly reduce the number of terms involved. Note that

$$\|\mathbf{f} - \mathbf{q}\|^2 = \sum_{\alpha \notin S_{\mathbf{f}}} (\mathbf{f}_\alpha - \mathbf{q}_\alpha)^2 + \sum_{\alpha \in S_{\mathbf{f}}} (\mathbf{f}_\alpha - \mathbf{q}_\alpha)^2 \quad (2.15)$$

Using equation (2.2) — $\mathbf{q}_\alpha = \mathbf{p}_{\alpha_0} \mathbf{p}_{\alpha_1}$ (suppressing superscripts to streamline notation) — together with the fact that $\mathbf{f}_\alpha = 0$ in every term of the first sum above, the first

sum reduces to

$$\begin{aligned}
\sum_{\langle \alpha_0, \alpha_1 \rangle \notin S_f} (\mathbf{p}_{\alpha_0} \mathbf{p}_{\alpha_1})^2 &= \sum_{\langle \alpha_0, \alpha_1 \rangle} (\mathbf{p}_{\alpha_0})^2 (\mathbf{p}_{\alpha_1})^2 - \sum_{\langle \alpha_0, \alpha_1 \rangle \in S_f} (\mathbf{p}_{\alpha_0} \mathbf{p}_{\alpha_1})^2 \\
&= \sum_g (\mathbf{p}_g)^2 - \sum_{\alpha \in S_f} (\mathbf{q}_\alpha)^2
\end{aligned} \tag{2.16}$$

It follows from (2.15) and (2.16) that

$$\begin{aligned}
\|\mathbf{f} - \mathbf{q}\|^2 &= \sum_g (\mathbf{p}_g)^2 + \sum_{\alpha \in S_f} (\mathbf{f}_\alpha - \mathbf{q}_\alpha)^2 - \sum_{\alpha \in S_f} (\mathbf{q}_\alpha)^2 \\
&= \sum_g (\mathbf{p}_g)^2 + \sum_{\alpha \in S_f} \mathbf{f}_\alpha (\mathbf{f}_\alpha - 2\mathbf{q}_\alpha)
\end{aligned} \tag{2.17}$$

which involves $2^\ell + |S_f|$ terms, assuming that S_f is known as a byproduct of computing \mathbf{f} . Therefore, (2.17) computes distance between finite and infinite population efficiently.

2.6 Simplification

Computations in the haploid case are simplified by equations (2.12) and (2.13) which follow from specializing to Vose's infinite population model and computing in the Walsh basis. Time switching between the standard basis and the Walsh basis is negligible; the fast Walsh transform (in dimension n) has complexity $n \log n$ [Shanks \(1969\)](#).

Only one mixing matrix as opposed to 2^ℓ matrices is needed to compute the next generation; evolution equation (2.13) references the same matrix for every g , whereas evolution equation (2.8) depends upon a different matrix M_g for each choice of g . The matrix is computed by a single sum as opposed to a triple sum; compare equation (2.12) with equation (2.10). Also, the relevant quadratic form is computed with a single sum as opposed to a double sum; computing via (2.13) is linear time in the

size of $g\mathcal{R}$ (for each g) as opposed to the quadratic time computation (for each g) represented by equation (2.8).

From a computational standpoint, the best-case scenario is where recomputation of the matrices mentioned in the previous paragraph is obviated by sufficient memory. The reduction from 2^ℓ matrices to one matrix helps significantly in that regard. To demonstrate this advantage in concrete terms, consider genomes of length $\ell = 14$. Using 2^{14} matrices each of which contains $2^{14} \times 2^{14}$ entries of type `double` requires 32 terabytes, whereas the mixing matrix at 2 gigabytes fits easily within the memory of a laptop. Moreover, for a population size of $N \leq 2^{20}$, the distance computation described in the previous section reduces the number of terms involved by a factor of $2^{28}/(2^{14} + 2^N) > 252$.

2.7 Convergence

This section presents a cursory numerical investigation of the convergence of finite diploid population short-term behaviour to that of the infinite diploid population model as described in section 2 (the underlying haploid model for the infinite population case is described in section 2.1).

Equations (2.2), (2.12), (2.13), (2.17) were employed to efficiently compute the distance

$$d = \|\mathbf{f}^n - \mathbf{q}^n\|$$

where \mathbf{f}^n and \mathbf{q}^n represent finite and infinite diploid populations (respectively) at generation $n \in \{1, 2, 4, 8, 16, 32, 64, 128\}$, beginning from a random initial population ($\mathbf{f}^0 = \mathbf{q}^0$). Genome lengths $\ell \in \{4, 6, 8, 10, 12, 14\}$ and population sizes $N = 2^i$ for integer $0 \leq i \leq 20$ were considered. The crossover distribution χ corresponds to independent assortment of bits, and the mutation distribution μ corresponds to

independent bit mutation probability 0.001,

$$\chi_m = 2^{-\ell}, \quad \mu_g = (0.001)^{\mathbf{1}^T g} (0.999)^{\ell - \mathbf{1}^T g}$$

(subscripts above on the left hand side of an equality are interpreted on the right hand side of the equality as column vectors in \mathbb{R}^ℓ). The finite population case is computed using the itemized procedural definition given in section 2.1; the transmission function (2.10) corresponds to μ and χ above (bits mutate independently and are freely assorted).

The data, presented in six surface graphs in **Figure 2.1** and organized by genome length, shows a near linear dependence of $\log d$ on $\log N$. As expected, the graphs show smoothing with increasing genome length (the computation of d involves averaging over ℓ components), and also with increased population size (as explained in Vose (1999), the initial transient of a finite haploid population trajectory converges as $N \rightarrow \infty$ to the corresponding infinite population model).

Of particular interest is the linear trend exhibited above. The slope m and intercept b of the regression line

$$\log d = m \log N + b \tag{2.18}$$

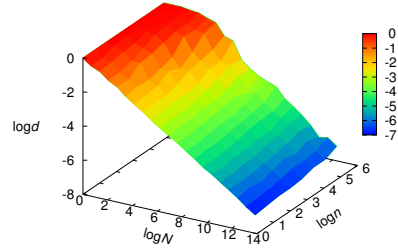
was computed using the data above; each was plotted against genome length ℓ and organized by generation n . The resulting graphs are displayed below.

Taking the exponential of the regression line (2.18) yields the estimate $d \approx N^m e^b$.

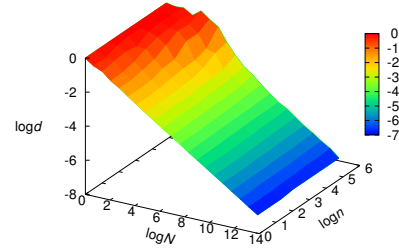
Slopes of the regression lines shown in **Figure 2.2** are approximately -0.5 , indicating

$$d \approx k/\sqrt{N}. \tag{2.19}$$

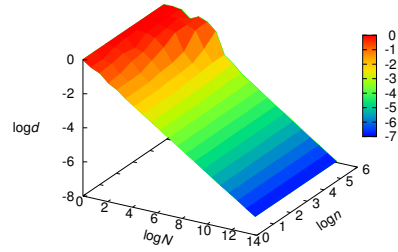
Equation 2.19 agrees with (1.3), (1.7) and theorem 3.1 from 'The Simple Genetic Algorithm: Foundations and Theory' (see Vose (1999)) which gives the bound for the expected rate of convergence for the single-step haploid case; the distance is inversely



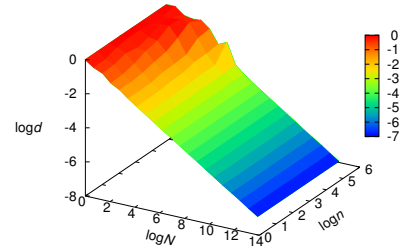
(a) $\ell = 4$.



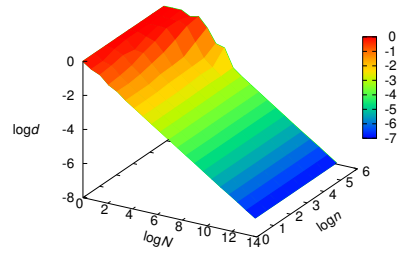
(b) $\ell = 6$.



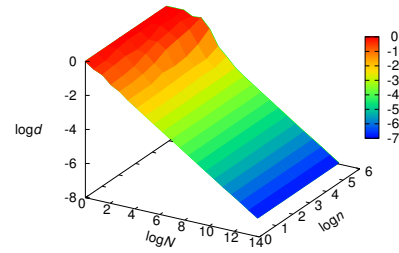
(c) $\ell = 8$.



(d) $\ell = 10$.

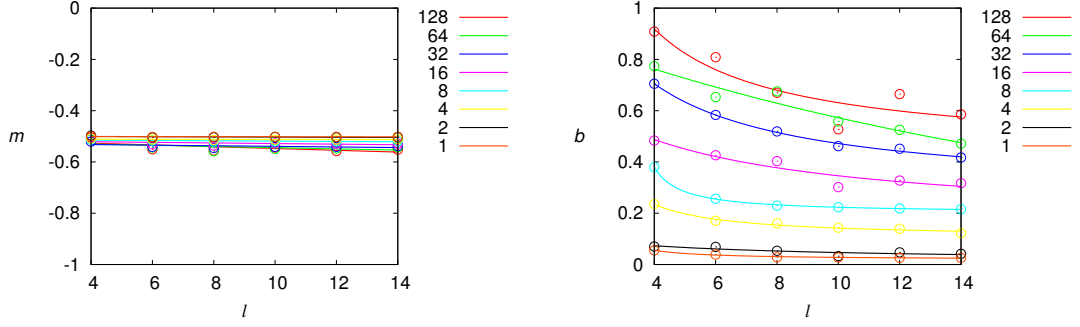


(e) $\ell = 12$.



(f) $\ell = 14$.

Figure 2.1: Convergence of finite population behaviour: d is distance between finite population \mathbf{f}^n and infinite population \mathbf{q}^n at generation n , population size N , for genome length ℓ (bits).



(a) Slope m , genome length ℓ .

(b) Intercept b , genome length ℓ .

Figure 2.2: Regression parameters: multi-plot of slope m and intercept b for generation $n \in \{1, 2, 4, 8, 16, 32, 64, 128\}$.

proportional to square root of population size. The consistent convergence rate across multiple generations shown in Figure (2.1) is somewhat surprising, simulation results above indicate it may persist to generation $n = 128$.

The intercept graphs in **Figure 2.2 b** show the constant of proportionality $k = e^b$ decreases monotonically with genome length ℓ , and increases monotonically with generation n . The increase in k for larger n seems to be a manifestation of the growing nonlinearity uniformly exhibited by the plots in **Figure 2.1** as n increases. It seems likely that the nonlinearity results from genetic drift experienced by finite populations (see [Crow and Kimura \(1970\)](#)).

2.8 Summary

In this chapter, we began with a description of simple diploid Markov model under mutation and crossover with no selective pressure. With reduction to haploid and specialization using masked base recombination operators, we showed Vose's infinite population model, can be extended to diploid case. Using computational benefits of this reduction (to haploid model and Walsh transform), we showed via experiment and regression of resulting data that distance between finite diploid population and

infinite diploid population decreases like $1/\sqrt{N}$ which is consistent with the single step-step haploid case convergence behavior predicted by Vose's infinite population model.

Chapter 3

Evolutionary Limits

This chapter investigates evolutionary limits predicted by Vose using infinite population model under no selective pressure. It uses computation to verify predicted infinite population limits and presents necessary and sufficient conditions obtained by Vose for convergence to periodic orbits. We investigate the convergence of finite population short-term behavior to infinite population evolutionary limits under no selective pressure. Infinite and finite population behavior when the necessary and sufficient conditions for convergence to periodic orbits are violated is also examined.

3.1 Limits

Vose states under mild assumptions on mutations (explained later), populations converge under repeated application of \mathcal{M} . Vose mentions that in the general case, periodic orbits are possible, but populations converge under repeated application of \mathcal{M}^2 and the limits $\mathbf{p}^* = \lim_{n \rightarrow \infty} \mathcal{M}^{2n}(\mathbf{p})$ and $\mathbf{q}^* = \lim_{n \rightarrow \infty} \mathcal{M}^{2n+1}(\mathbf{q})$ exist (see [Vose \(1999\)](#)).

In this section, operations ($+$ and \cdot) acting on elements of \mathcal{R} are component-wise addition and multiplication modulo 2.

Following Vose (see [Vose \(1999\)](#)), let $S_g = g\mathcal{R}/\{\mathbf{0}, g\}$, and let $|g|$ be the number of non zero bits in g .

$$\hat{\mathbf{p}}'_g = \begin{cases} 2^{\ell/2} & \text{if } g = 0 \\ x_g \hat{\mathbf{p}}_g + y_g(\hat{\mathbf{p}}_g) & \text{otherwise} \end{cases}$$

where,

$$x_g = 2\widehat{\mathcal{M}}_{g,0}, \quad y_g(z) = 2^{\ell/2} \sum_{i \in S_g} z_i z_{i+g} \widehat{\mathcal{M}}_{i,i+g}.$$

Moreover,

$$\begin{aligned} |g| &= 1 \Rightarrow y_g = 0 \\ |g| &> 0 \Rightarrow |x_g| \leq 1 \\ |x_g| &= 1 \Rightarrow y_g = 0 \end{aligned}$$

With above notations, limits can be expressed in Walsh basis by recursive equations (see [Vose \(1999\)](#))

$$\hat{\mathbf{p}}^*_g = \begin{cases} (x_g y_g(\hat{\mathbf{p}}^*) + y_g(\hat{\mathbf{q}}^*)) / (1 - x_g^2) & \text{if } |x_g| < 0 \\ \hat{p}_g & \text{otherwise} \end{cases} \quad (3.1)$$

$$\hat{\mathbf{q}}^*_g = \begin{cases} (x_g y_g(\hat{\mathbf{q}}^*) + y_g(\hat{\mathbf{p}}^*)) / (1 - x_g^2) & \text{if } |x_g| < 0 \\ \widehat{\mathcal{M}(\mathbf{p})}_g & \text{otherwise} \end{cases} \quad (3.2)$$

If $x_g \neq -1$ for all g , then $\mathbf{p}^* = \mathbf{q}^* = \lim_{n \rightarrow \infty} \mathcal{M}(\mathbf{p})$ is the limit of mixing. In other cases, mixing converges to a periodic orbit oscillating between \mathbf{p}^* and $\mathbf{q}^* = \mathcal{M}(\mathbf{p}^*)$.

Limits $\hat{\mathbf{p}}^*_g$ and $\hat{\mathbf{q}}^*_g$ can be computed considering g th components in order of increasing $|g|$. The necessary and sufficient condition for the sequence

$$\mathbf{p}, \mathcal{M}(\mathbf{p}), \mathcal{M}^2(\mathbf{p}), \dots$$

to converge to a periodic orbit given by Vose is that for some g

$$-1 = \sum_j (-1)^{g^T j} \mu_j = - \sum_{k \in \bar{g}\mathcal{R}} \chi_{k+g} + \chi_k \quad (3.3)$$

3.2 Computation of Mutation and Crossover Distribution

The following algorithm generates mutation and crossover distributions that satisfying equation (3.3) for evolution to converge to periodic orbits. Operations (+ and \cdot) acting on elements of \mathcal{R} in this section below are component-wise addition and multiplication modulo 2. Let μ_j and χ_k represent mutation and crossover distributions respectively where $j, k \in \mathcal{R}$ and $U01()$ is a random number between 0 and 1. For any $g \in \mathcal{R}$ and $g \neq 0$ and for all $j \in \mathcal{R}$,

$$\mu_j = \begin{cases} U01() & \text{if } (g^T \cdot j) \text{ is odd.} \\ 0 & \text{otherwise.} \end{cases}$$

This installs random values in some specific positions in μ according to value of g . Other components of μ are set to 0. Normalization of μ_j yields μ (the mutation distribution),

$$\mu_j = \mu_j / \sum_{j \in \mathcal{R}} \mu_j.$$

Hence

$$\sum_{j \in \mathcal{R}} \mu_j = 1.$$

Moreover, the values μ_j satisfy condition (3.3).

Condition $k \in \bar{g}\mathcal{R}$ in equation (3.3) can be simplified for computation as

$$k = \bar{g}i \text{ where } i \in \mathcal{R}$$

Logical bitwise ANDing both sides by \bar{g} ,

$$\bar{g}k = \bar{g}\bar{g}i = \bar{g}i = k$$

The crossover distribution can be generated as follows. For all $k \in \mathcal{R}$,

$$\begin{aligned}\chi_k &= U01() \\ \chi_{k+g} &= U01()\end{aligned}$$

where $k \in \bar{g}\mathcal{R}$, and

$$\chi_k = 0$$

for other values of k . This installs random values in some specific positions in χ according to the value of g . Other components of χ are set to 0. Normalization of χ_k yields χ (the crossover distribution),

$$\chi_k = \chi_k / \sum_{k \in \mathcal{R}} \chi_k.$$

Hence

$$\sum_{k \in \mathcal{R}} \chi_k = 1.$$

Moreover, the values χ_k satisfy condition (3.3).

3.3 Initial Population

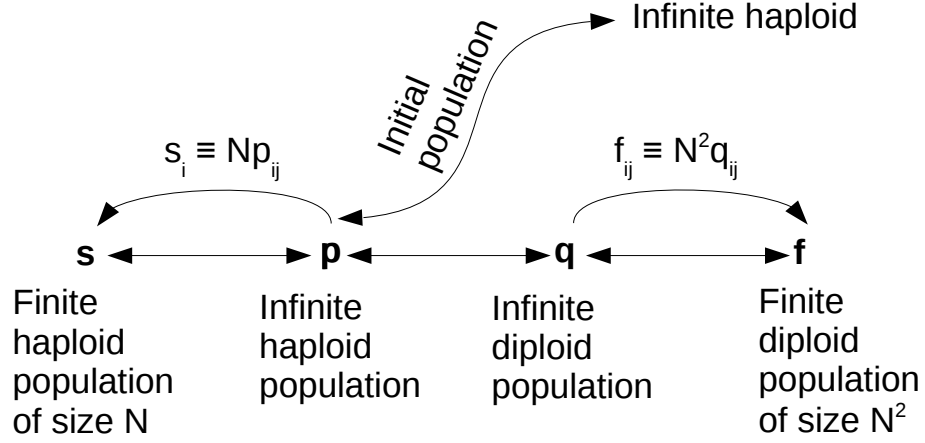


Figure 3.1: Initial population computation

Let finite haploid population \mathbf{s}^n , finite diploid population \mathbf{f}^n , infinite haploid population \mathbf{p}^n and infinite haploid population \mathbf{q}^n be considered with initial population \mathbf{s}^0 , \mathbf{f}^0 , \mathbf{p}^0 , \mathbf{q}^0 respectively. To investigate oscillating behavior of infinite population evolutionary limits and finite population behavior, it is desirable to have the same initial population.

For a length ℓ , $x = 2^\ell$ is the number of possible haploids. Let array \mathbf{t} represent a population size of N as follows: \mathbf{t}_j is the j th population member (some element of $\{0, \dots, x-1\}$ where elements are base 2 length ℓ binary strings). Array \mathbf{t} is generated as follows. First an arbitrary vector \mathbf{v} of size x is considered where

$$\mathbf{v}_i = U01(); \quad i = 0, 1, \dots, x-1$$

and $U01()$ is random number between 0 and 1.

$$\mathbf{t}_j = randp(\mathbf{v}); \quad j = 0, \dots, N-1$$

where $randp(\mathbf{v})$ returns random index i in array \mathbf{v} with probability \mathbf{v}_i .

Let \mathbf{c}_i represent count of haploid member i in population \mathbf{t} given by

$$\mathbf{c}_i = \sum_{j=0}^{N-1} [\mathbf{t}_j = i]; \quad i = 0, \dots, x-1$$

Then infinite population vector \mathbf{p} is calculated as

$$\mathbf{p}_i = \frac{\mathbf{c}_i}{\sum_{k=0}^{x-1} \mathbf{c}_k}$$

where $i = 0, \dots, x-1$ and $\sum_{k=0}^{x-1} \mathbf{c}_k = N$.

This \mathbf{p} is randomly generated initial infinite haploid population vector (\mathbf{p}^0) which corresponds to diploid infinite population vector \mathbf{q} and finite population vectors \mathbf{s} and \mathbf{f} .

Finite haploid population members \mathbf{t}_j s are generated again to match finite haploid population \mathbf{s}^0 with infinite haploid population \mathbf{p}^0 .

$$\mathbf{c}_i = N \cdot \mathbf{p}_i$$

$$\sum_{j=0}^{N-1} [\mathbf{t}_j = i] = \mathbf{c}_i; \quad i = 0, \dots, x-1$$

Initial infinite diploid population \mathbf{q}_0 is calculated corresponding to initial haploid population \mathbf{p}^0 as

$$\mathbf{q}_{i,j}^0 = \mathbf{p}_i^0 \cdot \mathbf{p}_j^0; \quad (0 \leq i, j < x).$$

Let \mathbf{v} represent finite diploid population member array of size N^2 and $\mathbf{d}_{i,j}$ represent count of diploid member $\langle i, j \rangle$ in \mathbf{v} . Then \mathbf{v} can be filled with population member to

match initial population vector \mathbf{p} generating diploid members such that

$$\begin{aligned} \mathbf{d}_{i,j} &= N \cdot \mathbf{p}_i \cdot N \cdot \mathbf{p}_j \\ \sum_{k=0}^{N^2-1} [\mathbf{v}_k = \langle i, j \rangle] &= \mathbf{d}_{i,j} \end{aligned}$$

Finite diploid population (proportion) vector \mathbf{f} can be obtained from finite diploid population member array \mathbf{v} using

$$f_{i,j} = \frac{\mathbf{d}_{i,j}}{\sum_{k=0}^{x-1} \sum_{h=0}^{x-1} \mathbf{d}_{k,h}}$$

where $i = 0, \dots, x-1$, $h = 0, \dots, x-1$ and $\sum_{k=0}^{x-1} \sum_{h=0}^{x-1} \mathbf{d}_{k,h} = N^2$.

Thus, initial infinite haploid population vector \mathbf{p}^0 corresponds to initial infinite diploid population vector \mathbf{q}^0 , initial finite haploid population vector with population size N and initial finite diploid population vector with population size N^2 .

3.4 Oscillation

Equations (3.1) and (3.2) were implemented with crossover distribution χ and mutation distribution μ satisfying condition (3.3) to investigate oscillating behavior of predicted infinite population evolutionary limits \mathbf{p}^* and \mathbf{q}^* and finite population under no selective pressure.

Infinite haploid population evolutionary limits \mathbf{p}_h^* and \mathbf{q}_h^* were computed using equations (3.1) and (3.2). Infinite diploid population evolutionary limits \mathbf{p}_d^* and \mathbf{q}_d^* as

$$\begin{aligned} (\mathbf{p}_d^*)_{\langle \gamma_0, \gamma_1 \rangle} &= (\mathbf{p}_h^*)_{\gamma_0} (\mathbf{p}_h^*)_{\gamma_1} \\ (\mathbf{q}_d^*)_{\langle \gamma_0, \gamma_1 \rangle} &= (\mathbf{q}_h^*)_{\gamma_0} (\mathbf{q}_h^*)_{\gamma_1} \end{aligned}$$

where $\gamma = \langle \gamma_0, \gamma_1 \rangle$ is diploid genome.

For every genome length ℓ , the same initial population (calculated as described in (3.3)) was used for the infinite population and all sizes of finite populations considered. Genome lengths $\ell \in \{8, 10, 12, 14\}$ were used. Base population size of $N_0 = 64$ was used for the finite haploid case to compute initial population vector. The population sizes considered for plotting graphs were $N \in \{1N_0^2, 10N_0^2, 20N_0^2\}$. The distances of \mathbf{p}^n and \mathbf{s}^n to haploid evolutionary limits \mathbf{p}_h^* and \mathbf{q}_h^* were plotted and the distances of \mathbf{q}^n and \mathbf{f}^n to diploid evolutionary limits \mathbf{p}_d^* and \mathbf{q}_d^* were plotted. Distance data of finite population to infinite population were also plotted.

Figures 3.2, 3.3, 3.4, 3.5, 3.6, 3.7, 3.8 and 3.9 arranged by genome length ℓ in ascending order. For each genome length, figures are split into two cases, haploid and diploid cases of population. In each figure for unique genome length ℓ and population case (either haploid or diploid), sub-figures are arranged by population size (N). In each figure, first three rows of sub-figures on left column shows distance of finite population to limits and sub-figure in fourth row on left column shows distance of infinite population to limits. These sub-figures depicts oscillating behavior of both infinite and finite population when necessary and sufficient condition 3.3 is met. Finite population oscillation in both the haploid and diploid case is sharper with increased population size. As population size increases, oscillation approaches the behavior exhibited by infinite population.

In each figure (3.2, 3.3, 3.4, 3.5, 3.6, 3.7, 3.8 and 3.9), first three rows of graphs on right side shows distance variation (difference in distance (d) and average distance (d_{avg})) where d is distance between finite and infinite populations and d_{avg} is average value of d . In fourth row on right, a single graph for distances (d) of different finite population sizes ($N = 1N_0^2, 10N_0^2, 20N_0^2$) to infinite population are plotted. The resulting graphs shows distance decreases as population size increases which is consistent with results from section 2.1. The distance graphs smooth as population size increases. The graphs of $d - d_{avg}$ decreases in amplitude as population size increases.

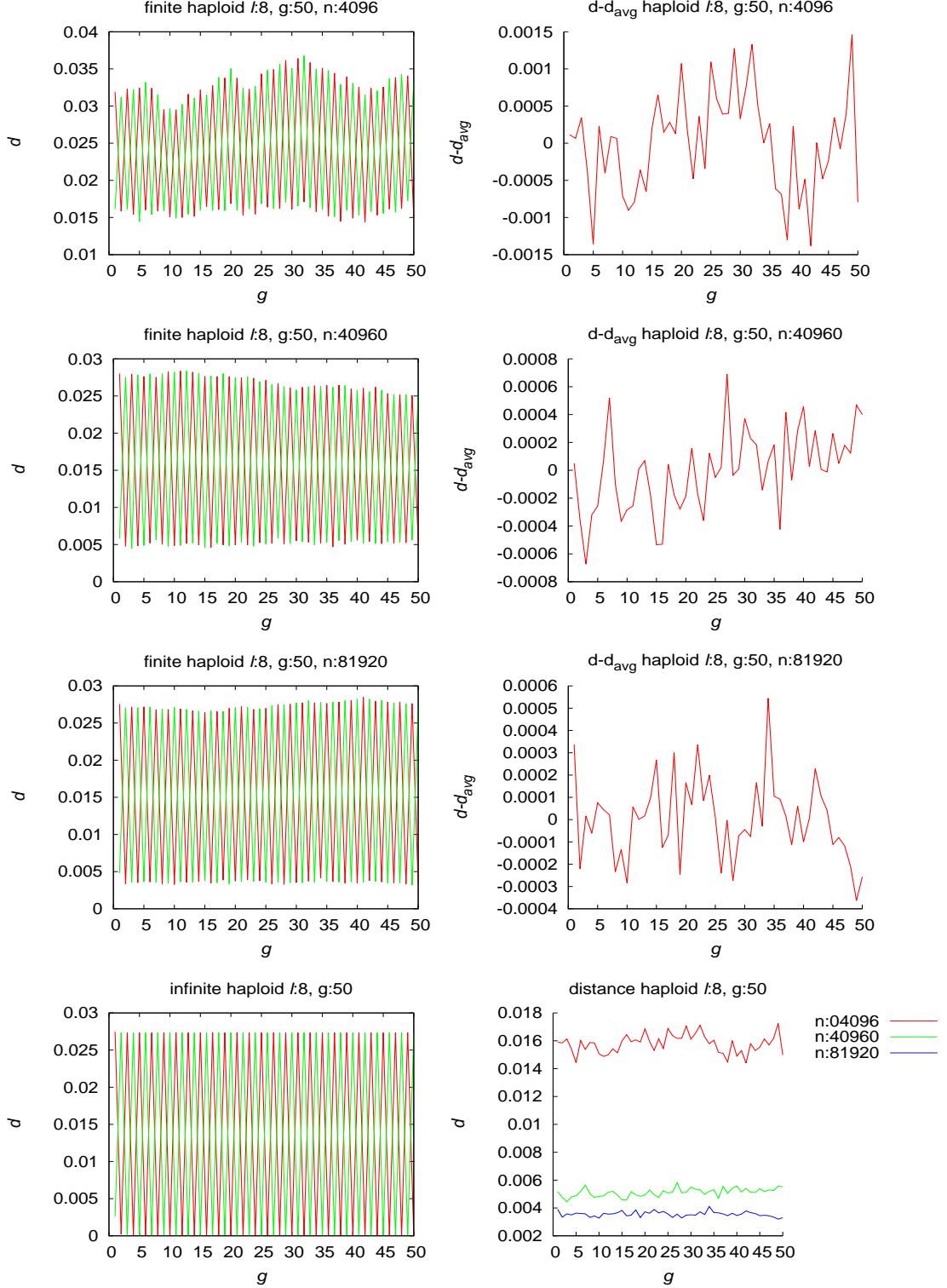


Figure 3.2: Infinite and finite haploid population oscillation behavior for genome length $\ell = 8$ (bits): In left column, d is distance of finite population of size n or infinite population to limits for g generations. In right column, d is distance of finite population to infinite population for g generations and d_{avg} is average of distance from 1 to 50 generations.

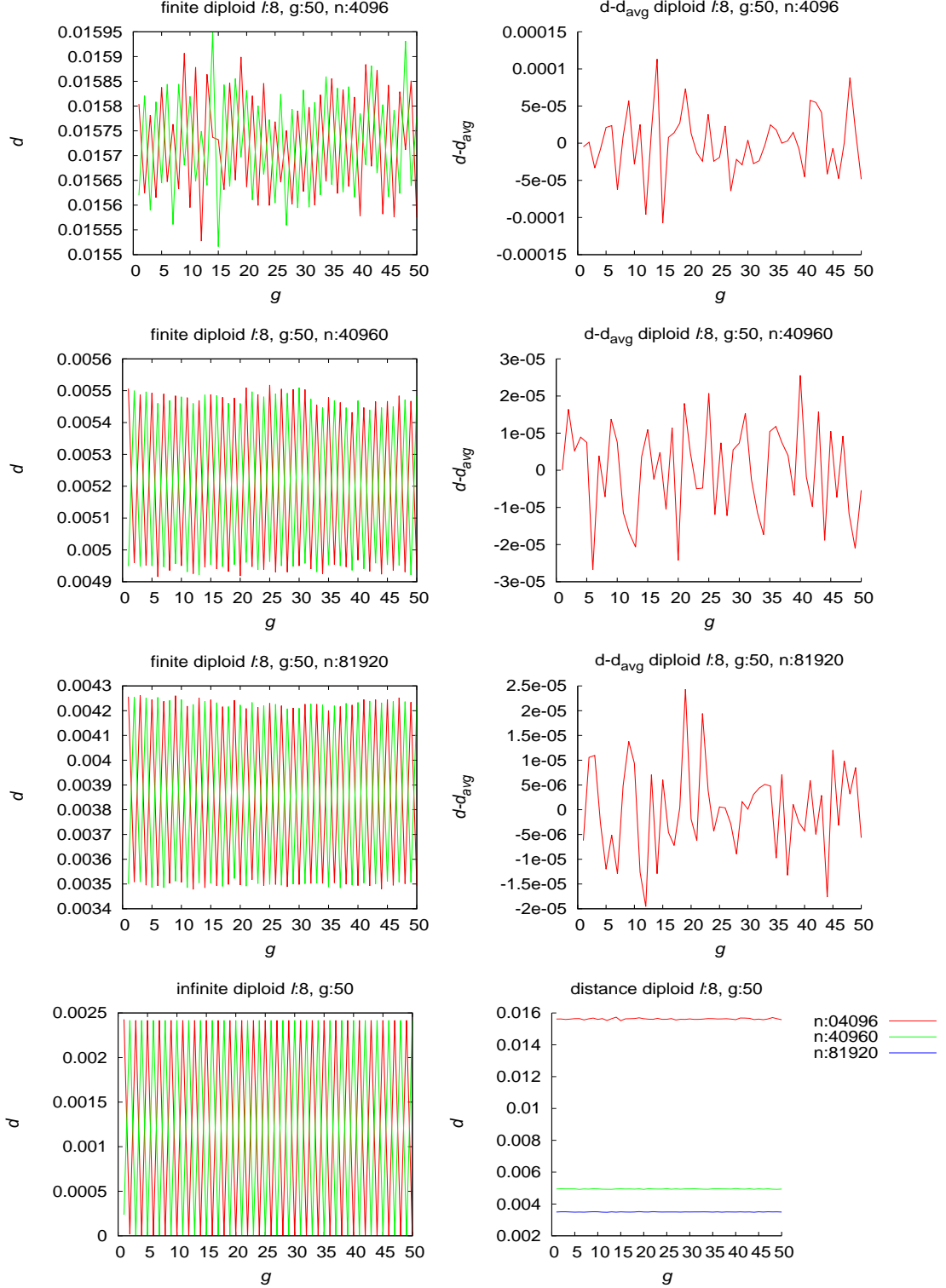


Figure 3.3: Infinite and finite diploid population oscillation behavior for genome length $\ell = 8$ (bits): In left column, d is distance of finite population of size n or infinite population to limits for g generations. In right column, d is distance of finite population to infinite population for g generations and d_{avg} is average of distance from 1 to 50 generations..

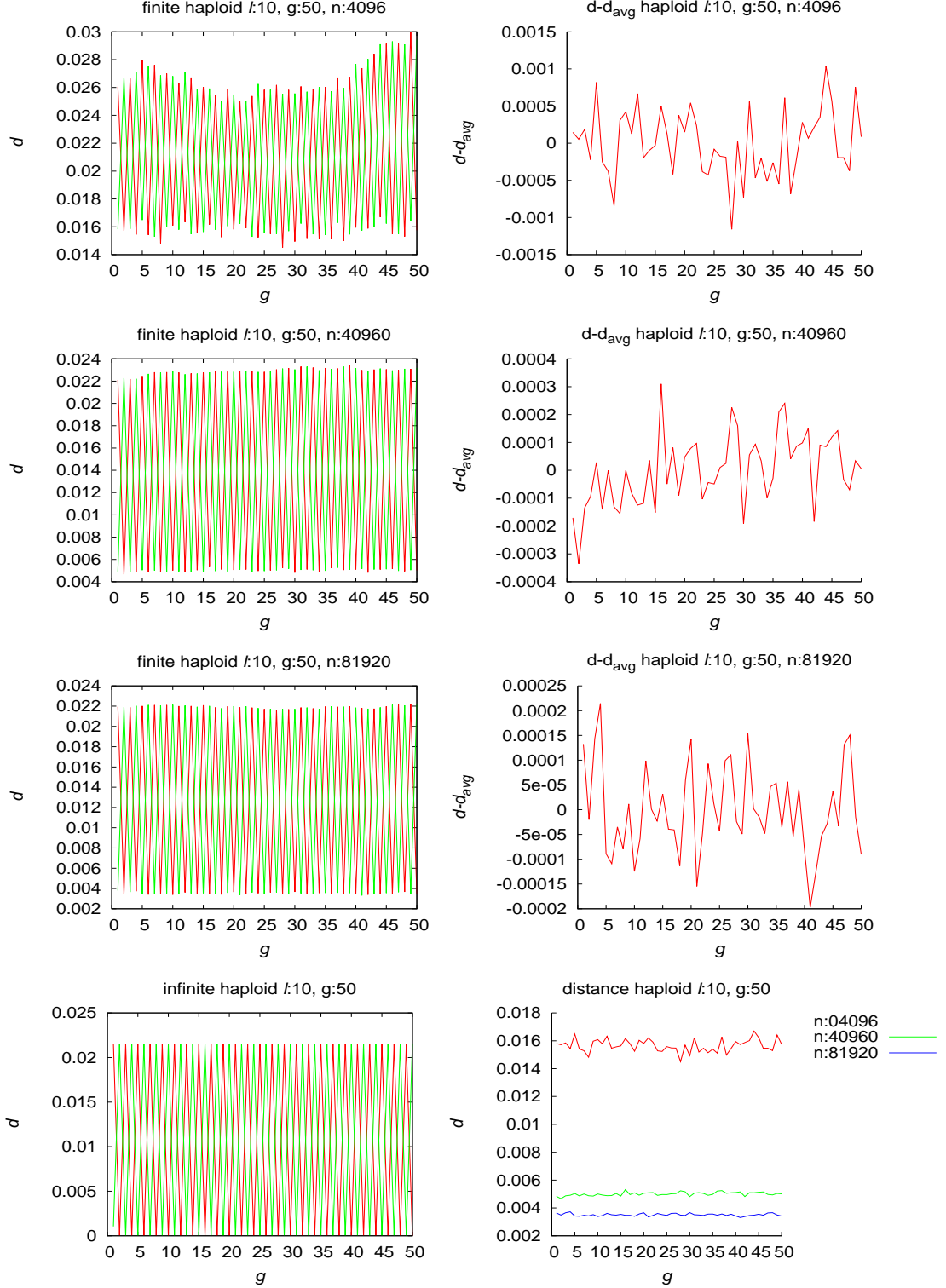


Figure 3.4: Infinite and finite haploid population oscillation behavior for genome length $\ell = 10$ (bits): In left column, d is distance of finite population of size n or infinite population to limits for g generations. In right column, d is distance of finite population to infinite population for g generations and d_{avg} is average of distance from 1 to 50 generations..

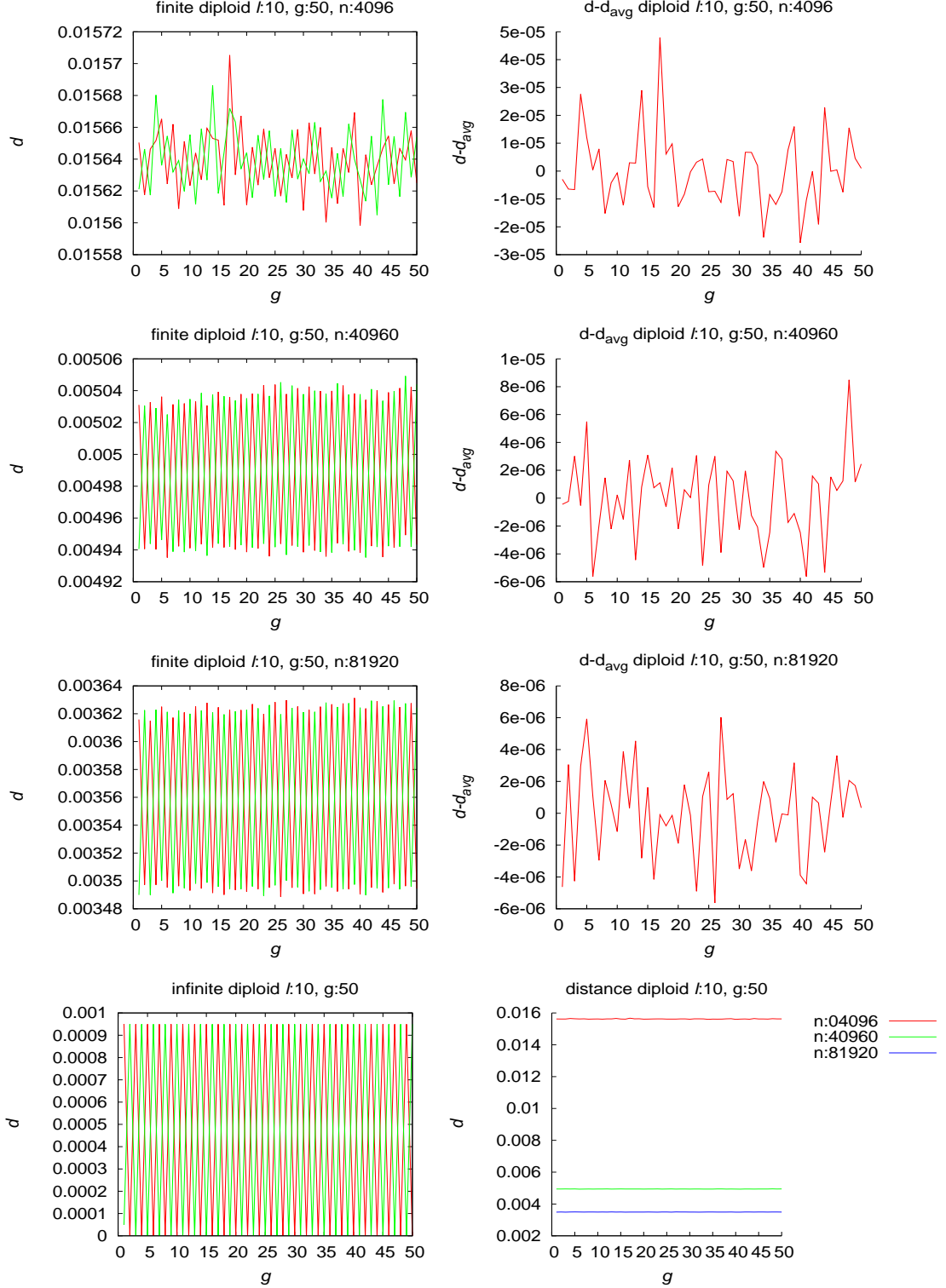


Figure 3.5: Infinite and finite population oscillation behavior for genome length $\ell = 10$ (bits): In left column, d is distance of finite population of size n or infinite population to limits for g generations. In right column, d is distance of finite population to infinite population for g generations and d_{avg} is average of distance from 1 to 50 generations..

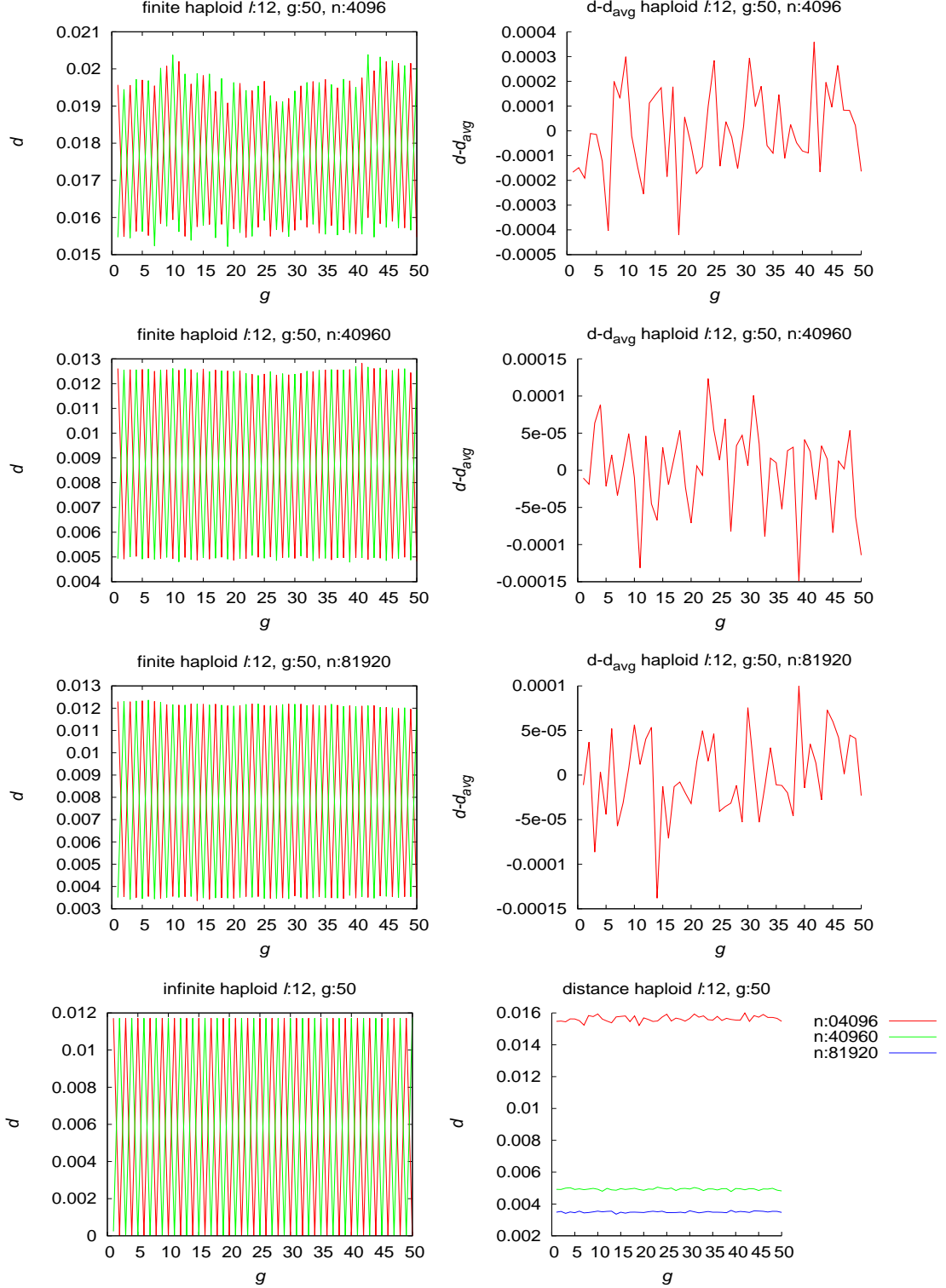


Figure 3.6: Infinite and finite haploid population oscillation behavior for genome length $\ell = 12$ (bits): In left column, d is distance of finite population of size n or infinite population to limits for g generations. In right column, d is distance of finite population to infinite population for g generations and d_{avg} is average of distance from 1 to 50 generations..

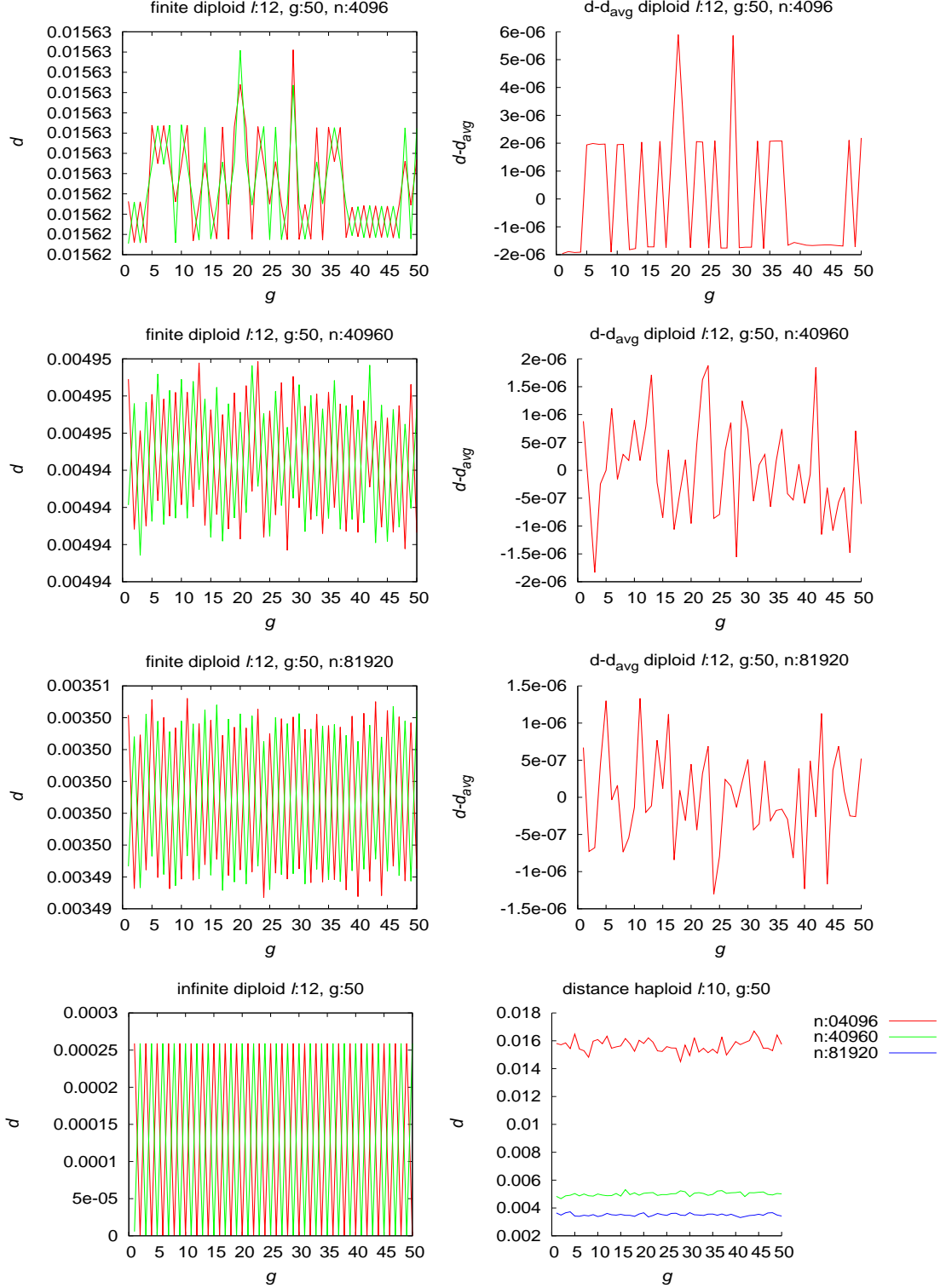


Figure 3.7: Infinite and finite diploid population oscillation behavior for genome length $\ell = 12$ (bits): In left column, d is distance of finite population of size n or infinite population to limits for g generations. In right column, d is distance of finite population to infinite population for g generations and d_{avg} is average of distance from 1 to 50 generations..

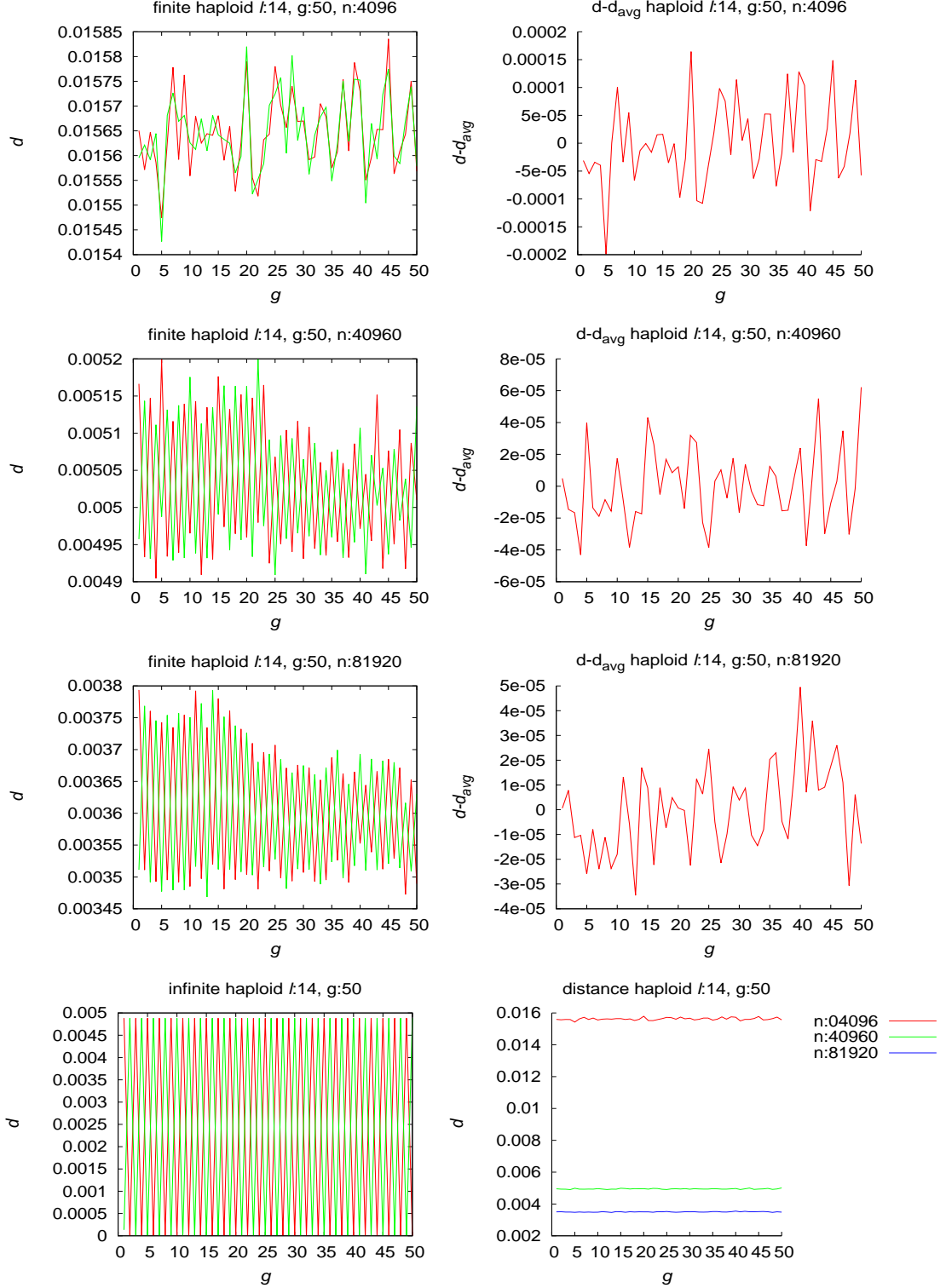


Figure 3.8: Infinite and finite haploid population oscillation behavior for genome length $\ell = 14$ (bits): In left column, d is distance of finite population of size n or infinite population to limits for g generations. In right column, d is distance of finite population to infinite population for g generations and d_{avg} is average of distance from 1 to 50 generations..

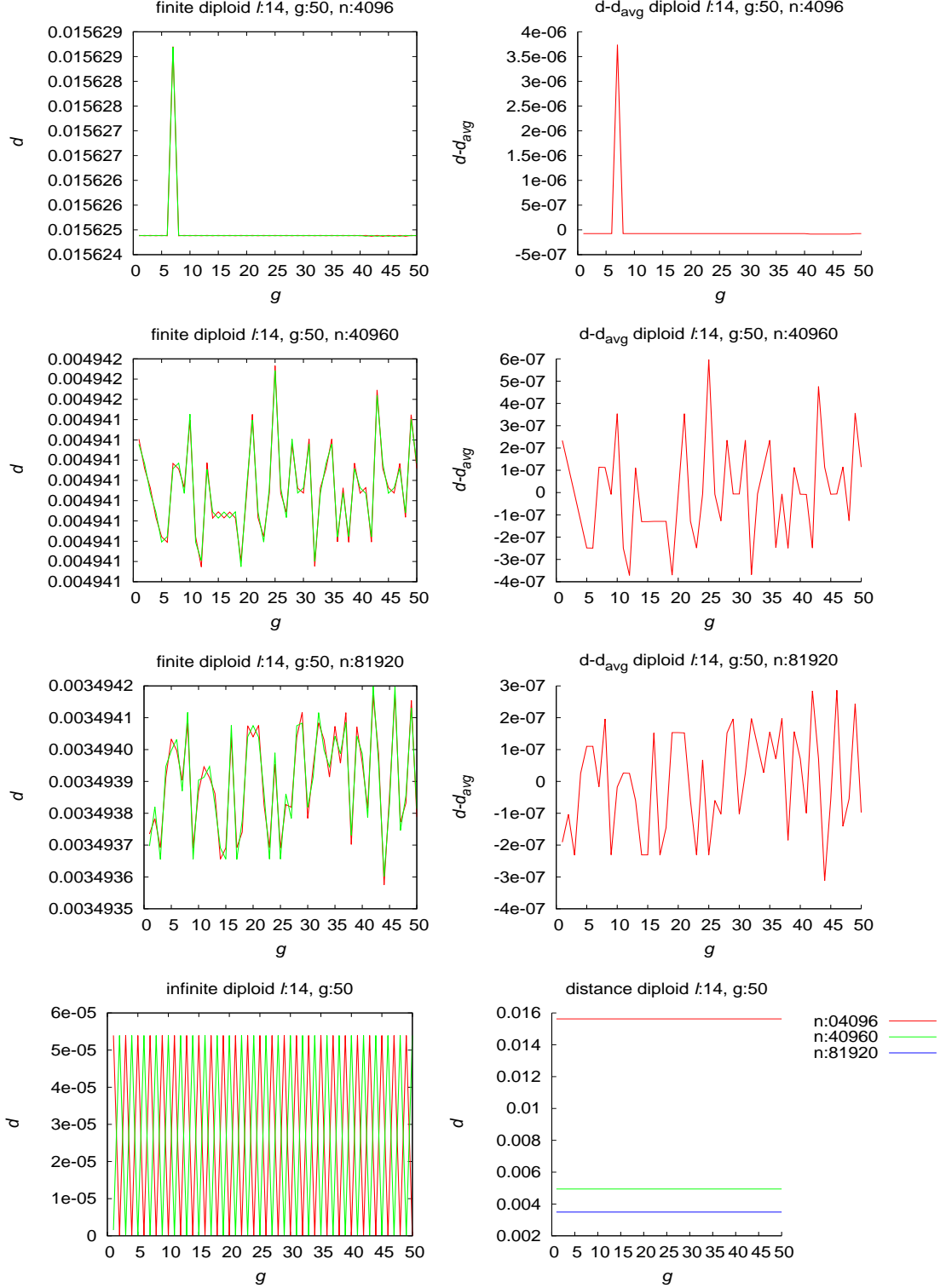


Figure 3.9: Infinite and finite diploid population oscillation behavior for genome length $\ell = 14$ (bits): In left column, d is distance of finite population of size n or infinite population to limits for g generations. In right column, d is distance of finite population to infinite population for g generations and d_{avg} is average of distance from 1 to 50 generations..

The numerator in equation ?? is approximately 1. So from ??, the expected single step distance between finite and infinite population, d , is

$$d \approx 1/\sqrt{N}$$

where N is the population size. This expected single step distance is shown in table 3.1. Distance data obtained from simulations are summarized in table 3.2. The last

Table 3.1: Expected single step distance d for population size N

N	4096	40960	81920
d	0.0156	0.0049	0.0035

three columns tabulate average distance values between finite and infinite population for population sizes $N = 4096$, $N = 40960$ and $N = 81920$ respectively. Results

Table 3.2: Experimental distance measured for oscillation: N is population size, ℓ is genome length and average distance between finite and infinite population is tabulated in the last three columns. {4096, 40960, 81920}

case	ℓ	$N = 4096$	$N = 40960$	$N = 81920$
haploid	8	0.0158	0.0051	0.0035
	10	0.0157	0.0050	0.0035
	12	0.0156	0.0049	0.0035
	14	0.0156	0.0049	0.0035
diploid	8	0.0156	0.0049	0.0035
	10	0.0156	0.0049	0.0035
	12	0.0156	0.0049	0.0035
	14	0.0156	0.0049	0.0035

from table 3.2 show average distance between finite and infinite population follows closely the expected single step distance given in table 3.1. The distance decreases as $1/\sqrt{N}$.

3.5 Violation

Previous results show that oscillation occurs when the crossover distribution χ , and the mutation distribution μ satisfy (3.3). Error ϵ was introduced to μ and χ distributions to violate condition (3.3). Consequently, $\mathbf{p}^* = \mathbf{q}^* = \mathbf{z}^*$. Going forward, we use 'limit \mathbf{z}^* ' to denote evolutionary limit when crossover distribution χ or mutation distribution μ is violated, and 'non-violation limits \mathbf{p}^* and \mathbf{q}^* ' to denote limits without violation.

3.5.1 Violation in Mutation Distribution

The mutation distribution μ was modified as follows

$$\mu_i = (1 - \epsilon)\mu_i; \quad i = \{0, 1, 2, \dots, 2^\ell - 1\}.$$

So that sum of μ distribution becomes,

$$1 - \epsilon = \sum_{i=0}^{2^\ell-1} \mu_i$$

Then set

$$\mu_0 = \epsilon$$

The mutation distribution μ is normalized such that $\sum_{i=0}^{2^\ell-1} \mu_i = 1$. This modification makes possible for any population member to mutate to any other possible population member via mutation. Let us explore for two cases of g in 3.3:

1. When g is all 1s:

Any mask j with only one 1 at any position k ($0 \leq k < \ell$) can mutate k th bit in string, and since mask all 0s in has positive probability, strings have option to not mutate. This gives possibility for any string to mutate to any possible string in population. Let us take an example with $\ell = 8$. Let $g = 11111111$. Then, mask $i = 00000100$

will have positive probability according to condition 3.3 implementation. Mask i can be used to mutate bit of a population string at position $k = 5$ from left.

2. When g has atleast one 0:

Any mask j with only one 1 at any position k ($0 \leq k < \ell$) where k th position in g also has 1 in it will have positive probability. So any string can mutate that particular k th bit using those masks. And any mask with 1 in any one of the position where g has one in it and 1 in any position where g has 0s in it can be used to mutate bits in string where g has 0s in it. Let us take an example with $\ell = 8$. Let $g = 11001111$. Then, mask $i = 00000100$ will have positive probability according to condition 3.3 implementation. Then, also mask $j = 00010100$ will have positive probability. Mask i can be used to mutate string at position $k = 5$. To mutate bit of any string at position $k = 3$, string can be mutated using mask i and mask j in sequence.

Population in state P can mutate to another state P' in next generation. Then it can mutate back to original state P or stay in its state P' or move to another state P'' in next generation. Since population state can mutate to any state, and any state can be returned to in any period of time, the Markov chain formed by GA is irreducible and aperiodic and hence, ergodic (see [Wikipedia \(2016c\)](#)). Because of ergodic Markov chain, steady state distribution with positive value components exists for the GA.

Simulations were repeated with the violations in (3.3) described above for mutation distribution. The distances of both infinite and finite populations to limit \mathbf{z}^* were plotted. The distances of both infinite and finite populations to non-violation limits \mathbf{p}^* and \mathbf{q}^* were also plotted.

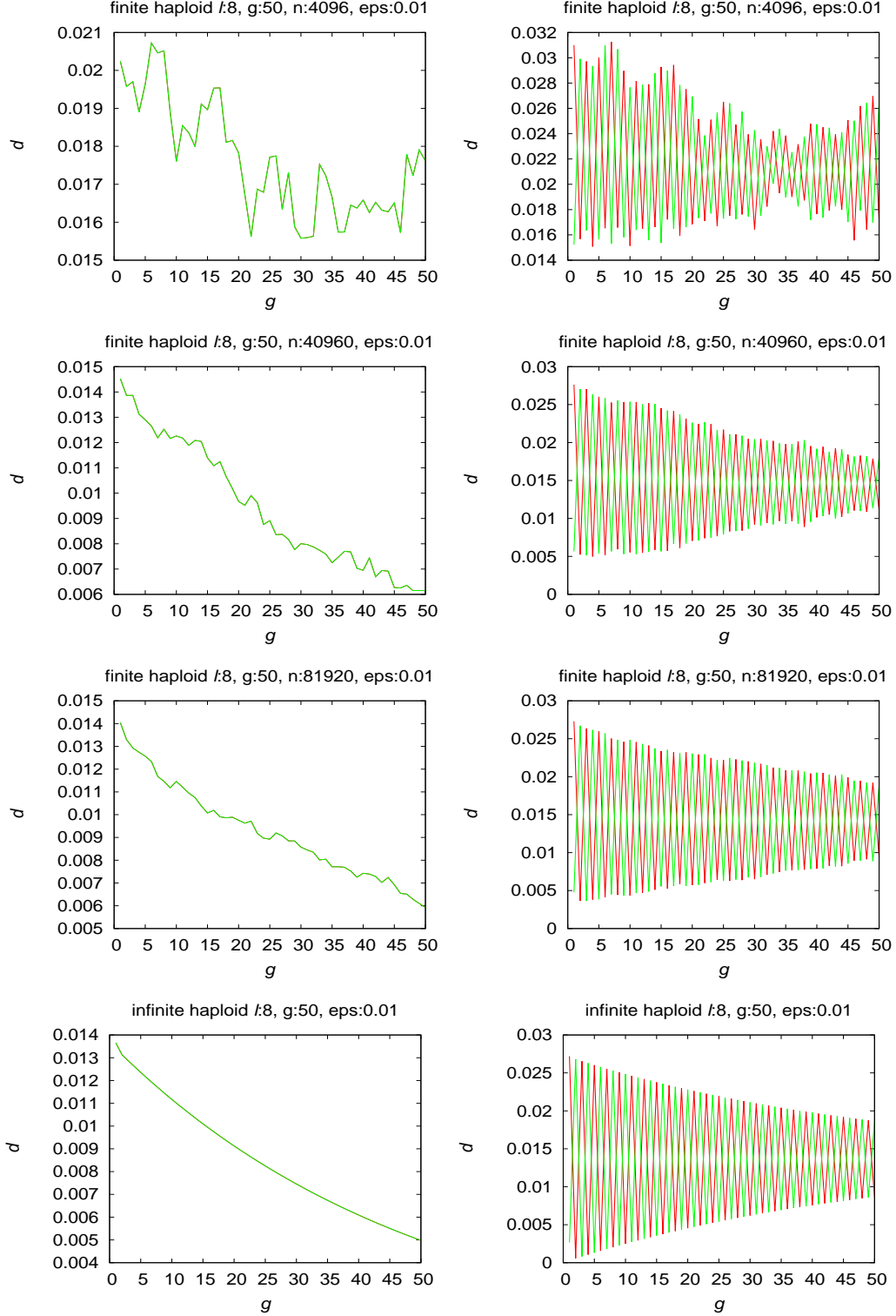


Figure 3.10: Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 8$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

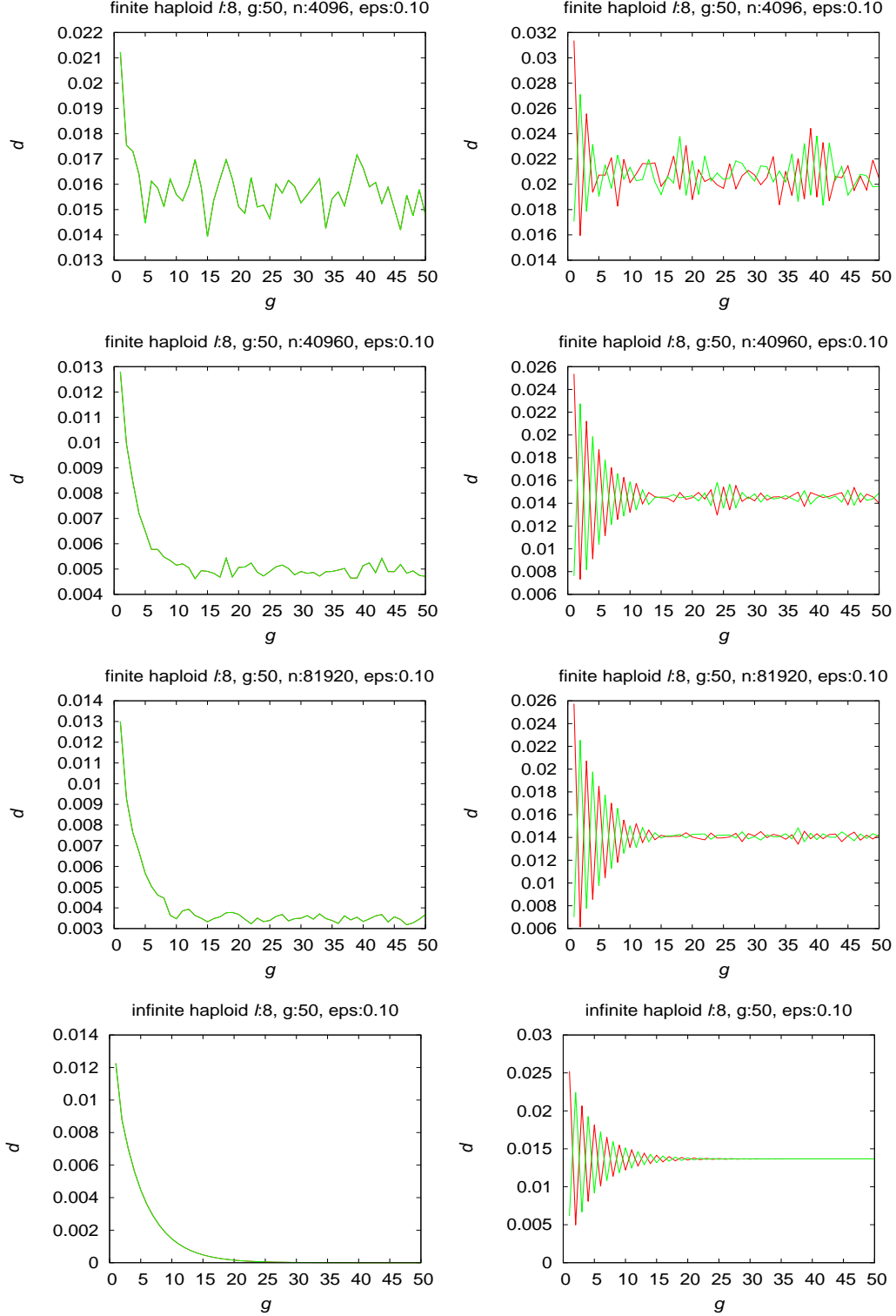


Figure 3.11: Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 8$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

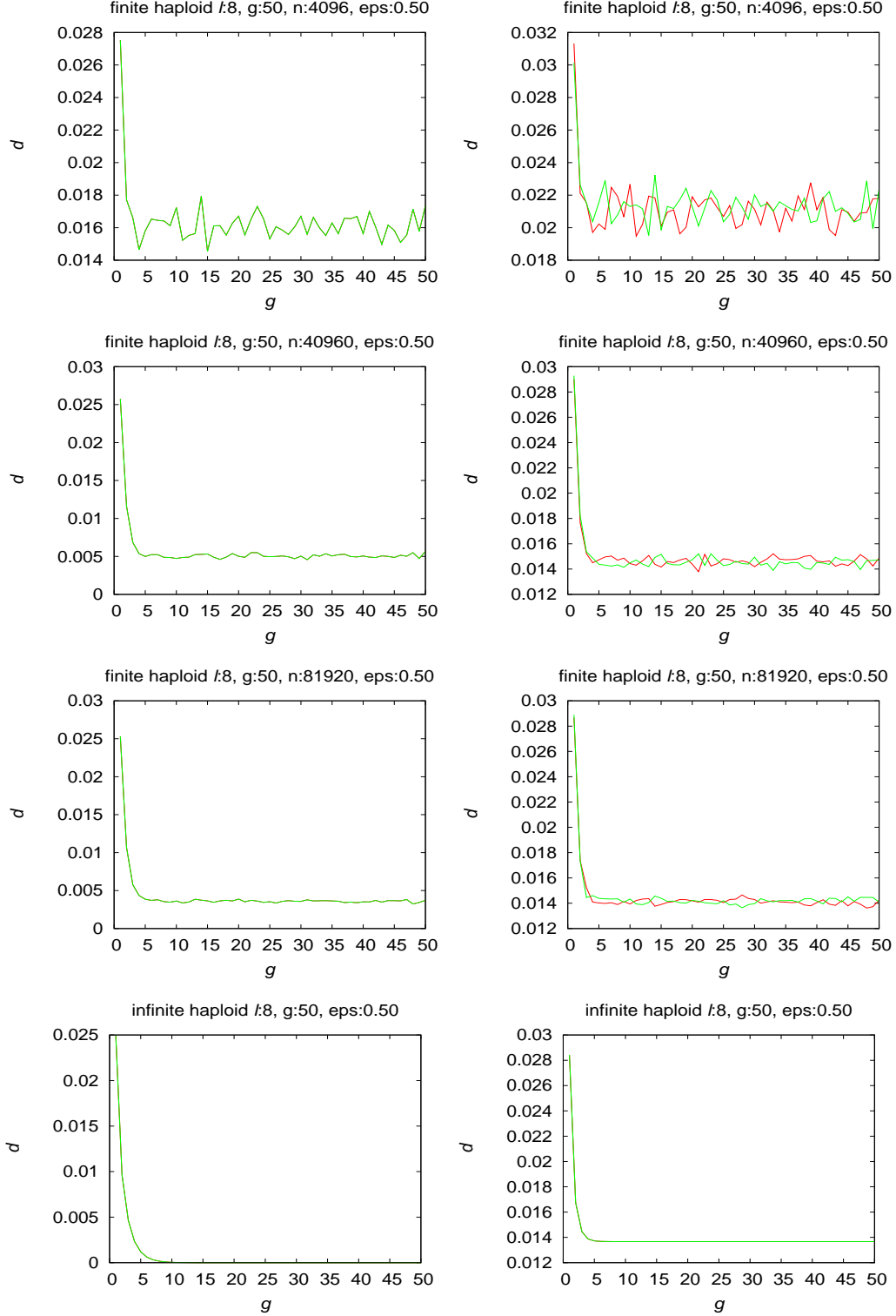


Figure 3.12: Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 8$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

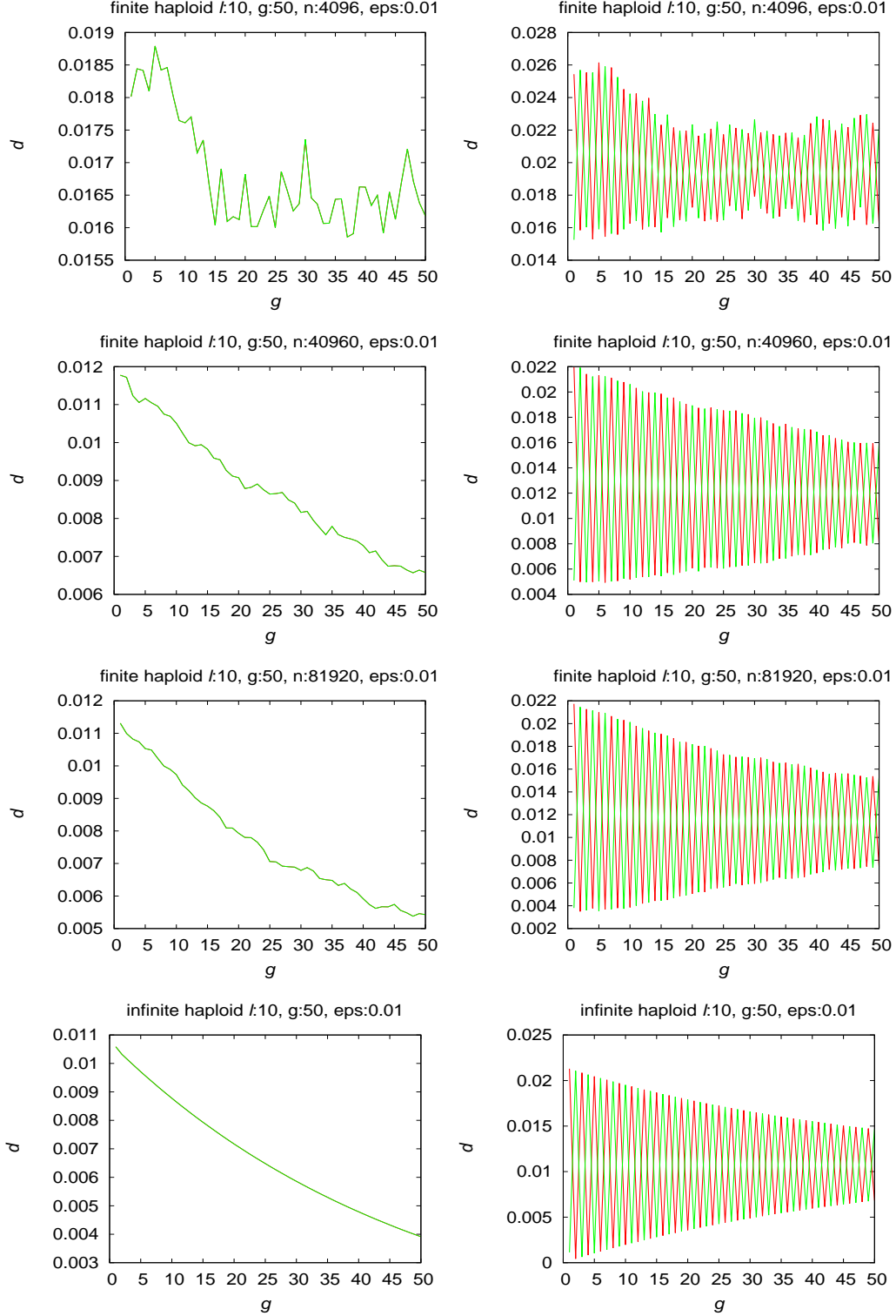


Figure 3.13: Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 10$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

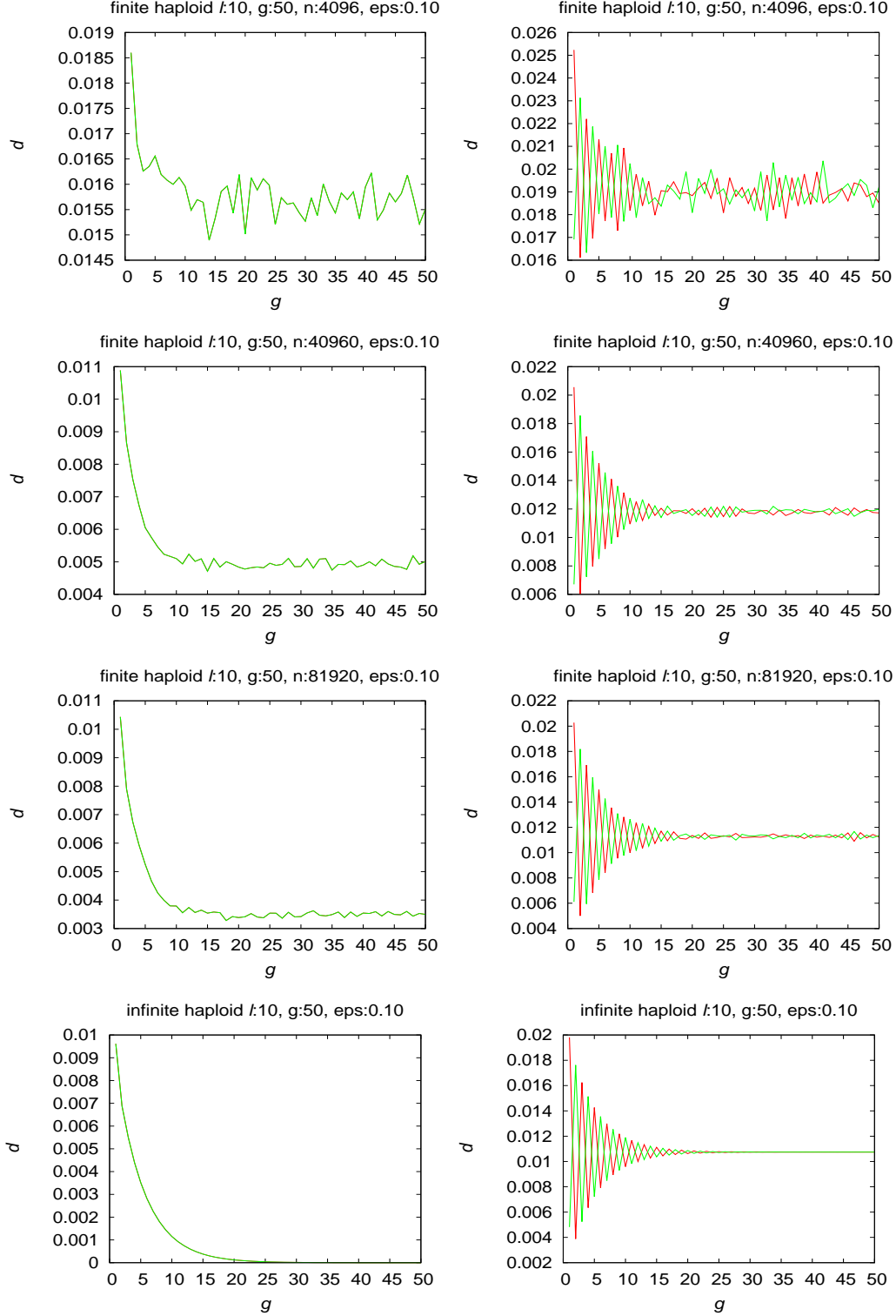


Figure 3.14: Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 10$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

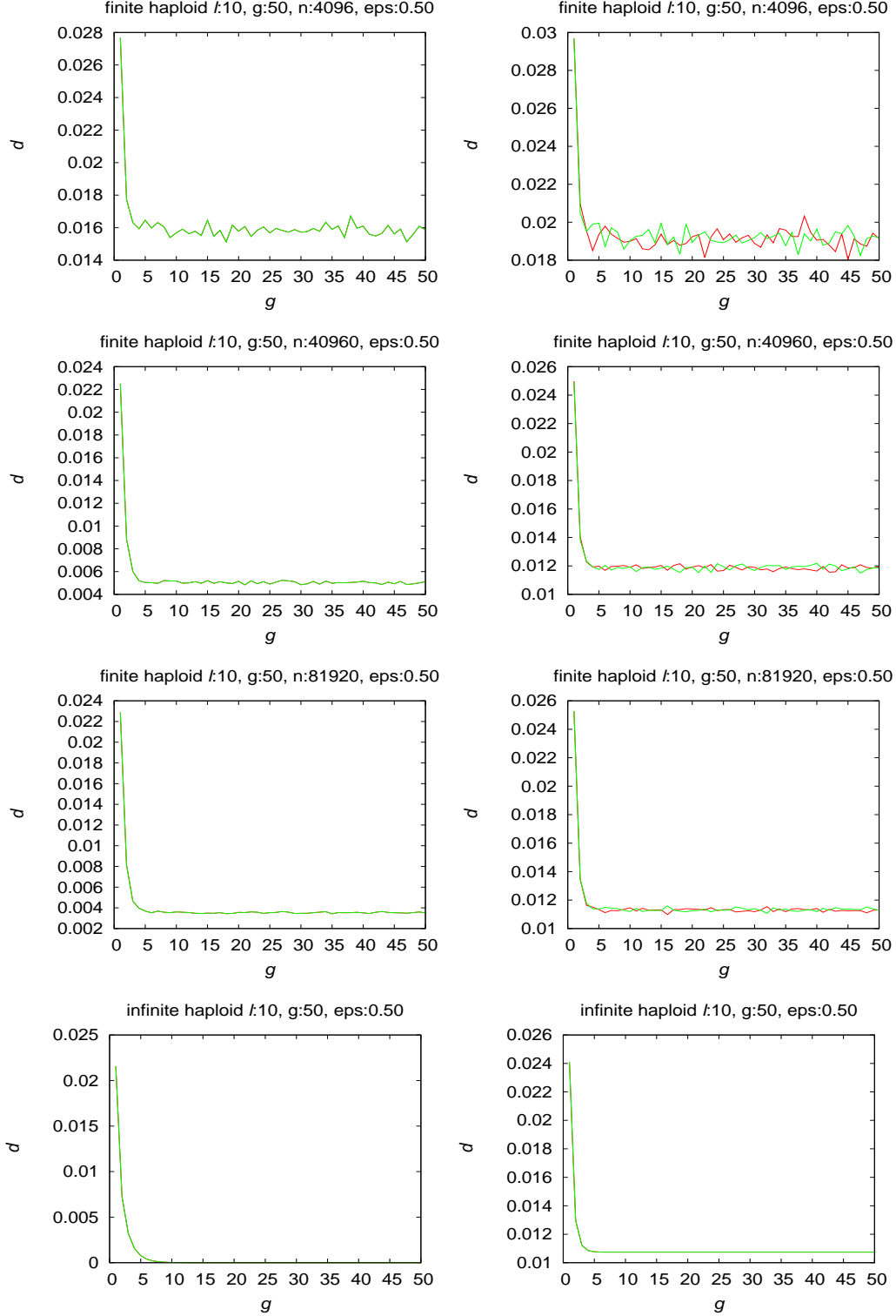


Figure 3.15: Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 10$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

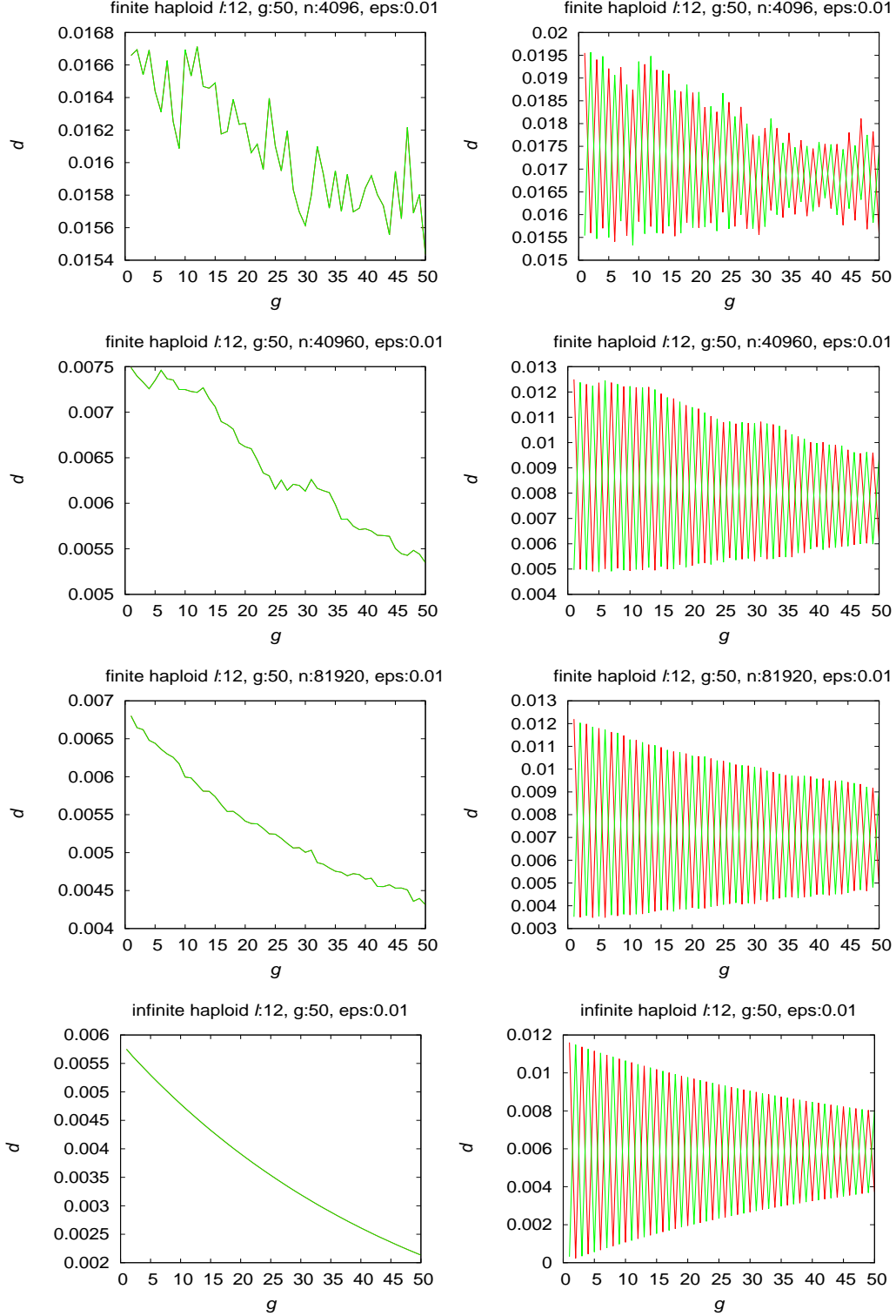


Figure 3.16: Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 12$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

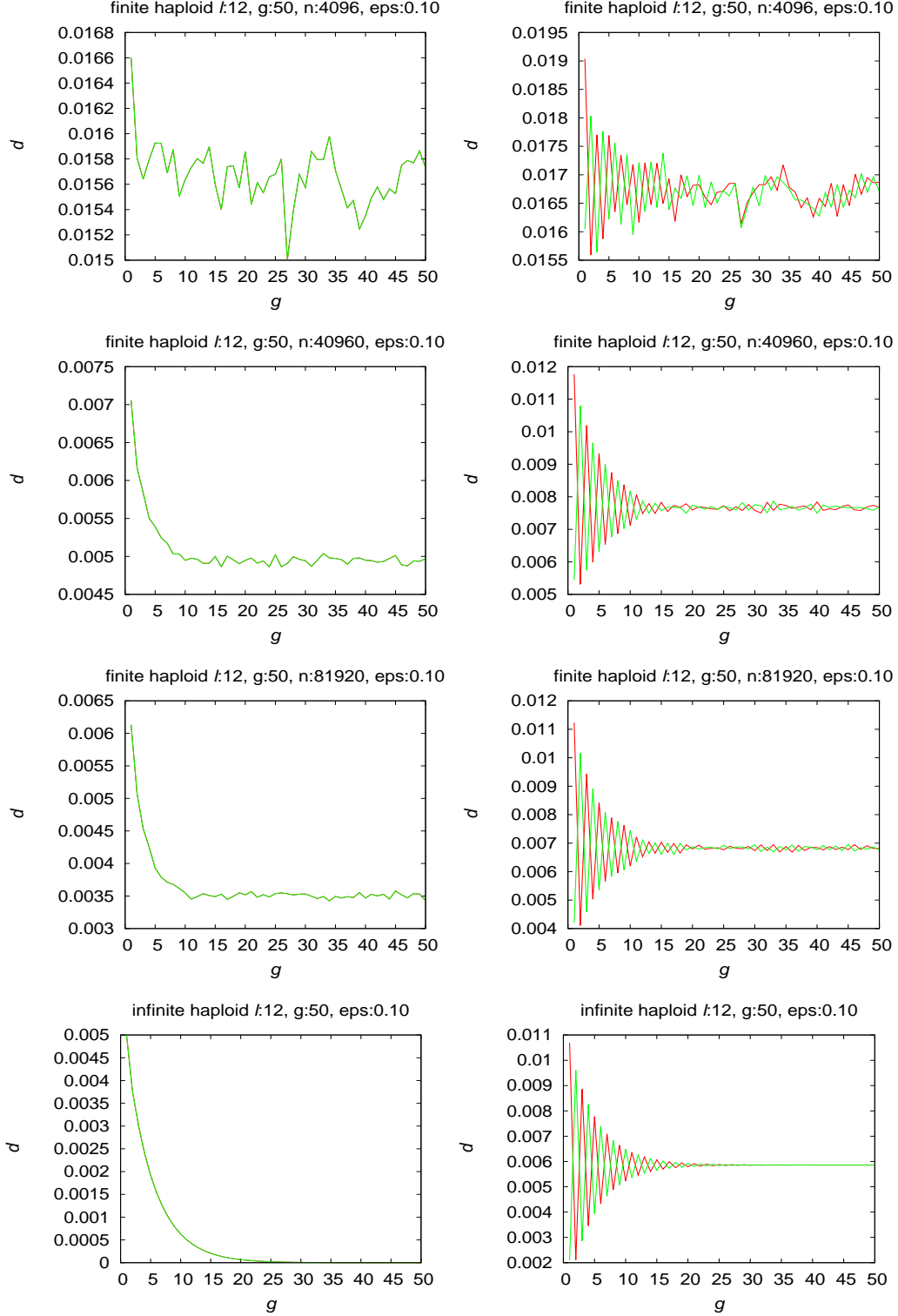


Figure 3.17: Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 12$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

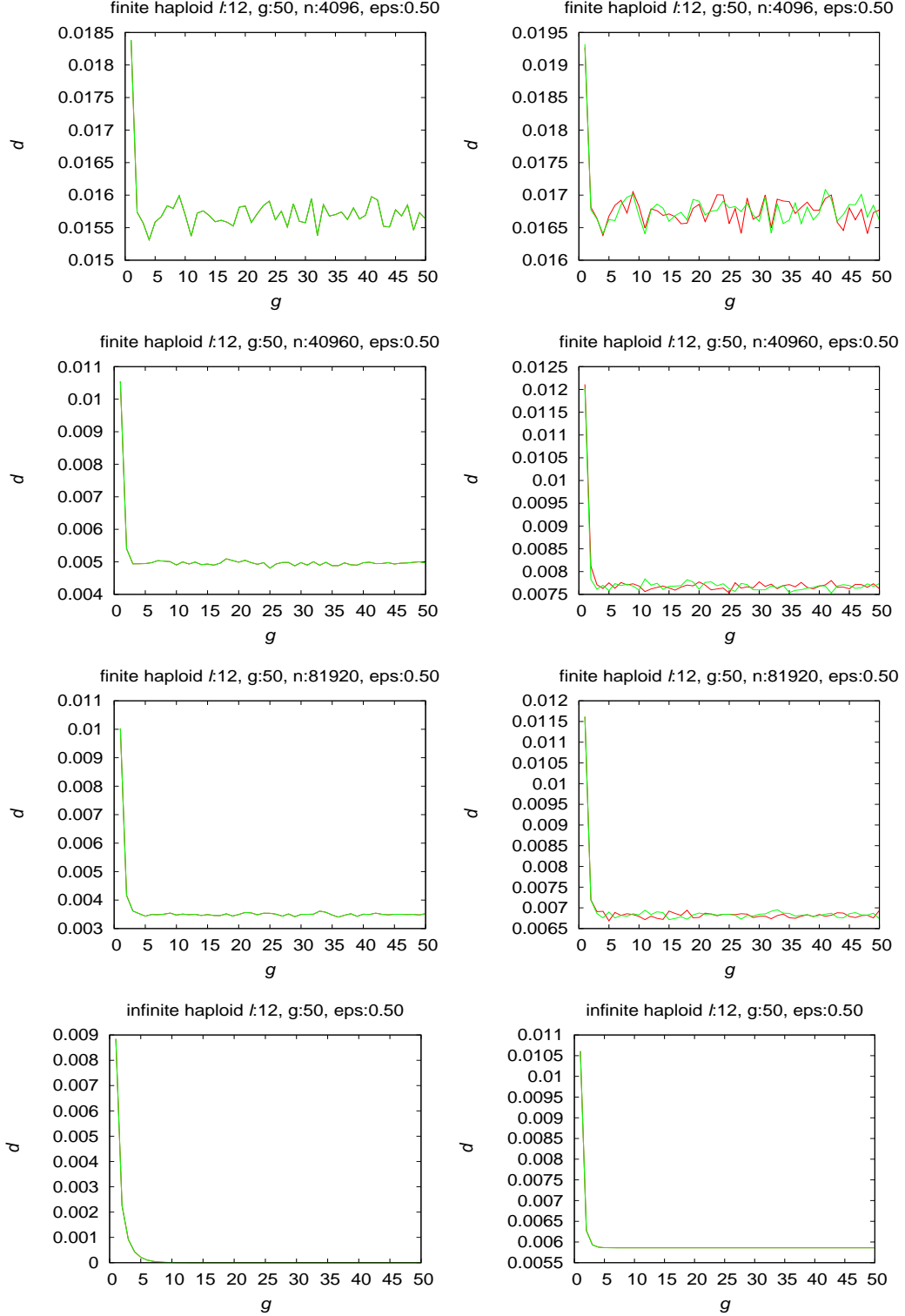


Figure 3.18: Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 12$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

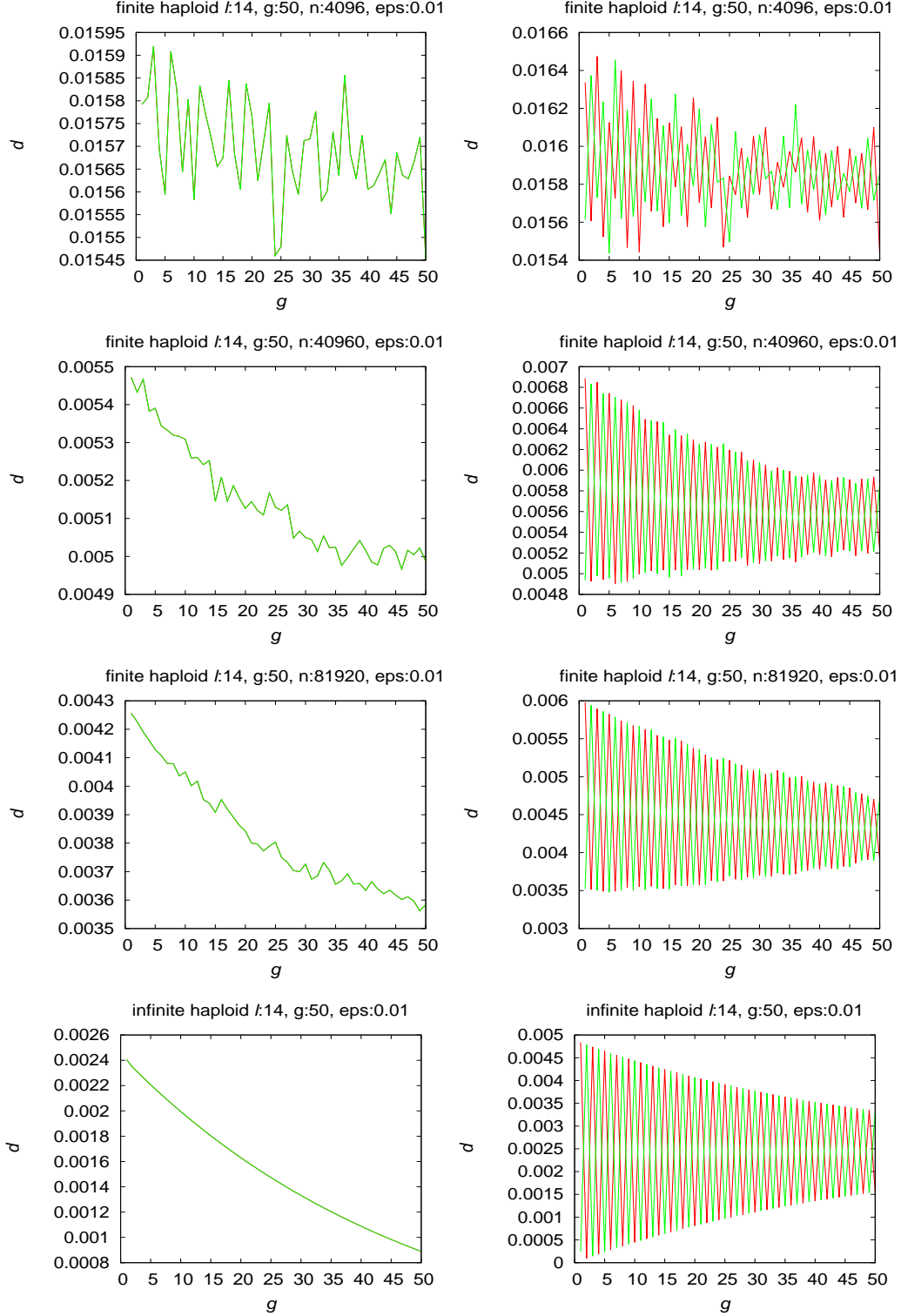


Figure 3.19: Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 14$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

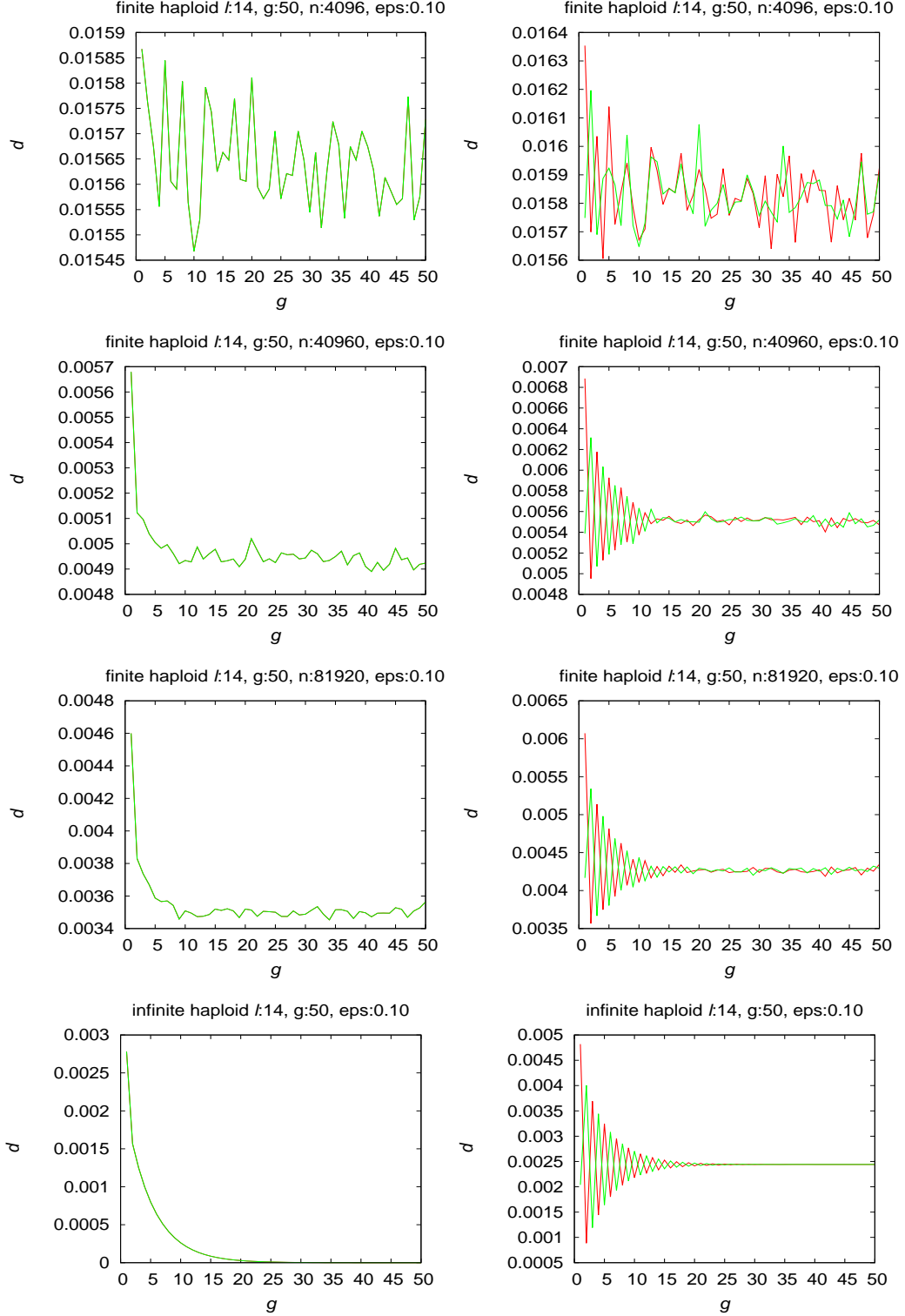


Figure 3.20: Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 14$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

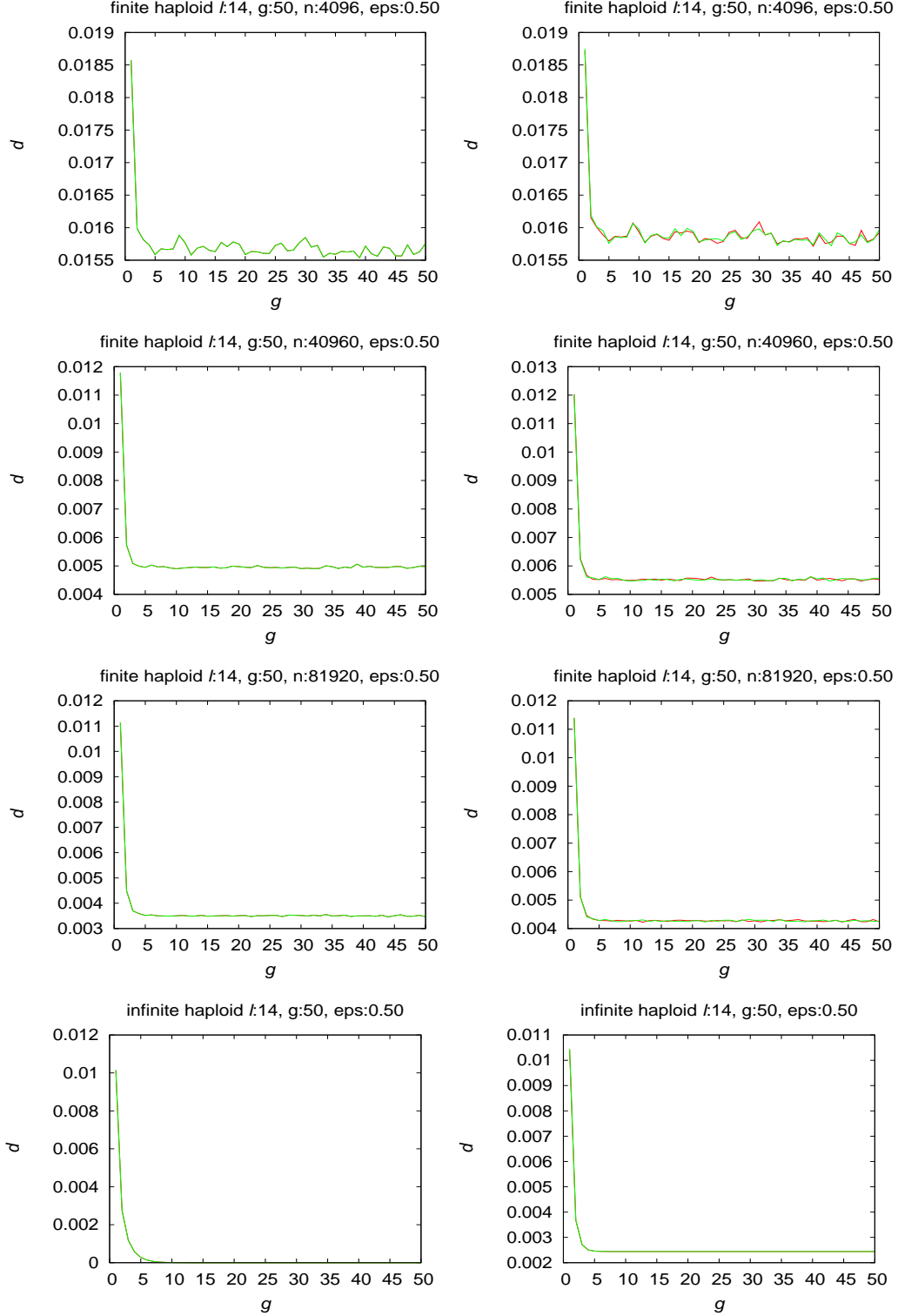


Figure 3.21: Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 14$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

The left column of figures 3.10 through 3.21 shows distance of finite and infinite haploid populations to limit \mathbf{z}^* (limit with violation in mutation distribution $\boldsymbol{\mu}$). The right column shows distance of finite and infinite haploid populations to non-violation limits \mathbf{p}^* and \mathbf{q}^* . Graphs in right column give picture of oscillating behavior of haploid population given violation. Graphs in left column in figures 3.10 through 3.21 show distance between finite population and limit \mathbf{z}^* (limit with violation in $\boldsymbol{\mu}$ distribution) for haploid case.

Both finite and infinite populations oscillate given violation. However, ripples die out quickly, and oscillation ceases for infinite populations. In the finite population case, although ripple amplitudes decrease, oscillation didn't die out completely.

Change in oscillating behavior of population with change in ϵ values $\{0.01, 0.1, 0.5\}$ were also studied. Results show ripples damp out faster with increase ϵ . With smaller values of ϵ , oscillations were sharper and as value of ϵ increased, rate of damping increased. With $\epsilon = 0.01$, oscillation is clearly visible and ripples are sharper; with $\epsilon = 0.1$ oscillation was visible but ripples were damping out quickly; with $\epsilon = 0.5$, oscillation was very minimal or not visible at all.

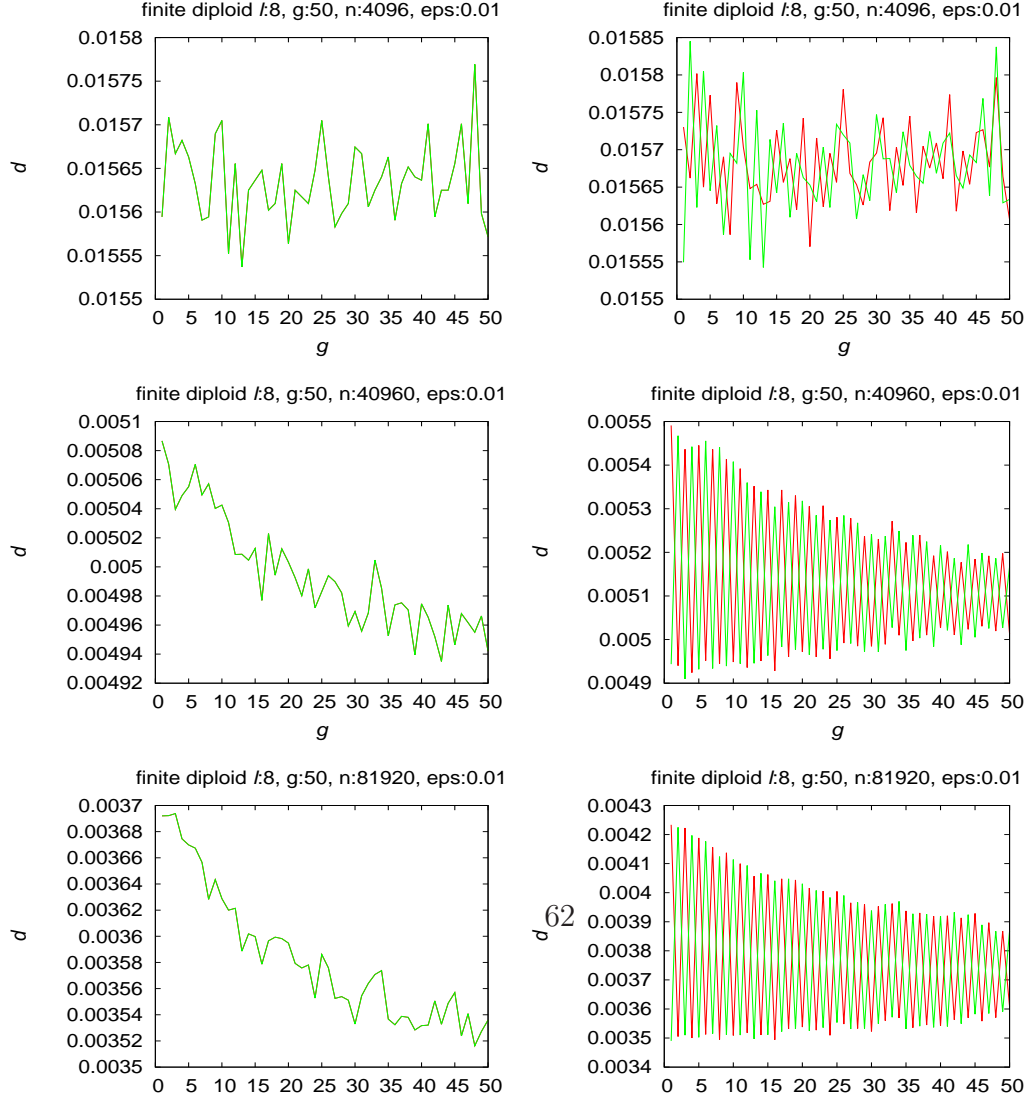
The distance between finite population and limit \mathbf{z}^* (limit with violation in $\boldsymbol{\mu}$ distribution) decreases as finite population size increases, and finite population shows behavior similar to infinite population behavior as finite population reach large number. Simulation results show infinite population converges to limit \mathbf{z}^* quicker with increase in ϵ .

The distance data for haploid population in case of violation in $\boldsymbol{\mu}$ distribution with different values of ϵ for different finite population size N are tabulated in table 3.3.

From table 3.3, average distance calculated for finite population size 4096 is 0.0158, for size 40960 is 0.0055 and for size 81920 is 0.0041. These results show average distance between finite population and limit \mathbf{z}^* closely follows expected single step distance between finite and infinite population given in 3.1. The distance decreased as $1/\sqrt{N}$.

Table 3.3: Experimental distance measured for violation in μ for haploids: ℓ is genome length, ϵ is error introduced to μ for violation, and average distance between finite and infinite populations is tabulated in last three columns.

ϵ	case	ℓ	N 4096	$=$ N 40960	$=$ N 81920
0.01	haploid	8	0.0176	0.0094	0.0093
		10	0.0168	0.0088	0.0077
		12	0.0161	0.0064	0.0053
		14	0.0157	0.0051	0.0038
0.1	haploid	8	0.0158	0.0054	0.0041
		10	0.0158	0.0053	0.0039
		12	0.0157	0.0051	0.0036
		14	0.0156	0.0050	0.0035
0.5	haploid	8	0.0161	0.0056	0.0042
		10	0.0161	0.0055	0.0040
		12	0.0157	0.0051	0.0036
		14	0.0157	0.0051	0.0037



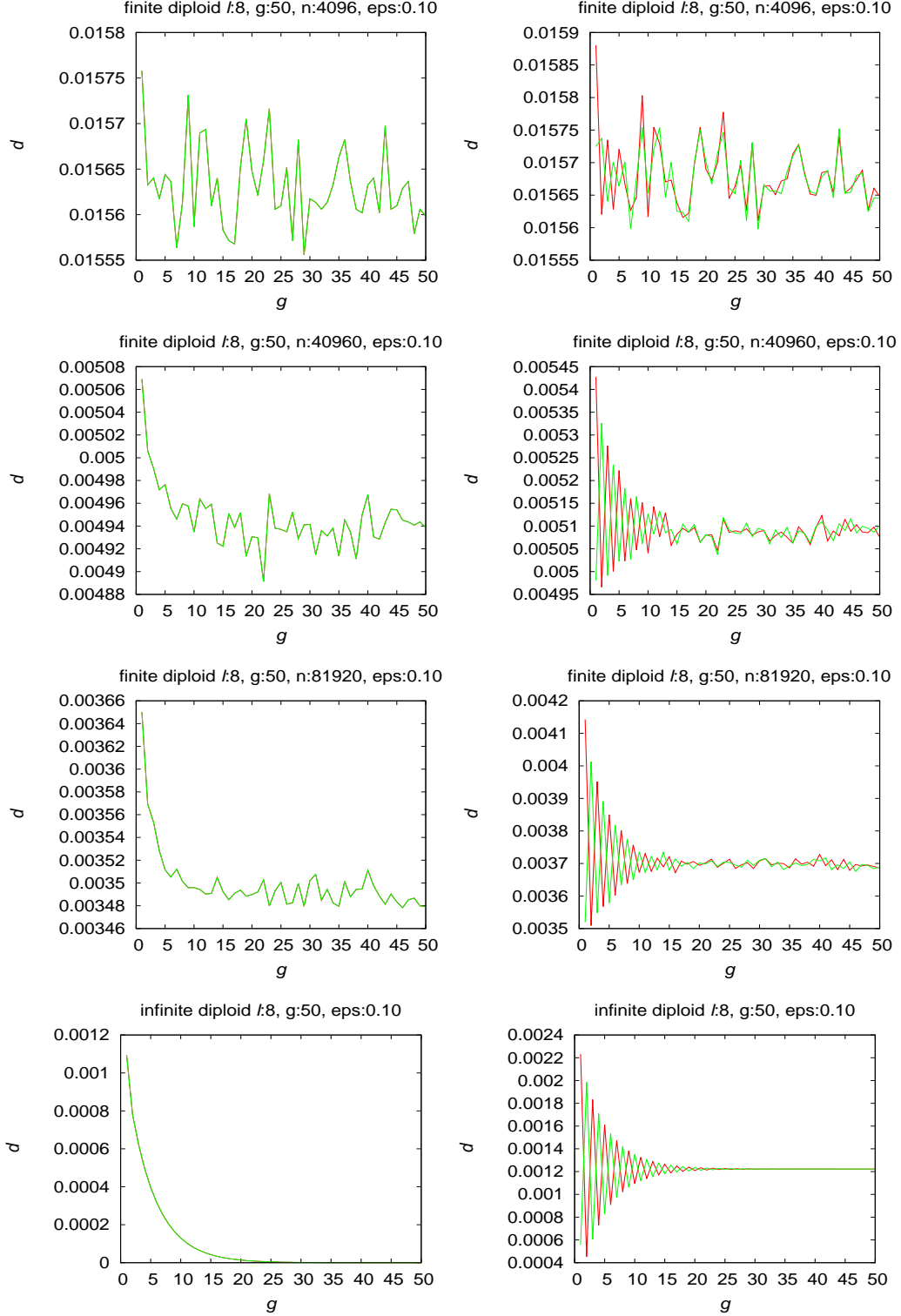


Figure 3.23: Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 8$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

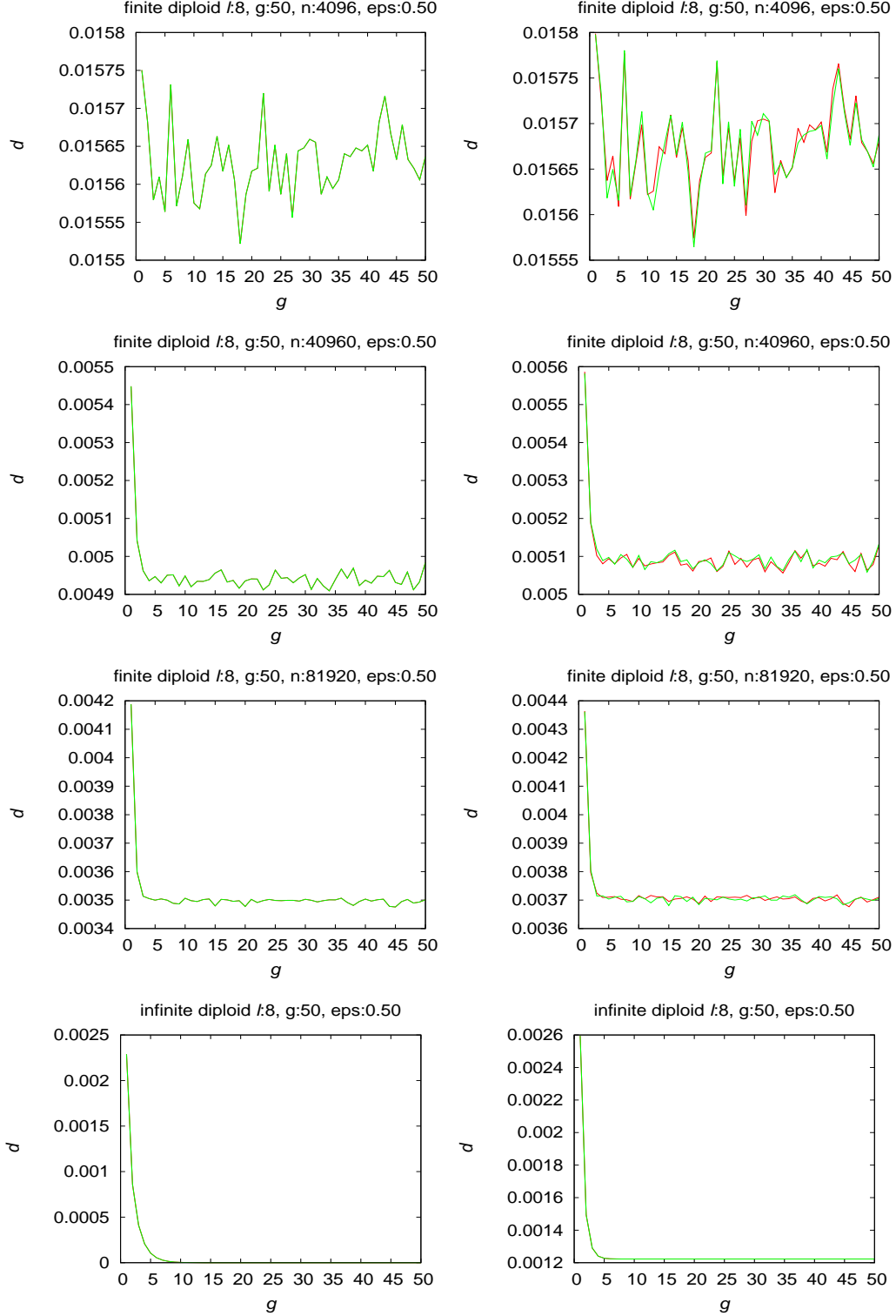


Figure 3.24: Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 8$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

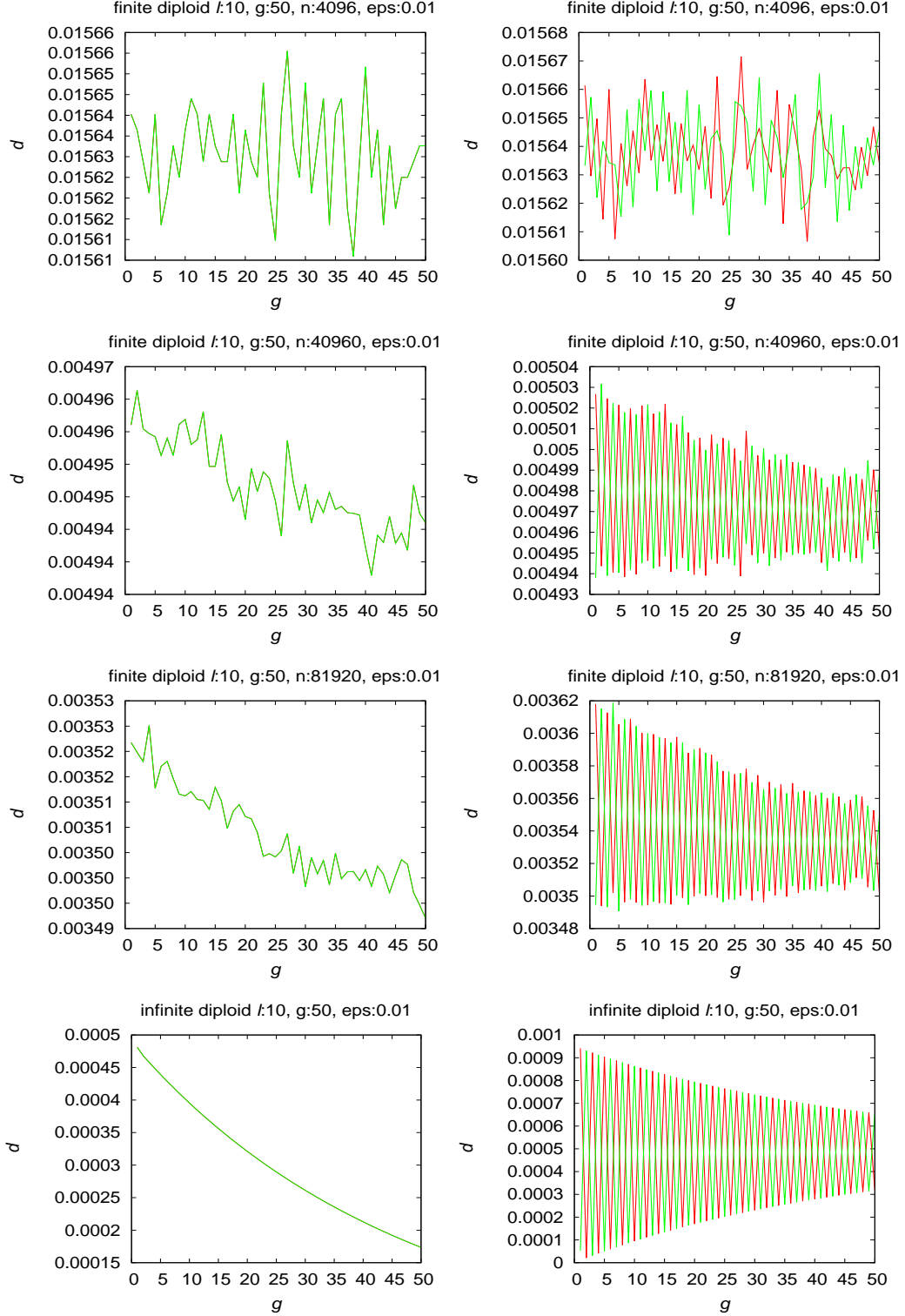


Figure 3.25: Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 10$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

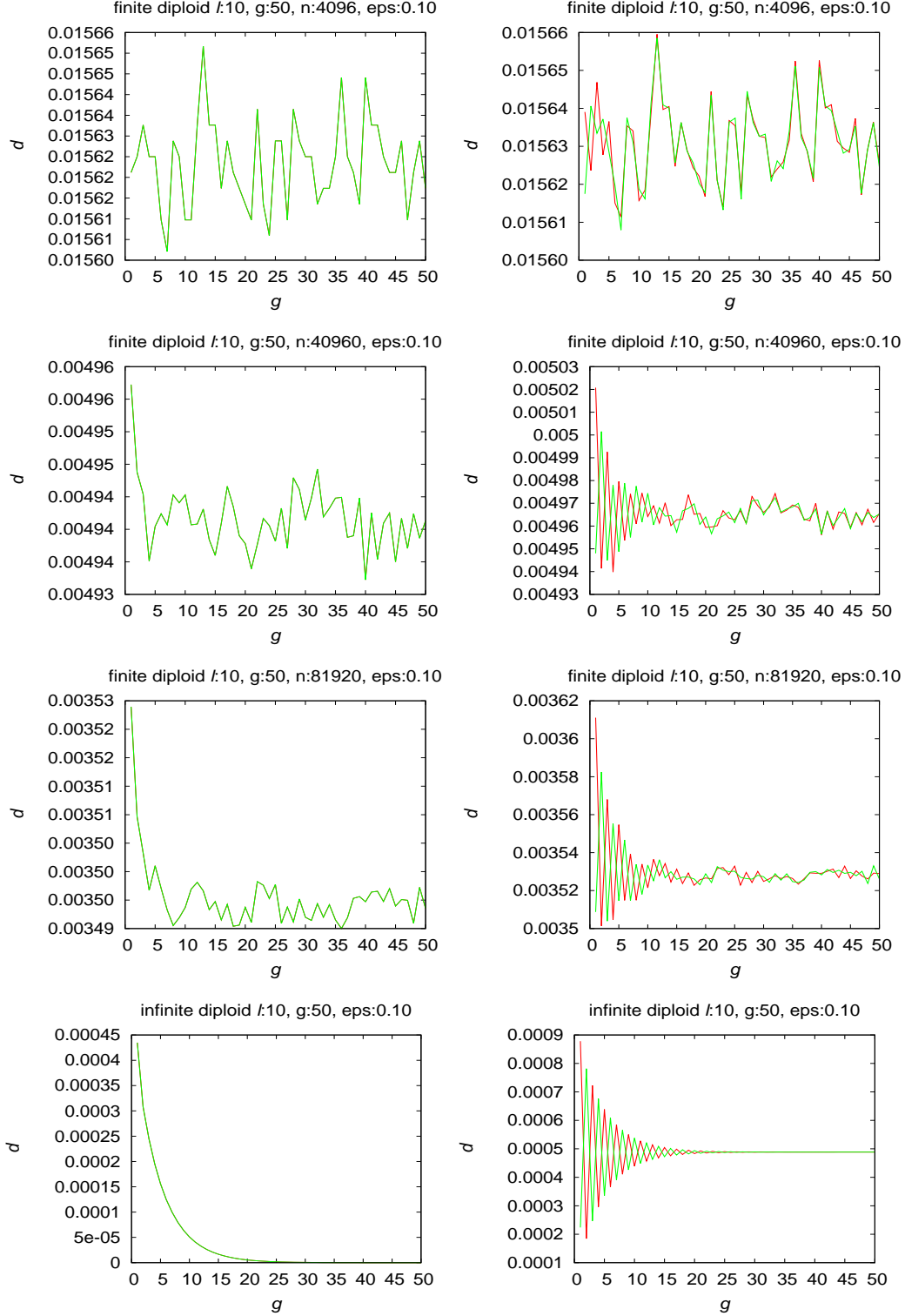


Figure 3.26: Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 10$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

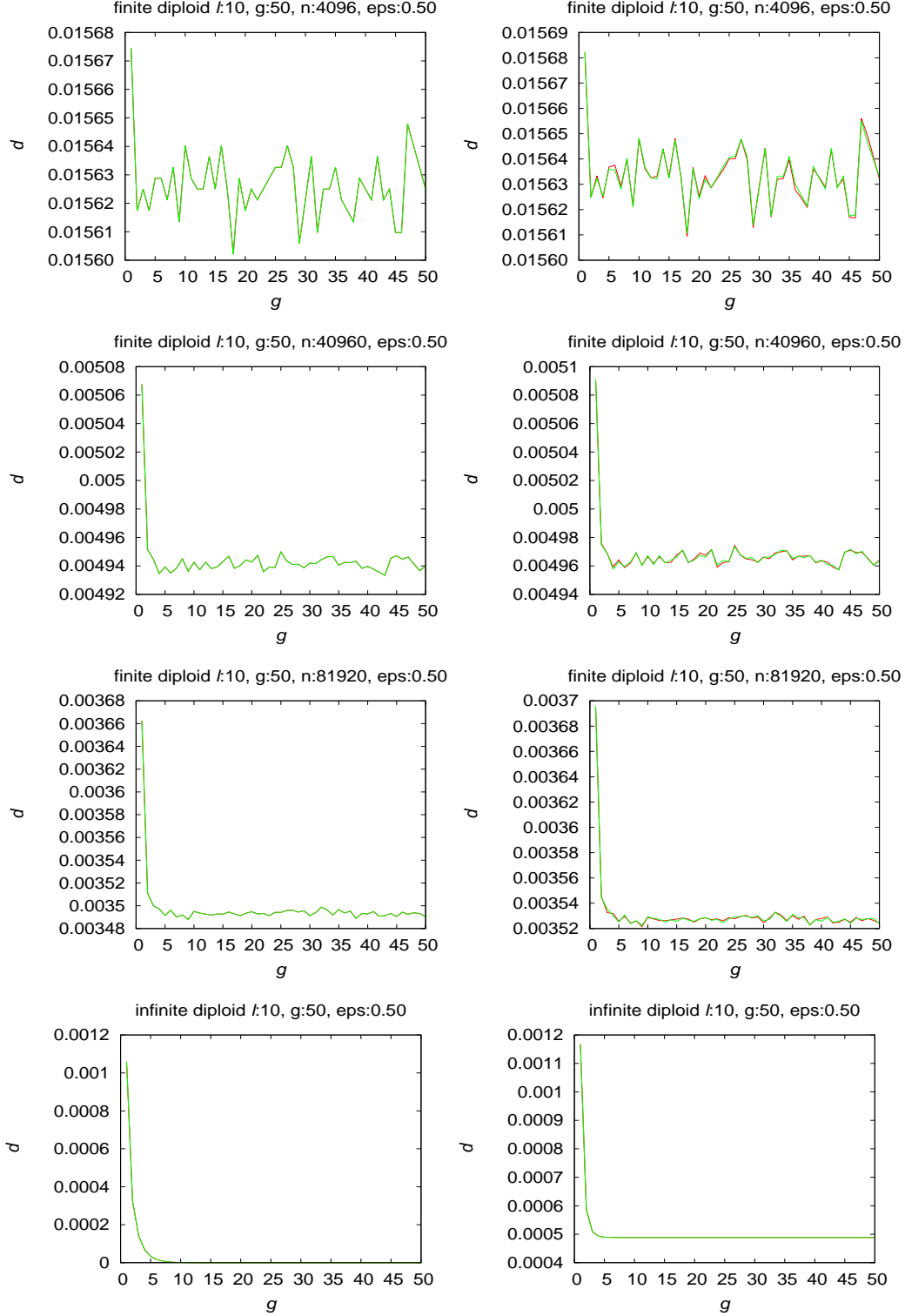


Figure 3.27: Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 10$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

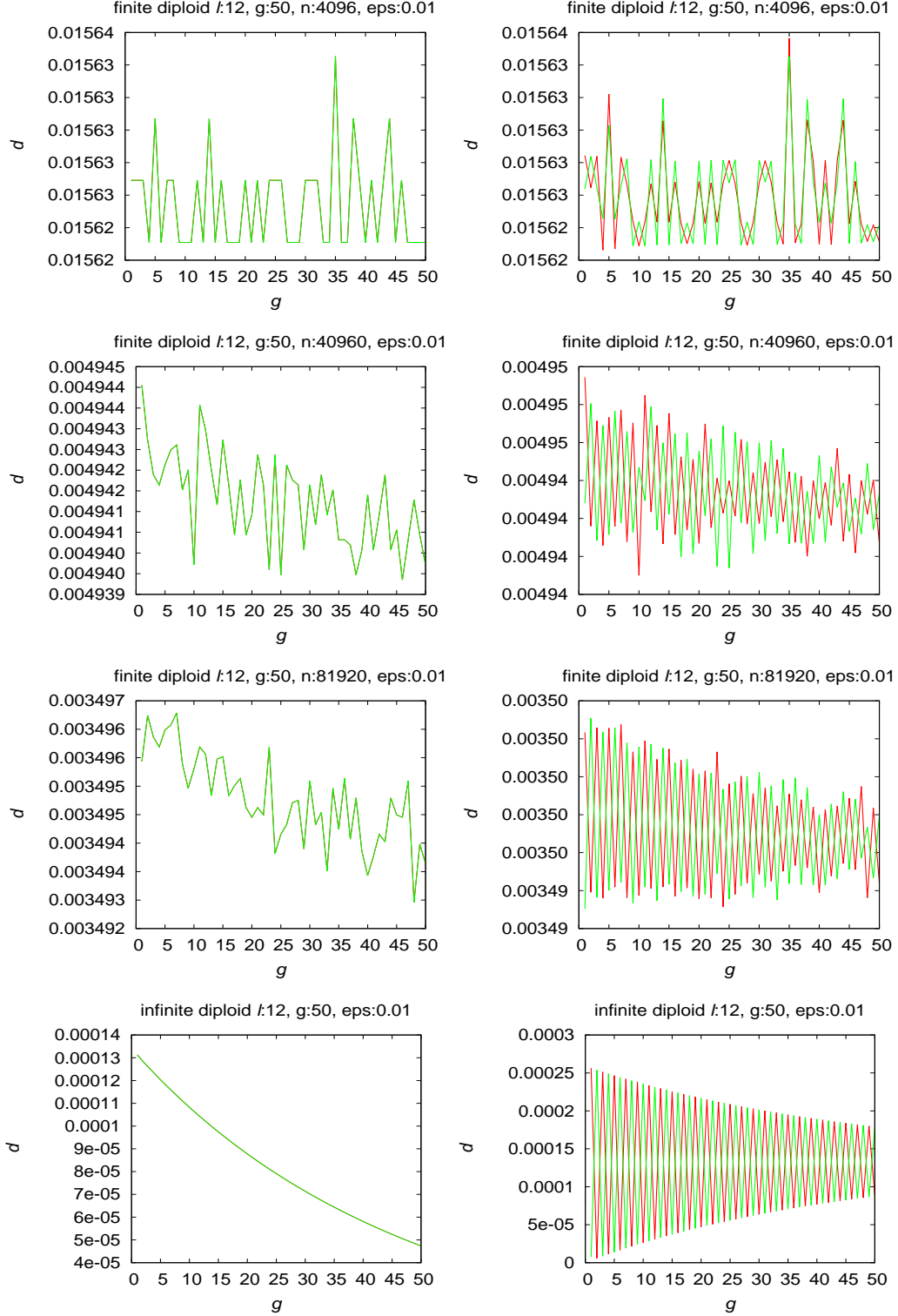


Figure 3.28: Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 12$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

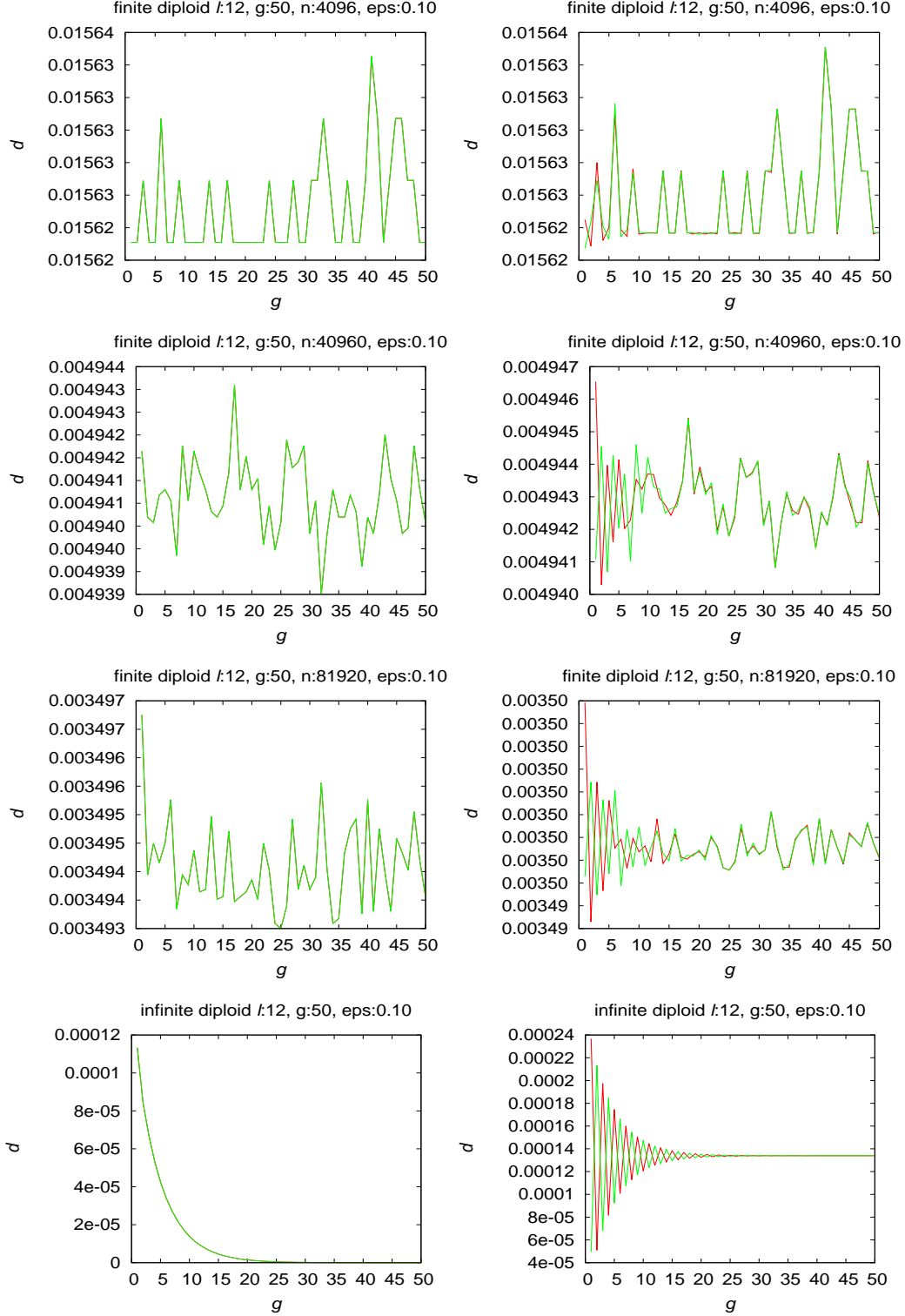


Figure 3.29: Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 12$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

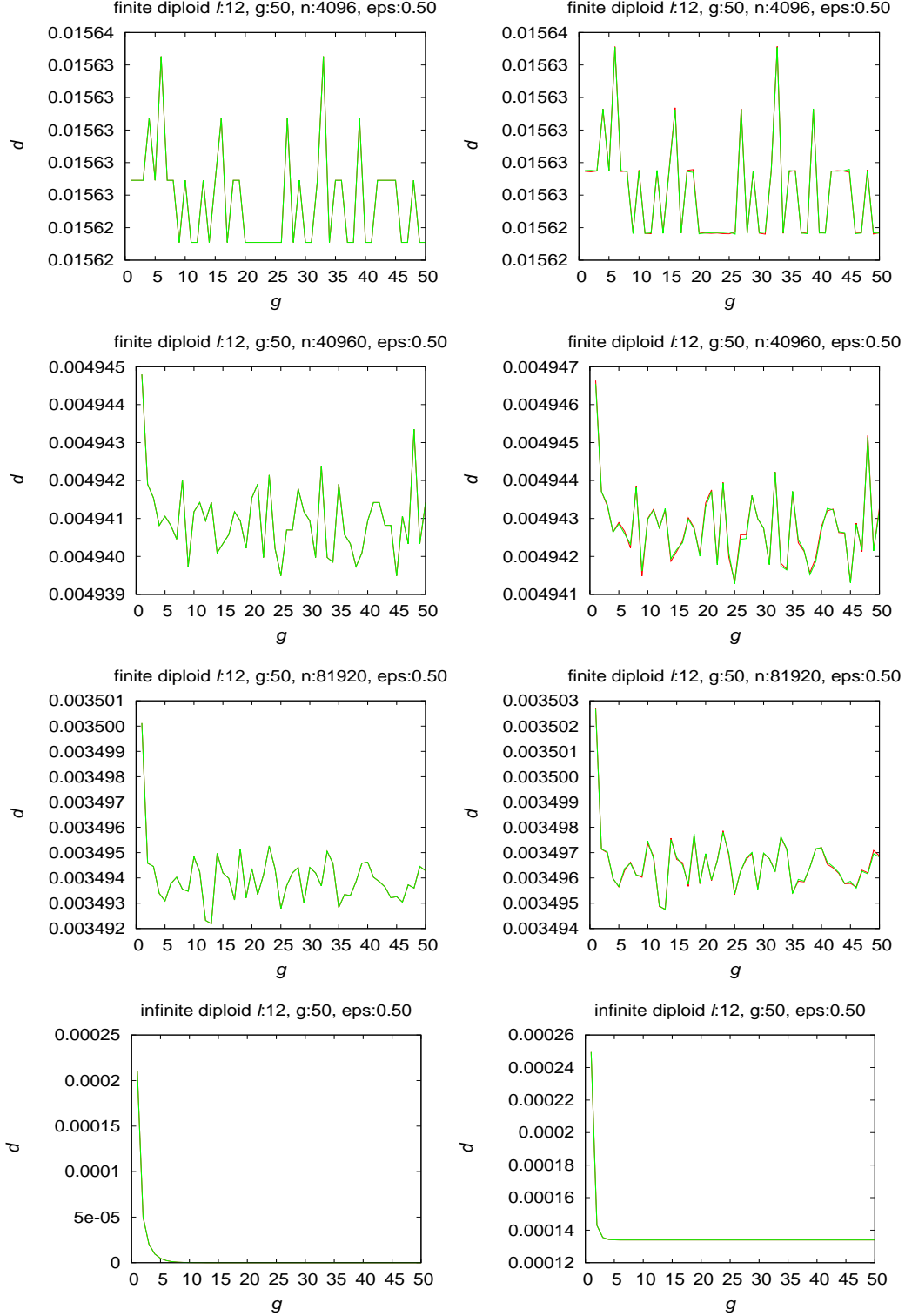


Figure 3.30: Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 12$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

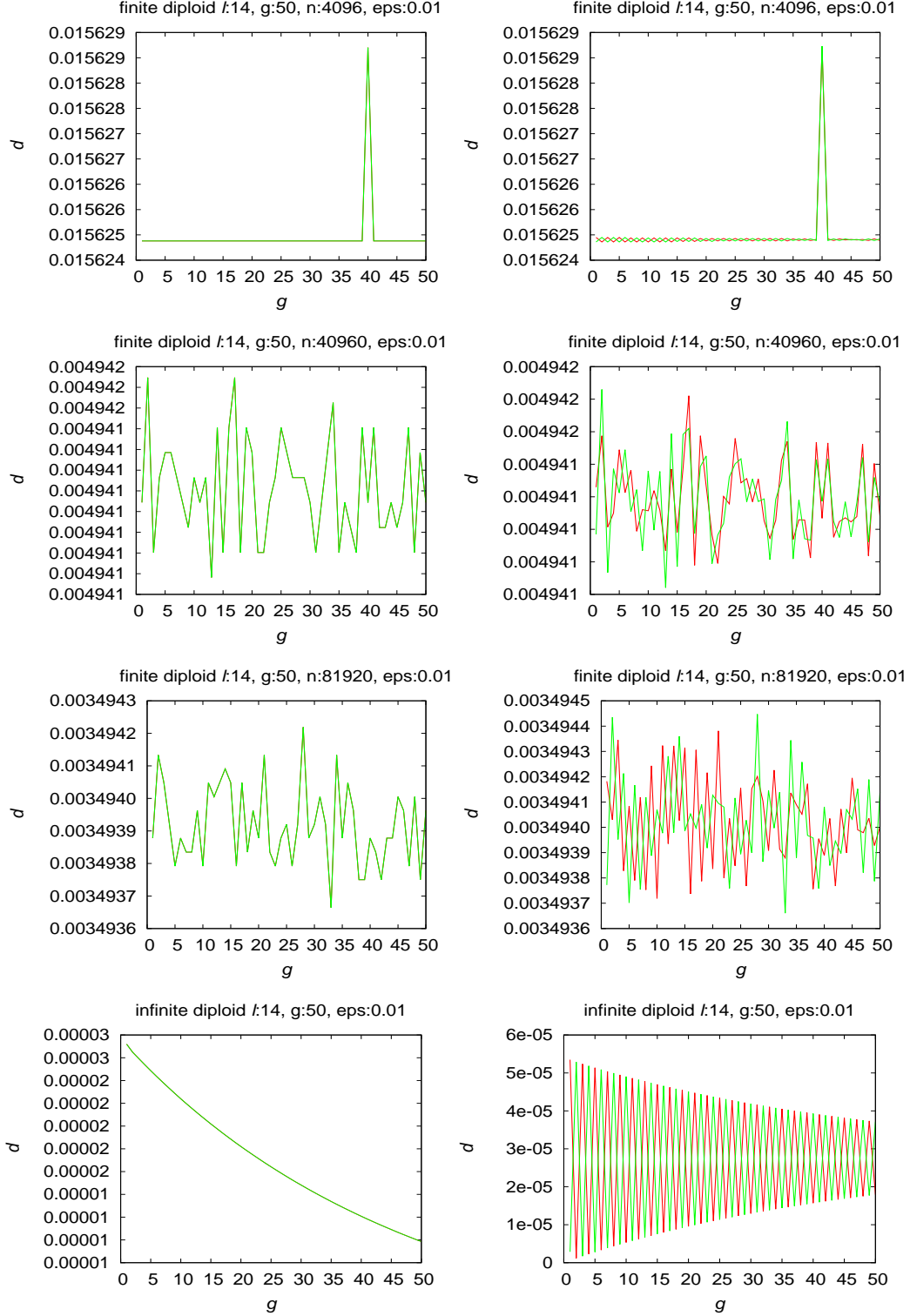


Figure 3.31: Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 14$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

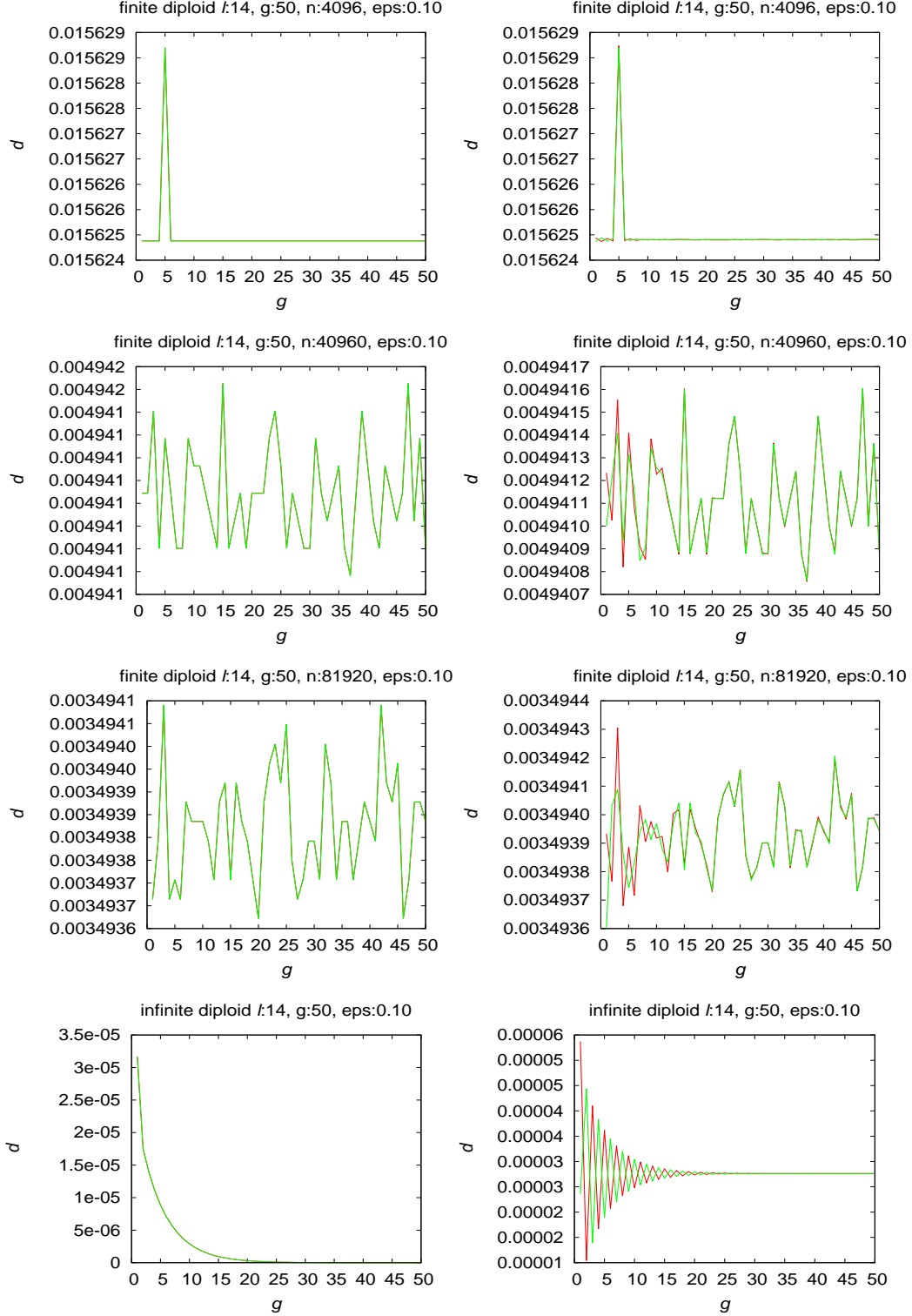


Figure 3.32: Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 14$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

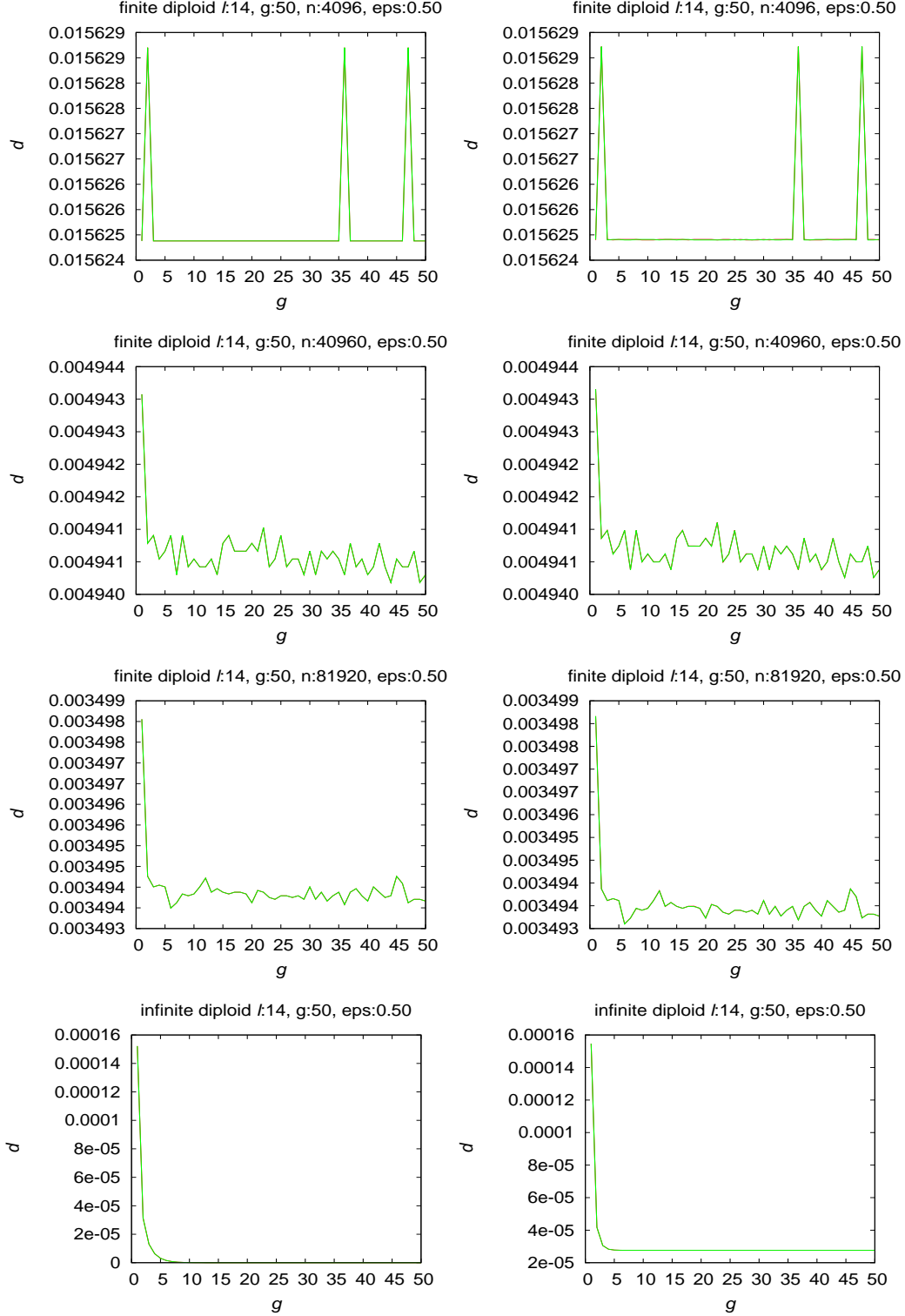


Figure 3.33: Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 14$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

The left column of figures 3.22 through 3.33 shows distance of finite and infinite diploid populations to limit \mathbf{z}^* (limit with violation in $\boldsymbol{\mu}$ distribution). The right column shows distance of finite and infinite diploid populations to non-violation limits \mathbf{p}^* and \mathbf{q}^* . Graphs in right column give picture of oscillating behavior of diploid population given violation. Graphs in left column in figures 3.22 through 3.33 show distance between finite population and limit \mathbf{z}^* (limit with violation in $\boldsymbol{\mu}$ distribution) for diploid case.

Like in haploid case, both finite and infinite populations oscillate given violation. Ripples die out quickly, and oscillation ceases for infinite populations, and in the finite population case, although ripple amplitudes decrease, oscillation didn't die out completely.

Change in oscillating behavior of population with change in ϵ values $\{0.01, 0.1, 0.5\}$ were also studied. Results show ripples damp out faster with increase ϵ . With smaller values of ϵ , oscillations were sharper and as value of ϵ increased, rate of damping increased. With $\epsilon = 0.01$, oscillation is clearly visible and ripples are sharper; with $\epsilon = 0.1$ oscillation was visible but ripples were damping out quickly; with $\epsilon = 0.5$, oscillation was very minimal or not visible at all.

Like just in haploid case, the distance between finite population and limit \mathbf{z}^* (limit with violation in $\boldsymbol{\mu}$ distribution) decreases as finite population size increases, and finite population shows behavior similar to infinite population behavior as finite population reach large number. Simulation results show infinite population converges to limit \mathbf{z}^* quicker with increase in ϵ .

The distance data for diploid population in case of violation in $\boldsymbol{\mu}$ distribution with different values of ϵ for different finite population size N are tabulated in table 3.4.

From table 3.4, average distance calculated for finite population size 4096 is 0.0158, for size 40960 is 0.0055 and for size 81920 is 0.0041. These results show average distance between finite population and limit \mathbf{z}^* closely follows expected single step distance between finite and infinite population given in 3.1. The distance decreased as $1/\sqrt{N}$.

Table 3.4: Experimental distance measured for violation in μ for diploids: ℓ is genome length, ϵ is error introduced to μ for violation, and average distance between finite and infinite populations is tabulated in last three columns.

ϵ	case	ℓ	N 4096	= N 40960	= N 81920
0.01	diploid	8	0.0156	0.0050	0.0035
		10	0.0156	0.0049	0.0035
		12	0.0156	0.0049	0.0035
		14	0.0156	0.0049	0.0035
0.1	diploid	8	0.0156	0.0049	0.0035
		10	0.0156	0.0049	0.0035
		12	0.0156	0.0049	0.0035
		14	0.0156	0.0049	0.0035
0.5	diploid	8	0.0156	0.0049	0.0035
		10	0.0156	0.0049	0.0035
		12	0.0156	0.0049	0.0035
		14	0.0156	0.0049	0.0035

Violation introduced to the mutation distribution μ creates new masks (different than in case of without violation) to be used in mutation during transmission. With small ϵ , the probability of using the new masks available due to violation is very small and those masks might not be used at all during mutation in finite population, but with higher values of ϵ , those new masks have higher chance of usage during mutation which cause oscillation to damp out quickly or cause no oscillation at all.

From graphs on the right in figures 3.10 through 3.21, and 3.22 through 3.33, haploid population shows sharper and more distinct oscillations for same size of population size than diploid population; diploid population requires larger population size to exhibit sharper oscillations. And from graphs on the left in figures 3.10 through 3.21, and 3.22 through 3.33, diploid population converges to limit z^* more quickly than haploid population.

3.5.2 Violation in Crossover Distribution

The crossover distribution χ was modified as

$$\chi_i = (1 - \epsilon)\chi; \quad i = \{1, 2, \dots, 2^\ell - 1\}$$

So that

$$\chi_i + \chi_{i+g} = 1 - \epsilon; \quad g \text{ is defined in section } \textcolor{blue}{3.1}$$

Then j is chosen where $\chi_j = 0$ and set $\chi_j = \epsilon$.

Simulations were repeated with the violations in (3.3) described above for crossover distribution. The distances of both infinite and finite populations to limit \mathbf{z}^* were plotted. The distances of both infinite and finite populations to non-violation limits \mathbf{p}^* and \mathbf{q}^* were also plotted.

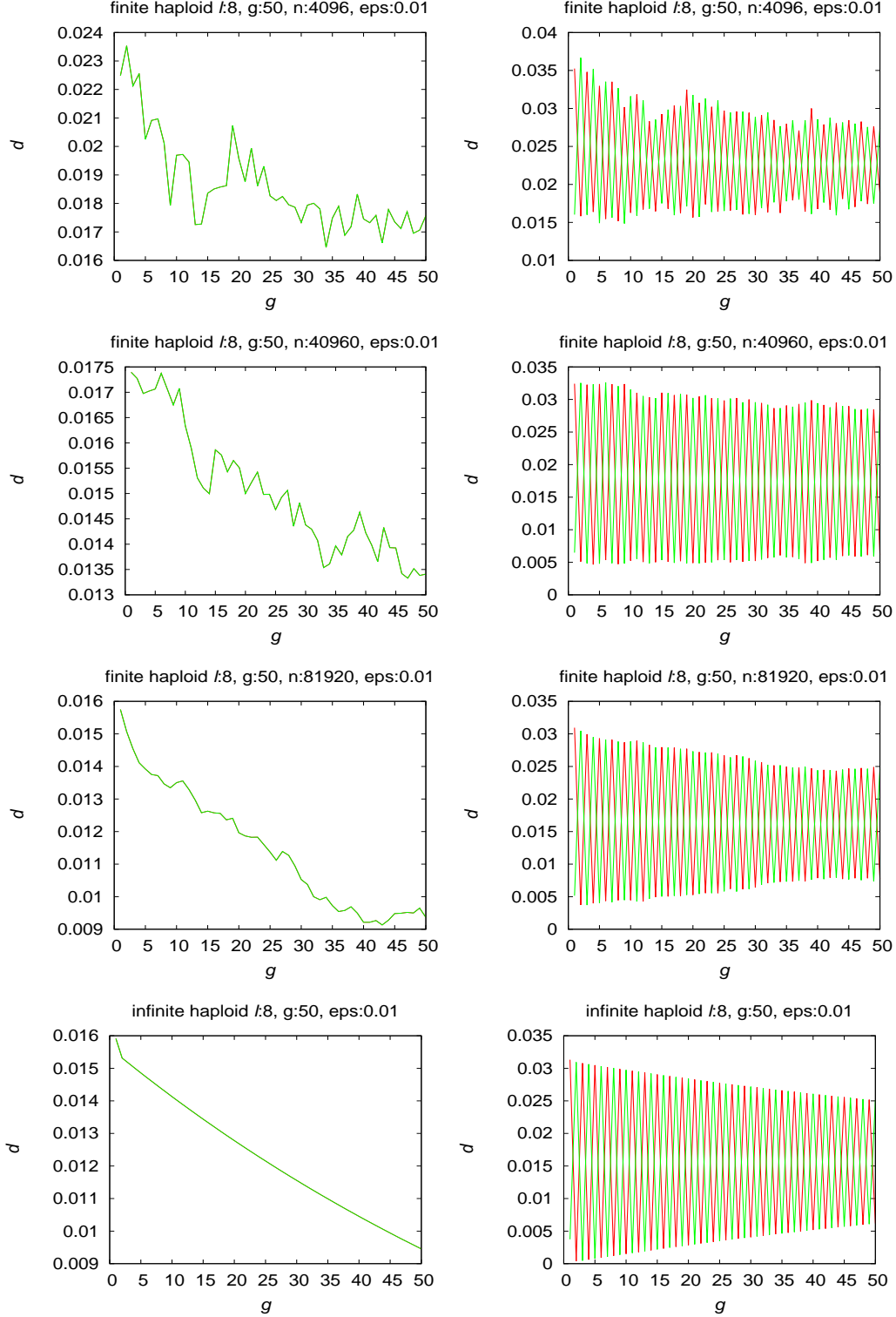


Figure 3.34: Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 8$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

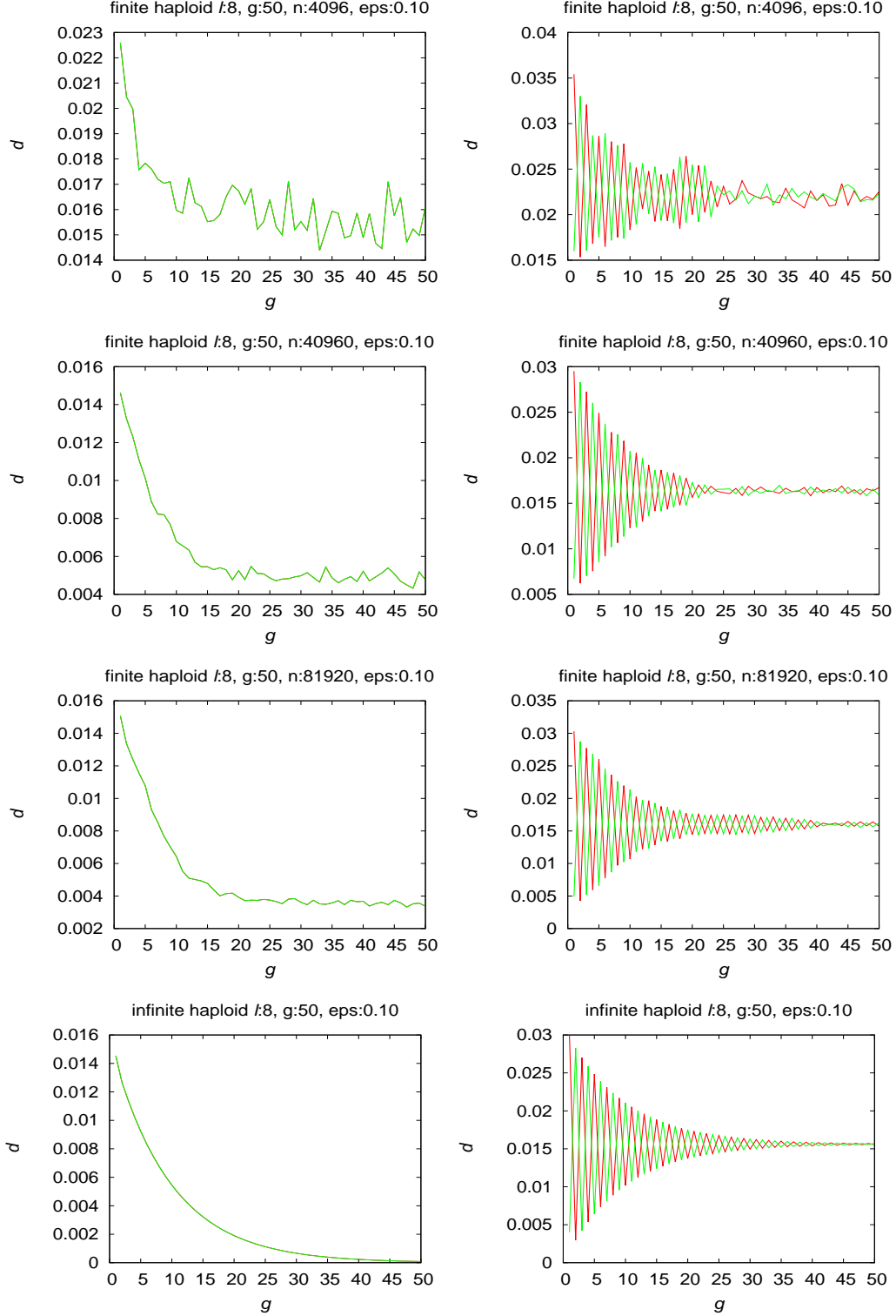


Figure 3.35: Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 8$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

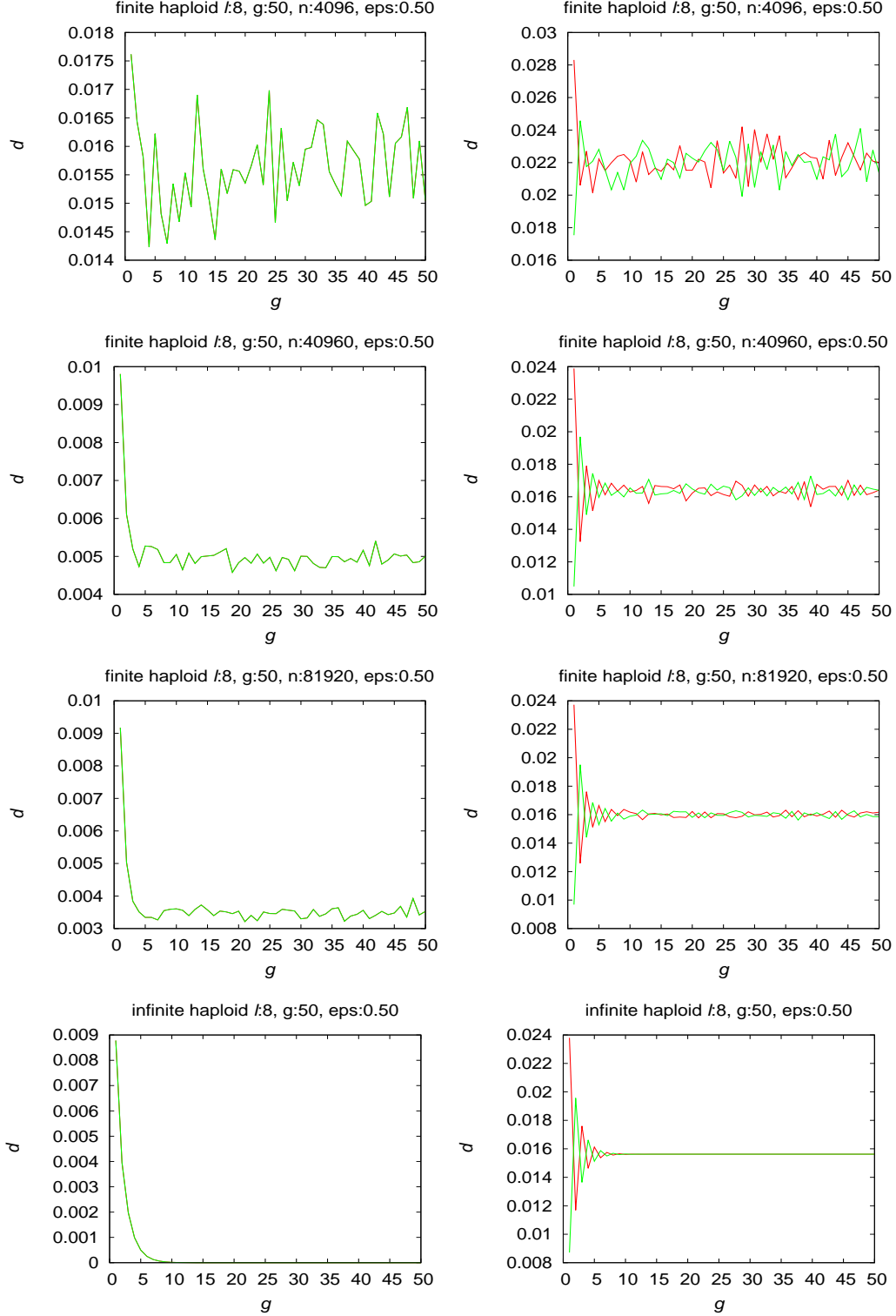


Figure 3.36: Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 8$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

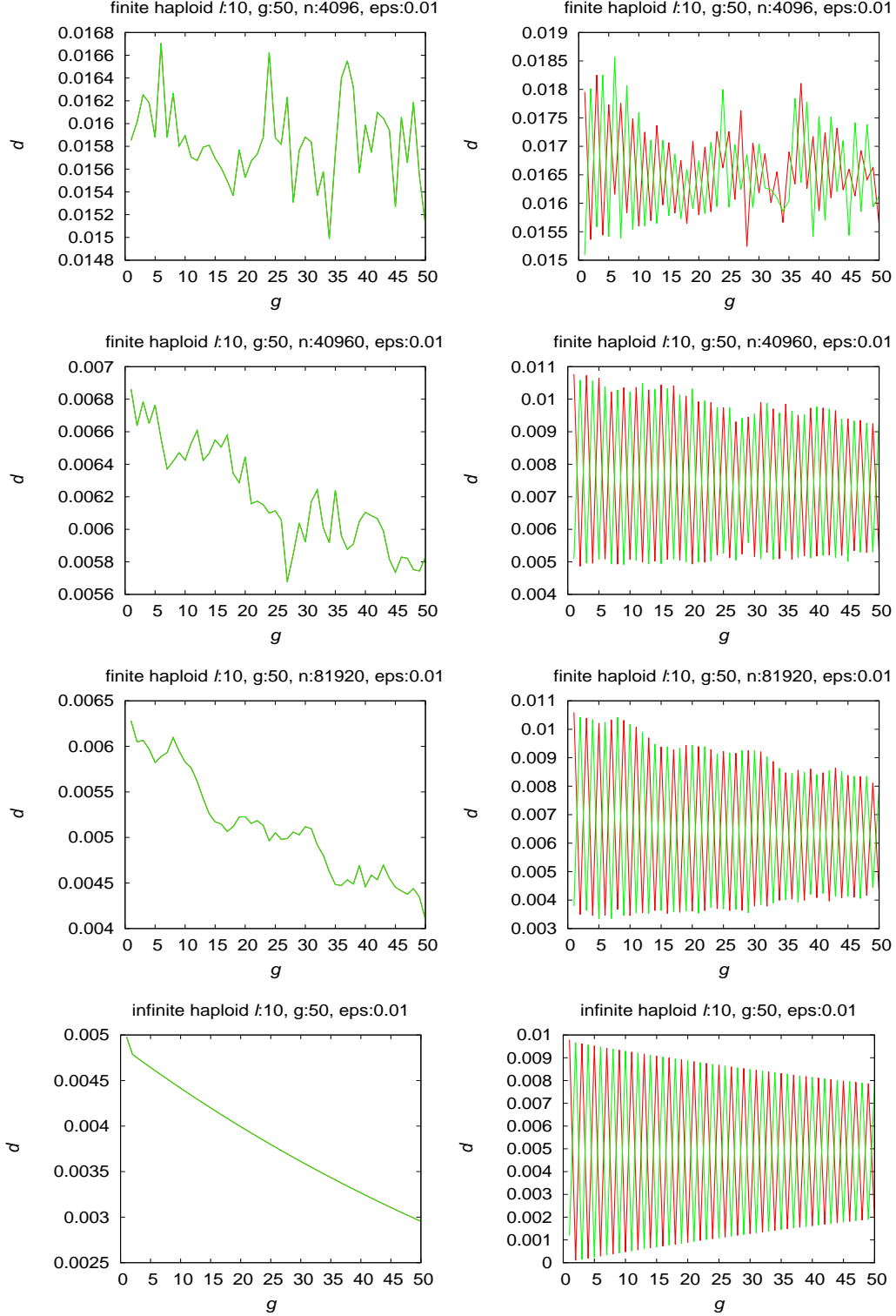


Figure 3.37: Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 10$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

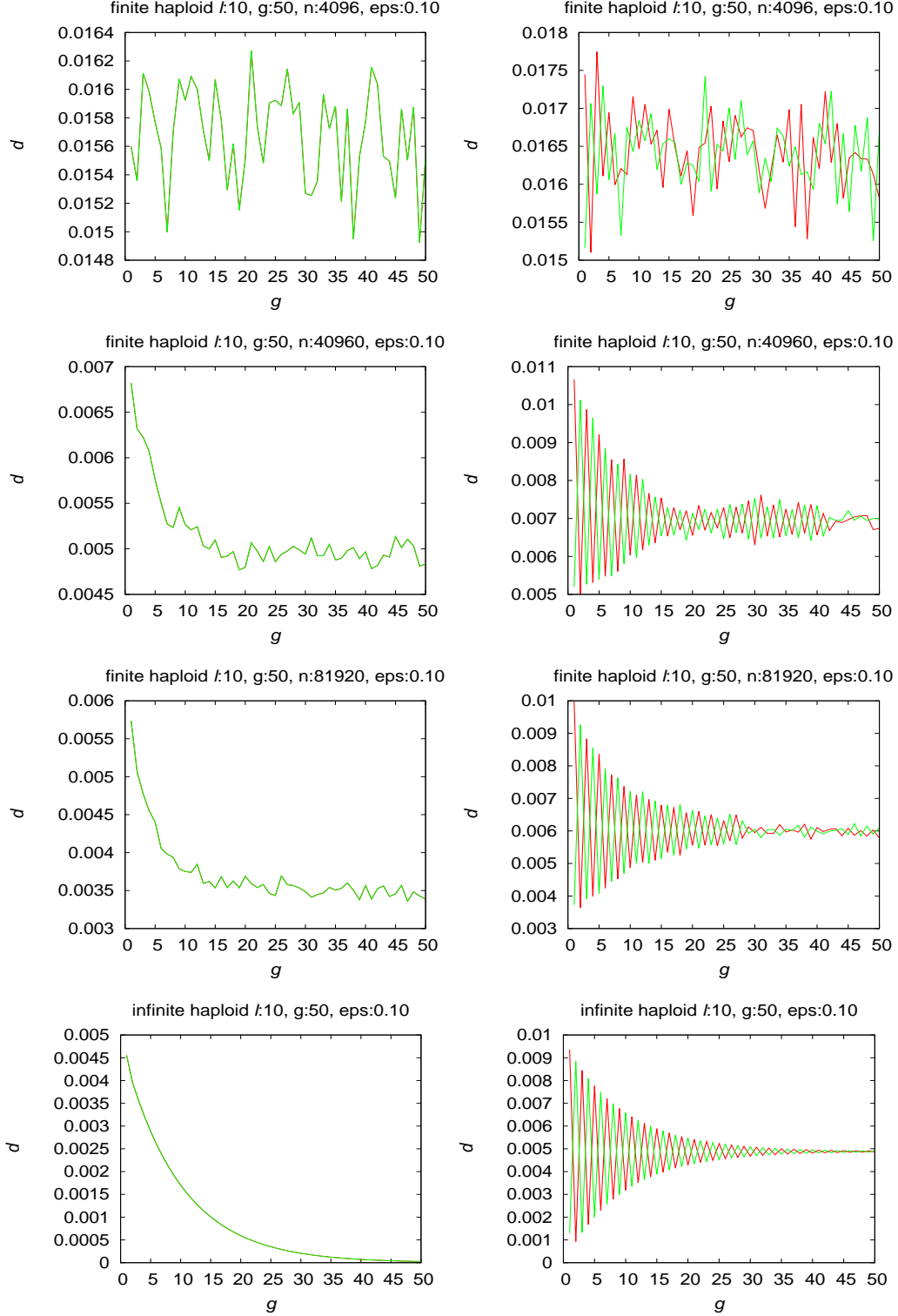


Figure 3.38: Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 10$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

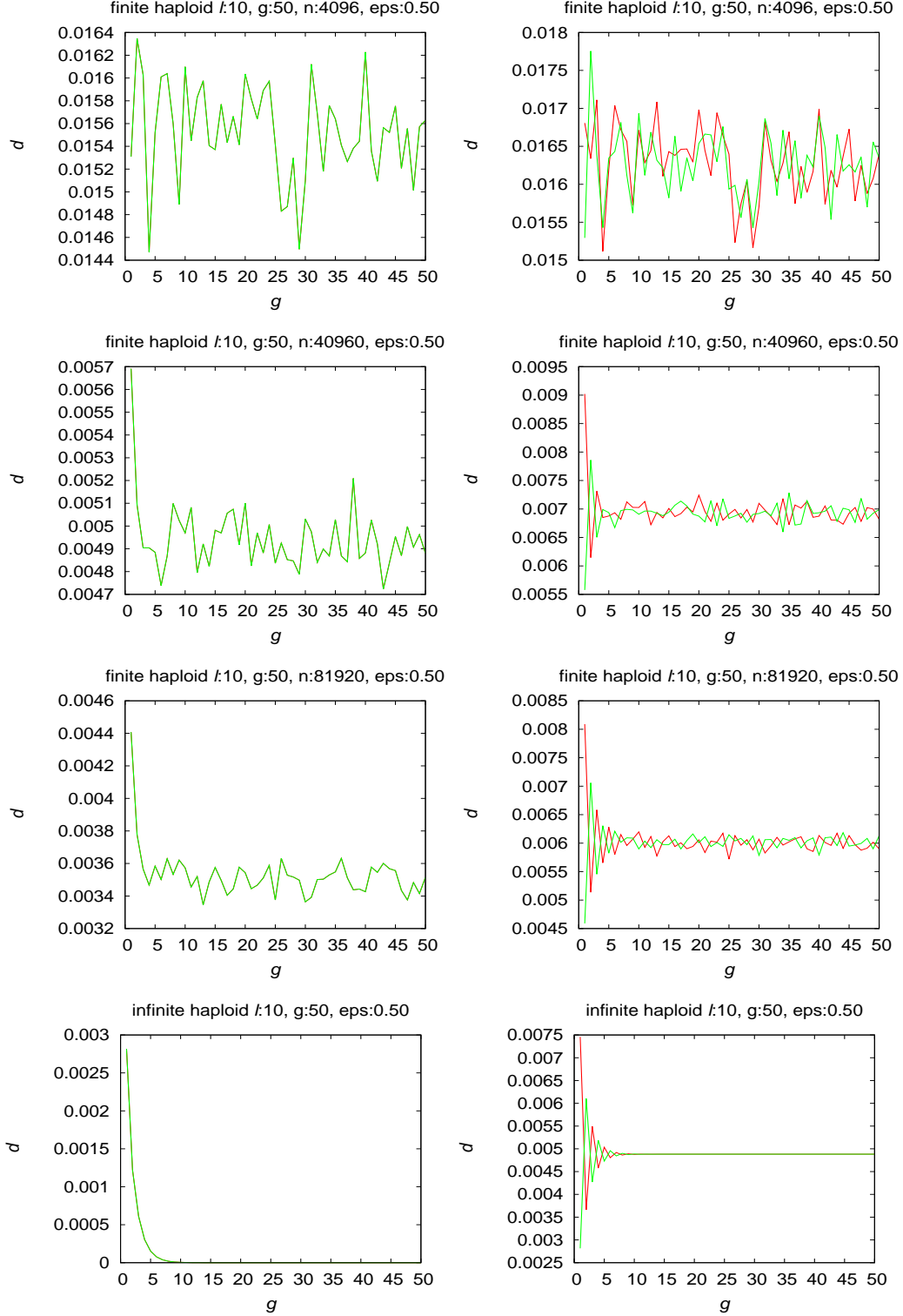


Figure 3.39: Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 10$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

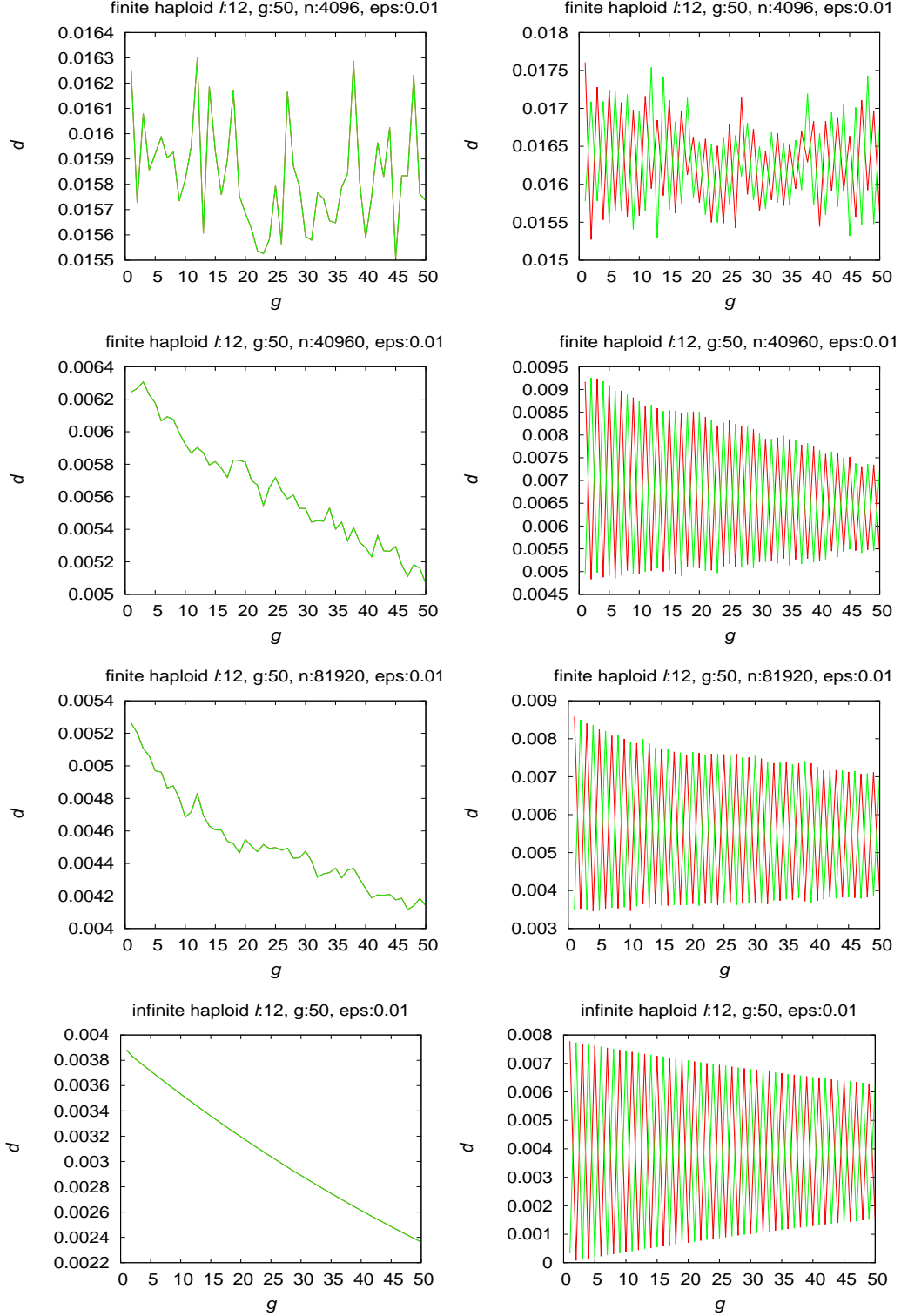


Figure 3.40: Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 12$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

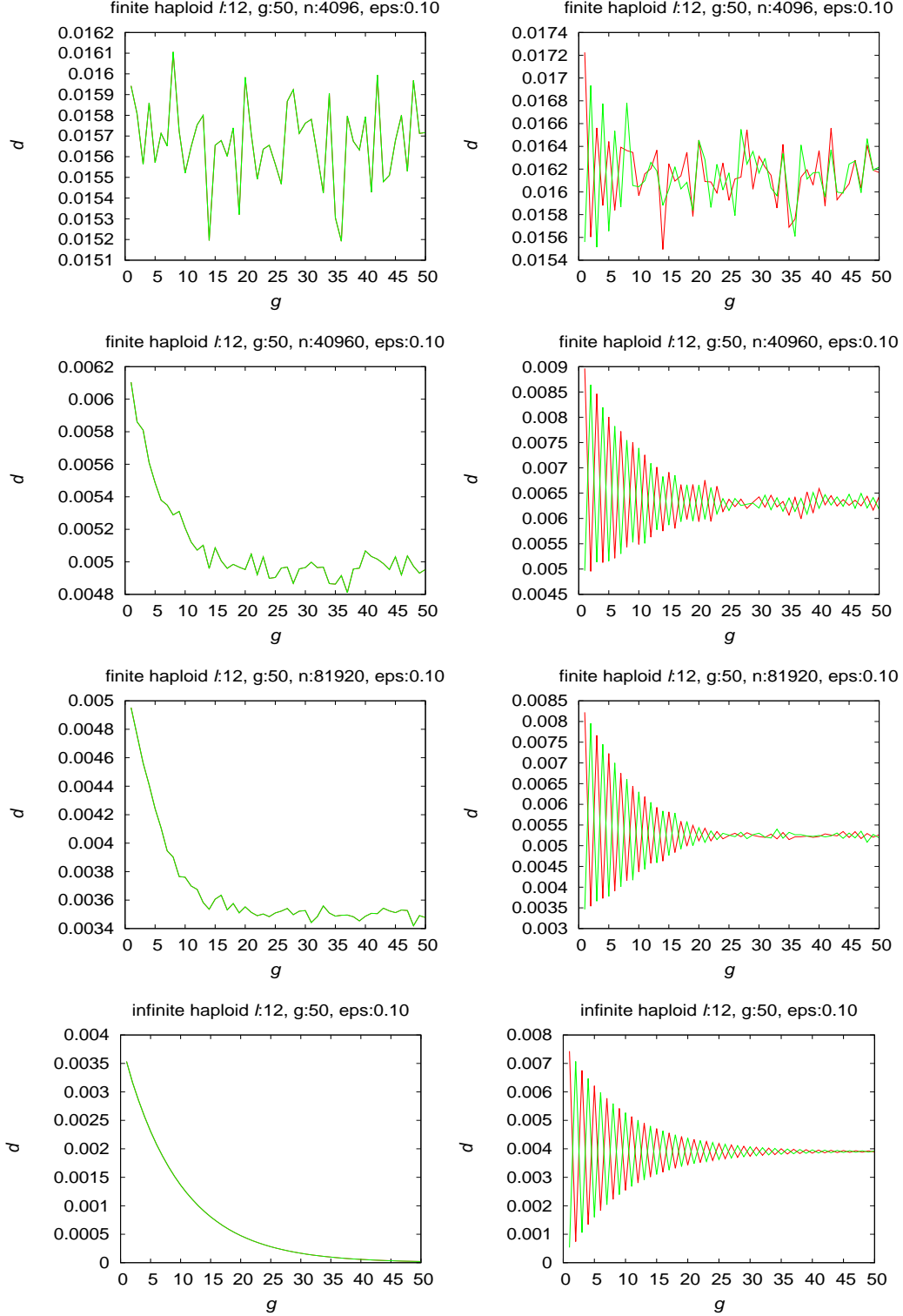


Figure 3.41: Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 12$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

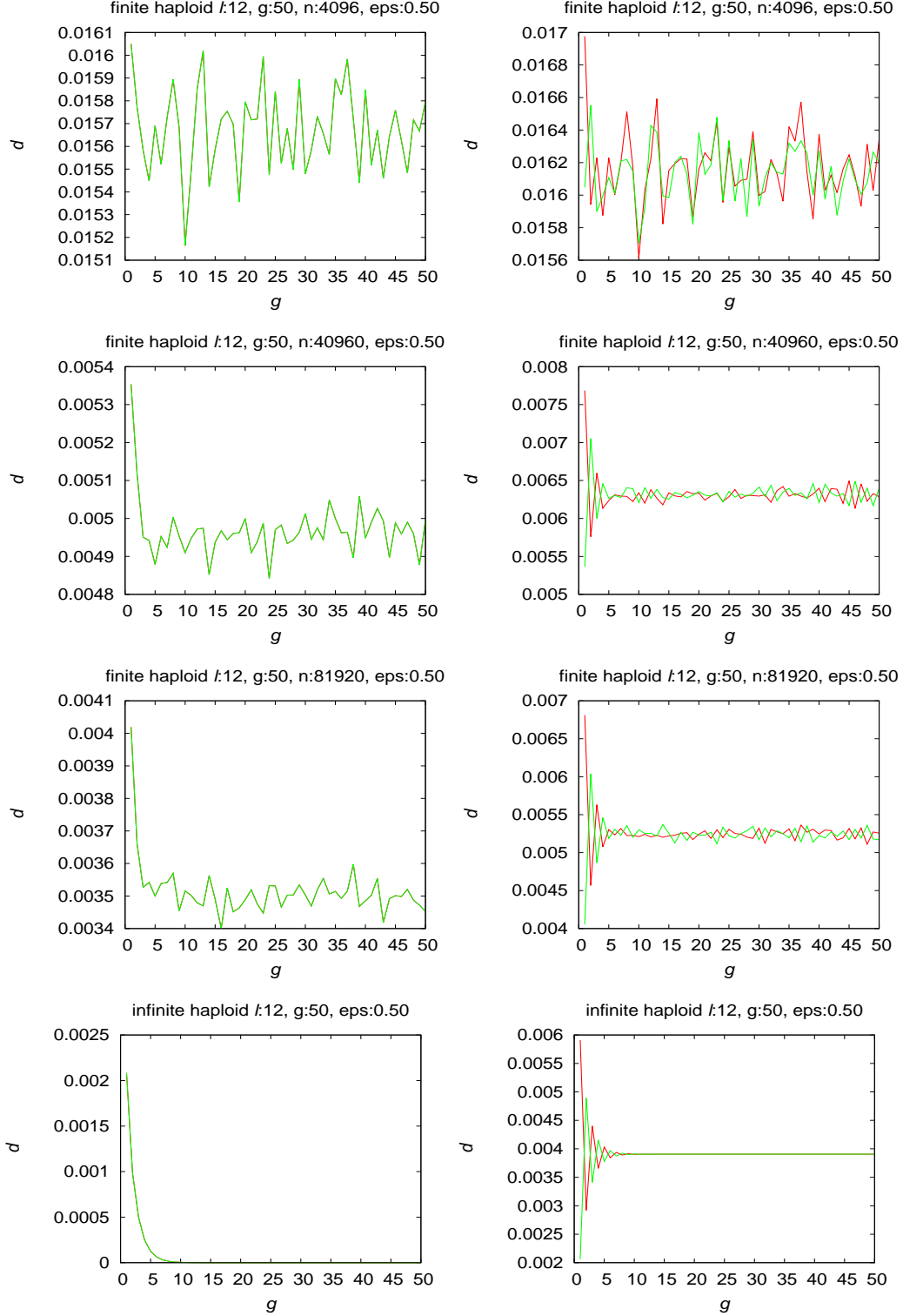


Figure 3.42: Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 12$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

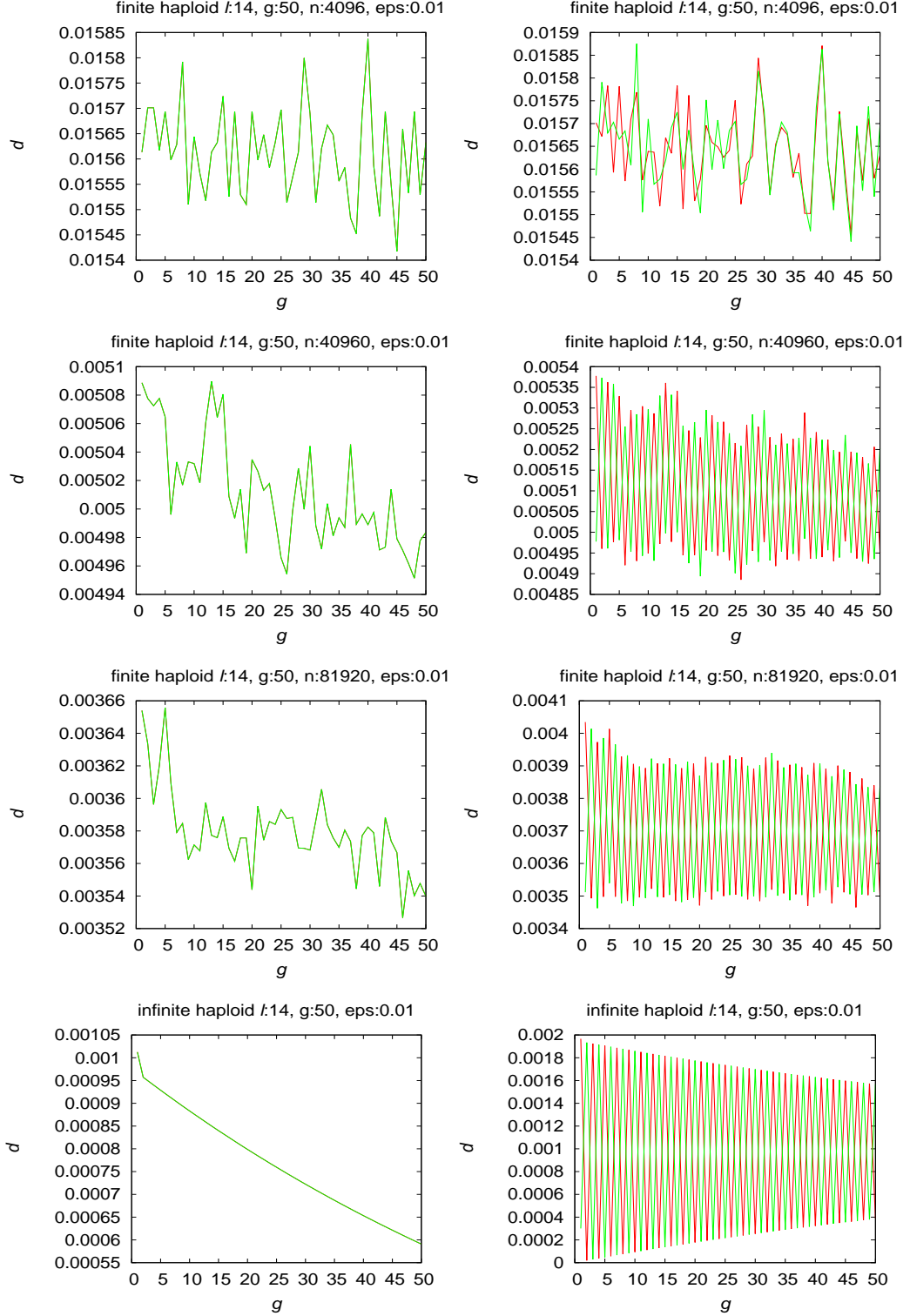


Figure 3.43: Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 14$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

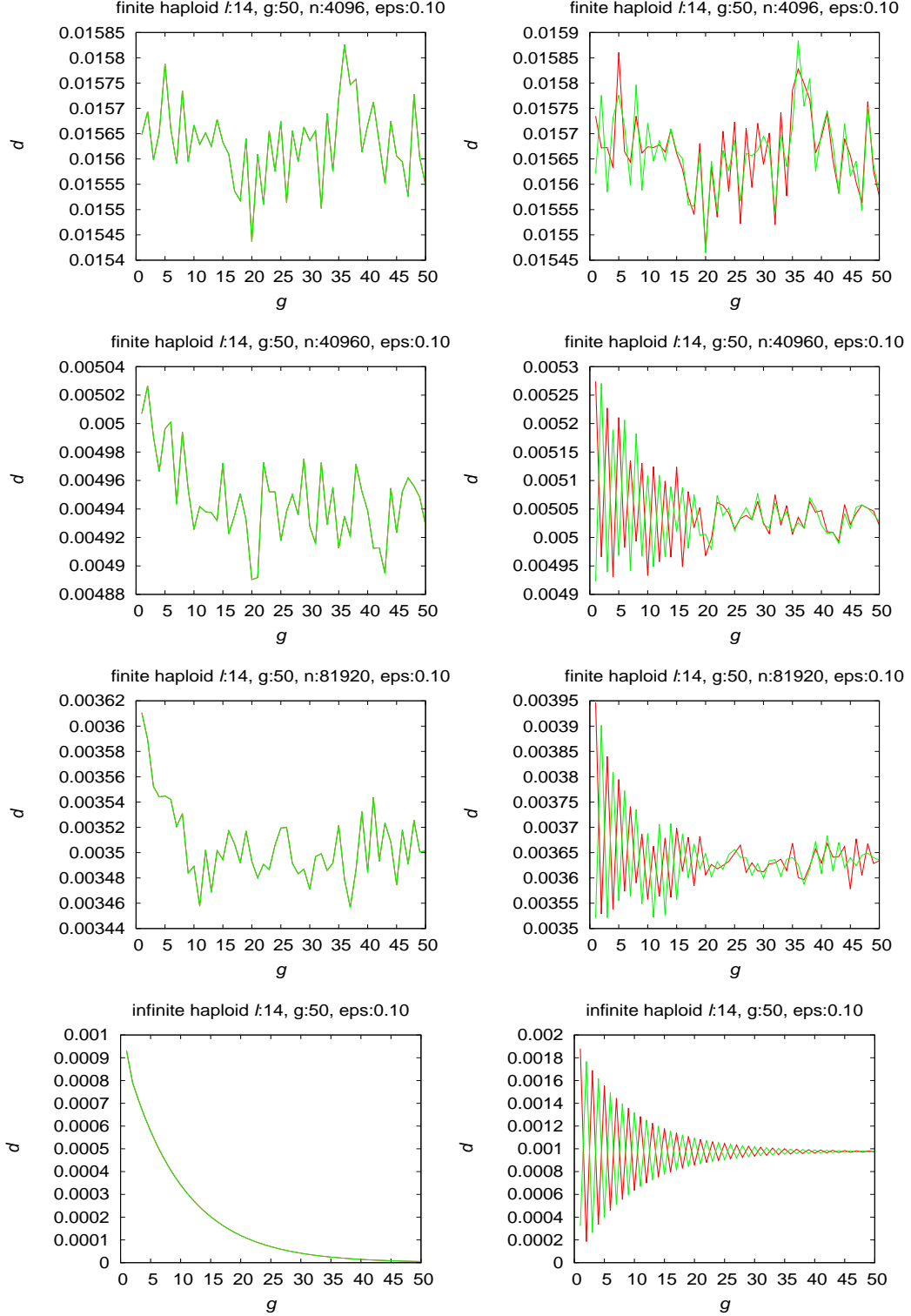


Figure 3.44: Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 14$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

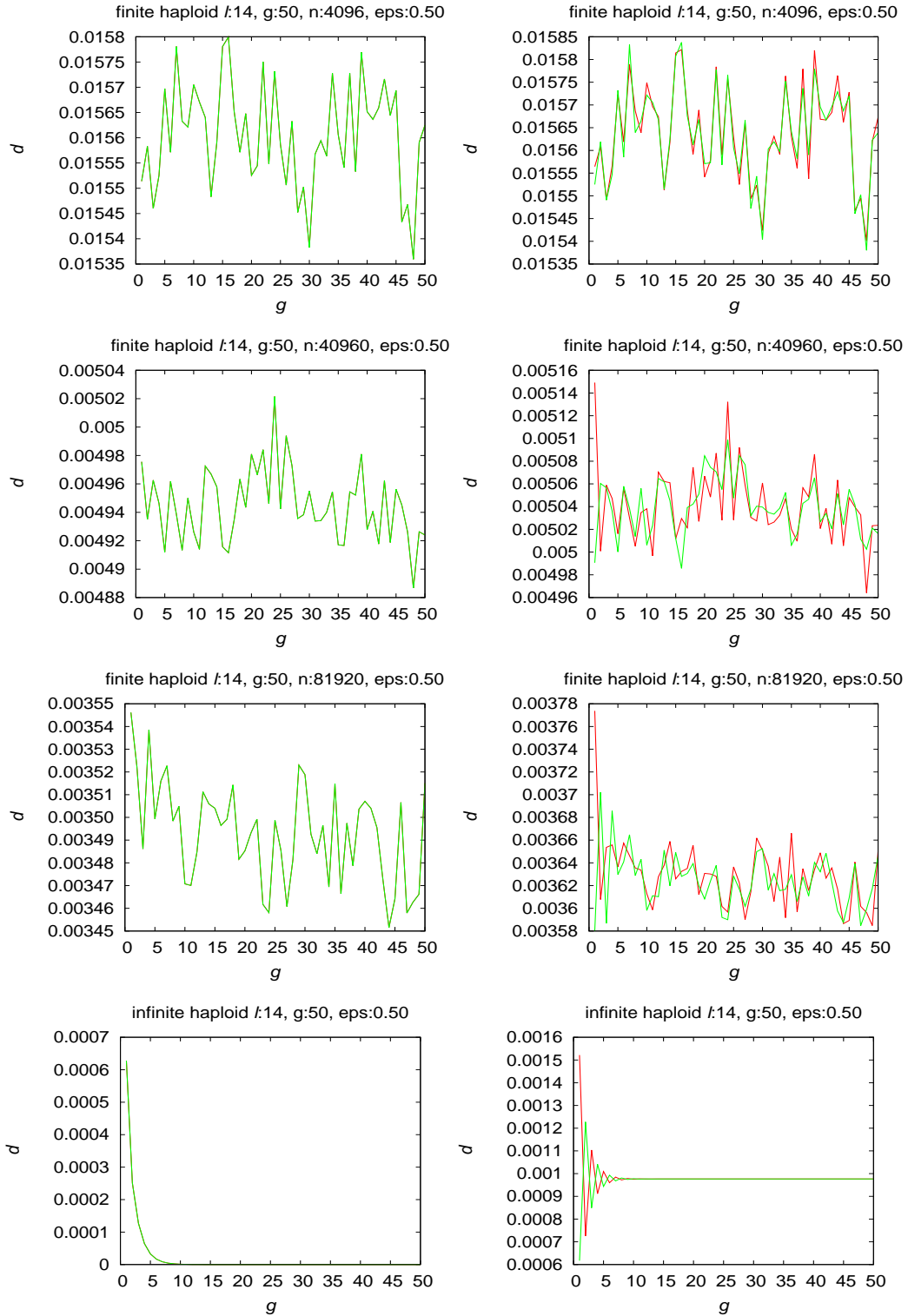


Figure 3.45: Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 14$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

The left column of figures 3.34 through 3.45 shows distance of finite and infinite haploid populations to limit z^* (limit with violation in crossover distribution χ). The right column shows distance of finite and infinite haploid populations to non-violation limits p^* and q^* . Graphs in right column give picture of oscillating behavior of haploid population given violation. Graphs in left column in figures 3.34 through 3.45 show distance between finite population and limit z^* (limit with violation in χ distribution) for haploid case.

Both finite and infinite populations oscillate given violation. However, ripples die out quickly, and oscillation ceases for infinite populations. In the finite population case, although ripple amplitudes decrease, oscillation didn't die out completely.

Change in oscillating behavior of population with change in ϵ values $\{0.01, 0.1, 0.5\}$ were also studied. Results show ripples damp out faster with increase ϵ . With smaller values of ϵ , oscillations were sharper and as value of ϵ increased, rate of damping increased. With $\epsilon = 0.01$, oscillation is clearly visible and ripples are sharper; with $\epsilon = 0.1$ oscillation was visible but ripples were damping out quickly; with $\epsilon = 0.5$, oscillation was very minimal or not visible at all.

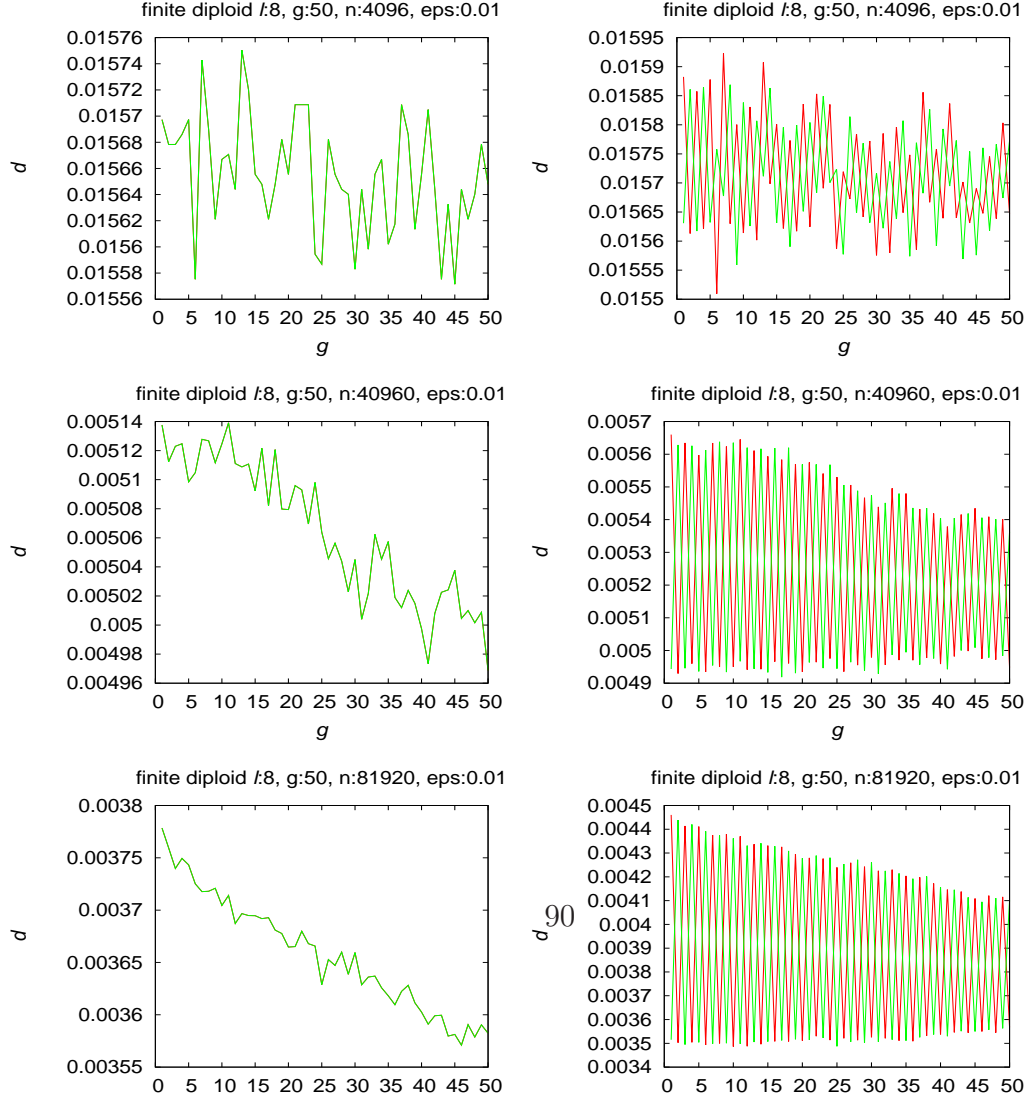
The distance between finite population and limit z^* (limit with violation in χ distribution) decreases as finite population size increases, and finite population shows behavior similar to infinite population behavior as finite population reach large number. Simulation results show infinite population converges to limit z^* quicker with increase in ϵ .

The distance data for haploid population in case of violation in χ distribution with different values of ϵ for different finite population size N are tabulated in table 3.5.

From table 3.5, average distance calculated for finite population size 4096 is 0.0158, for size 40960 is 0.0055 and for size 81920 is 0.0040. These results show average distance between finite population and limit z^* closely follows expected single step distance between finite and infinite population given in 3.1. The distance decreased as $1/\sqrt{N}$.

Table 3.5: Experimental distance measured for violation in χ for haploids:
 ℓ is genome length, ϵ is error introduced to χ for violation, and average distance between finite and infinite populations is tabulated in last three columns.

ϵ	case	ℓ	N 4096	$=$ N 40960	$=$ N 81920
0.01	haploid	8	0.0186	0.0150	0.0115
		10	0.0158	0.0062	0.0051
		12	0.0158	0.0056	0.0045
		14	0.0156	0.0050	0.0036
0.1	haploid	8	0.0163	0.0061	0.0051
		10	0.0157	0.0051	0.0037
		12	0.0157	0.0051	0.0037
		14	0.0156	0.0049	0.0035
0.5	haploid	8	0.0156	0.0051	0.0036
		10	0.0155	0.0049	0.0035
		12	0.0157	0.0050	0.0035
		14	0.0156	0.0049	0.0035



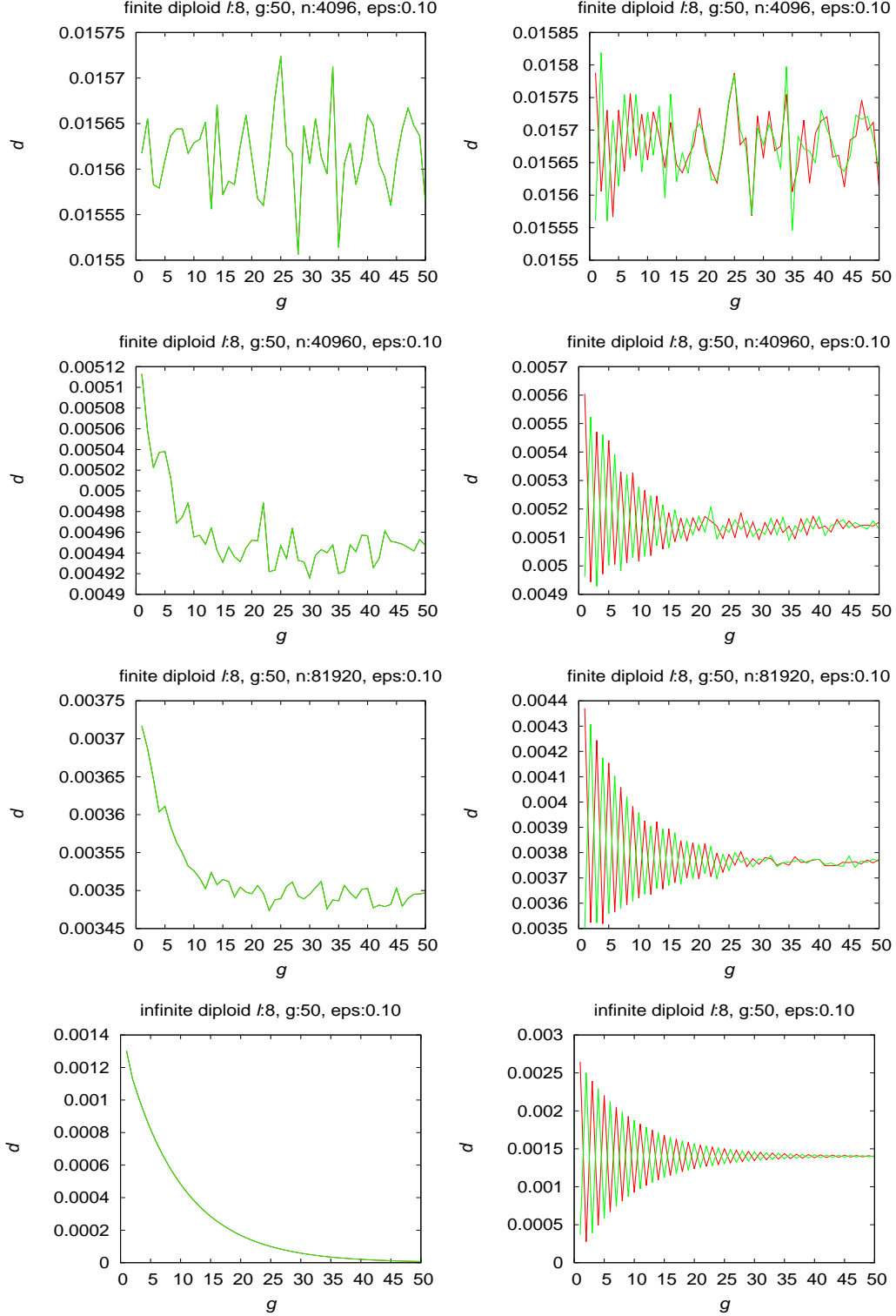


Figure 3.47: Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 8$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

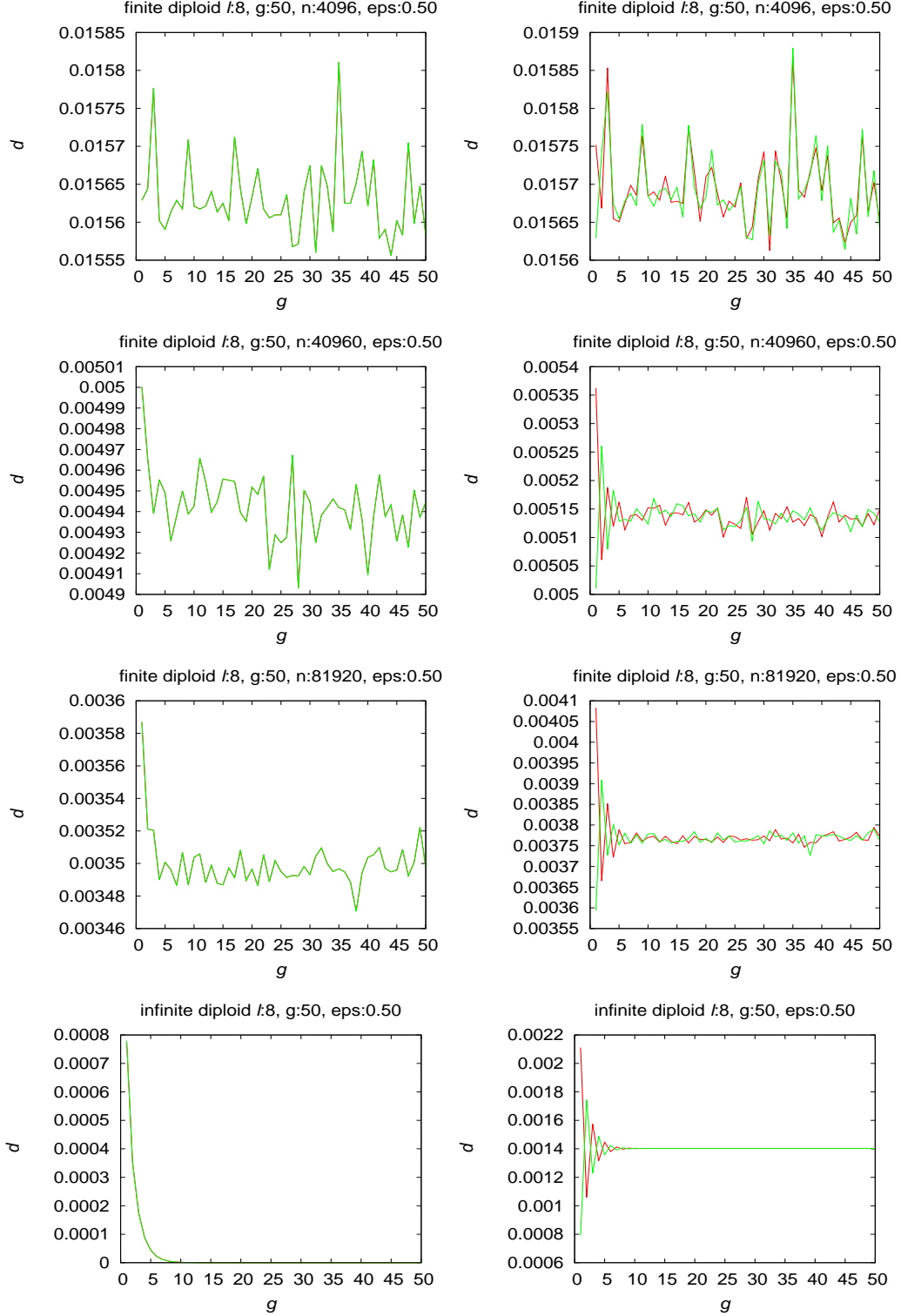


Figure 3.48: Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 8$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

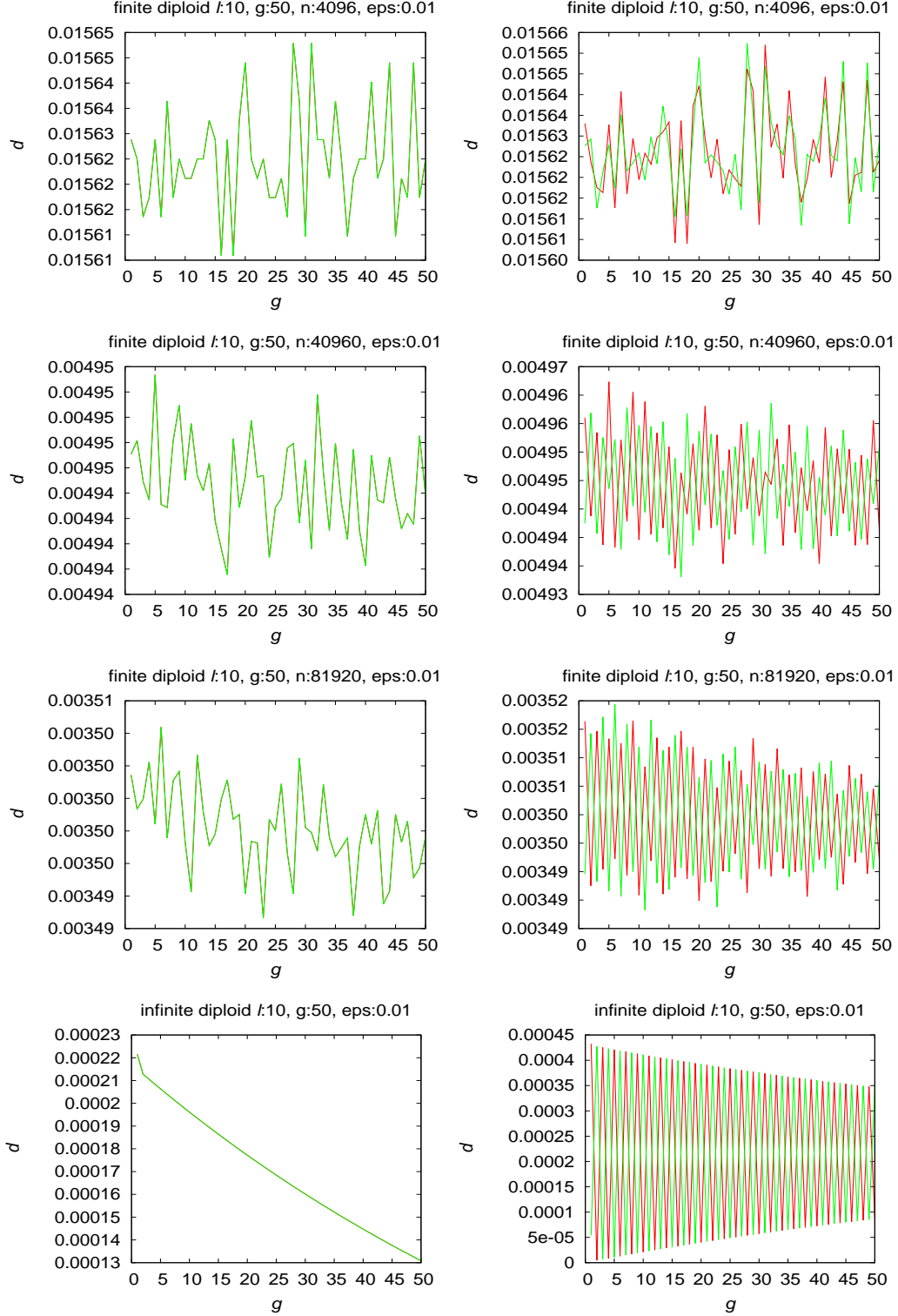


Figure 3.49: Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 10$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

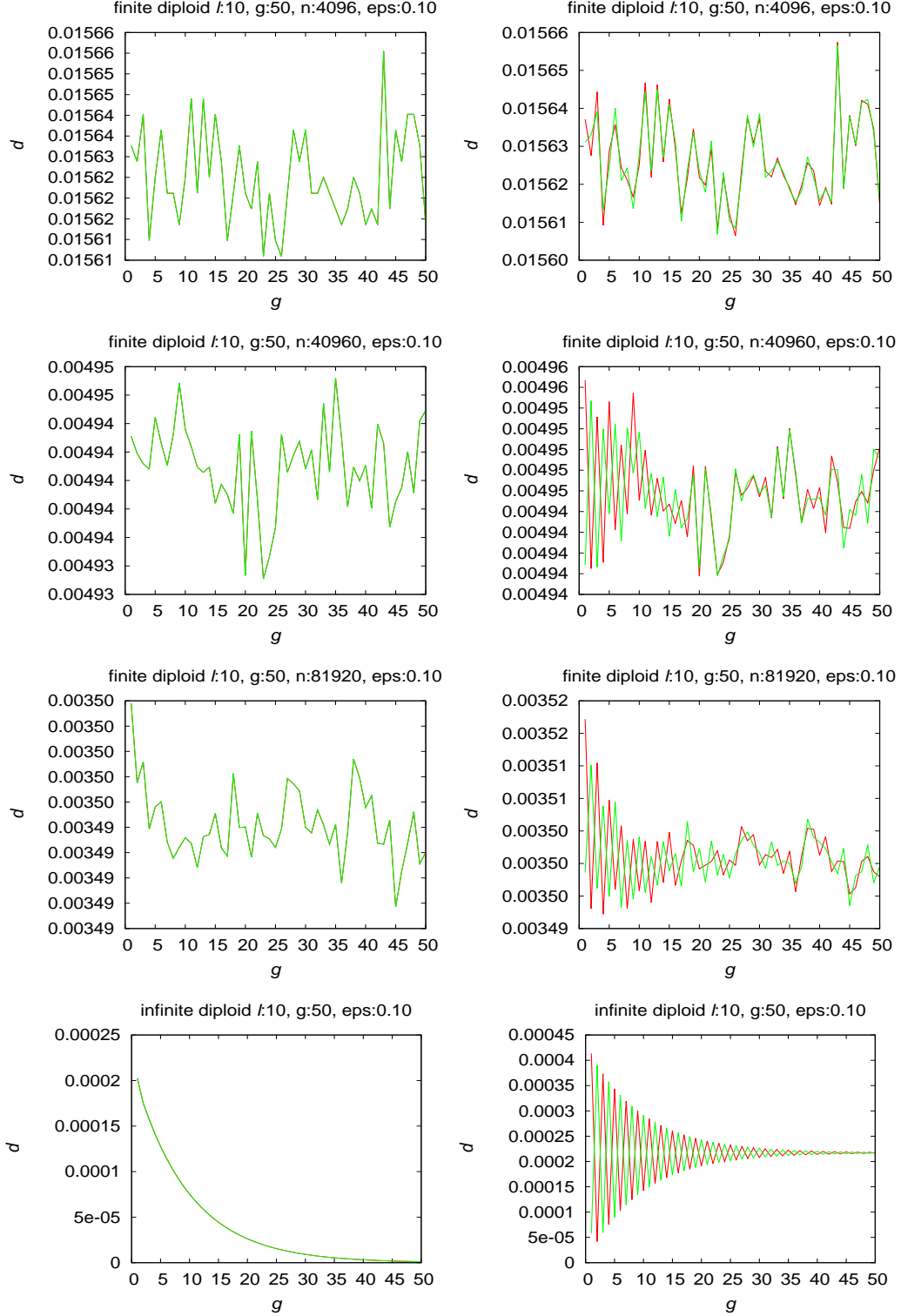


Figure 3.50: Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 10$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

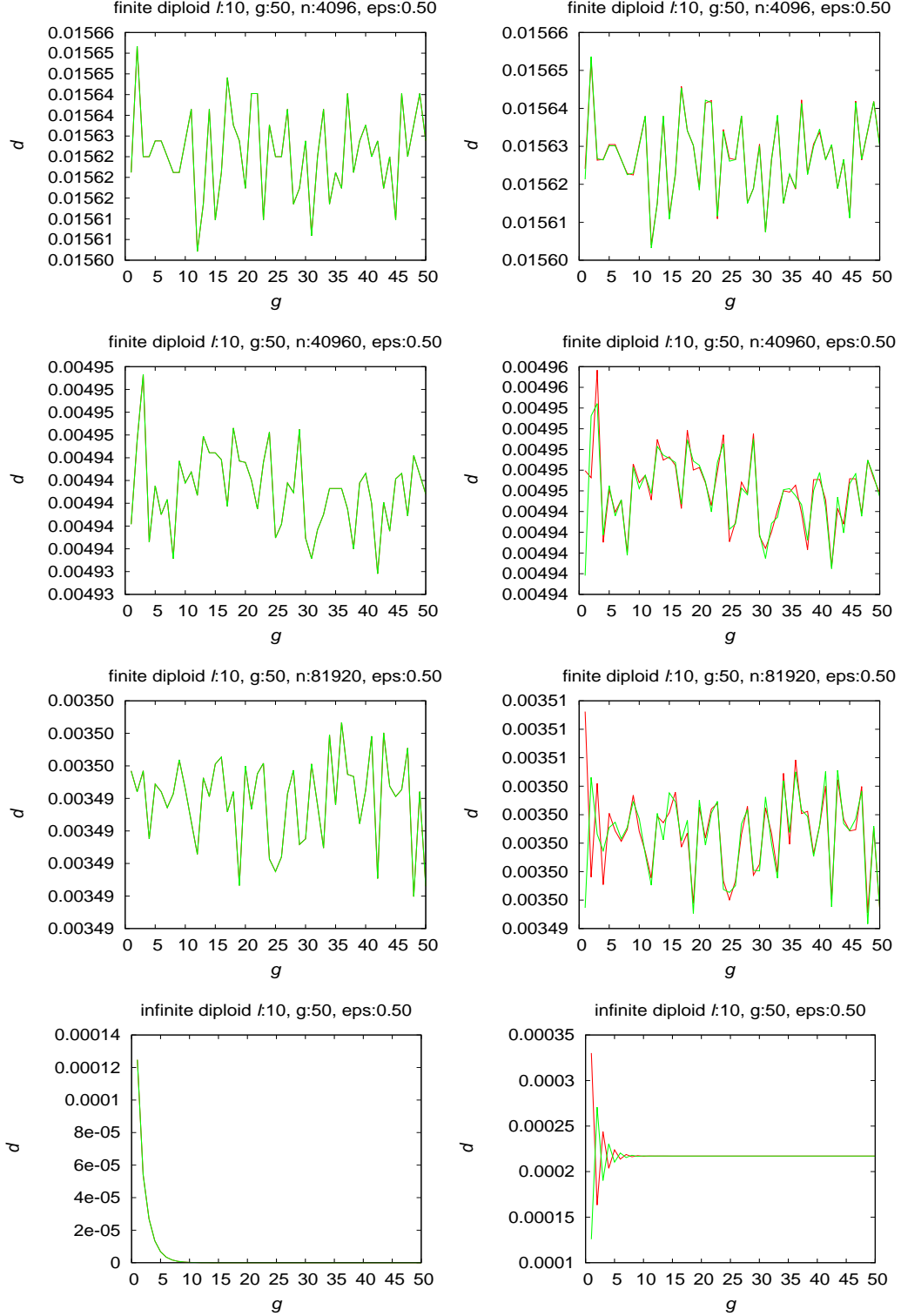


Figure 3.51: Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 10$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

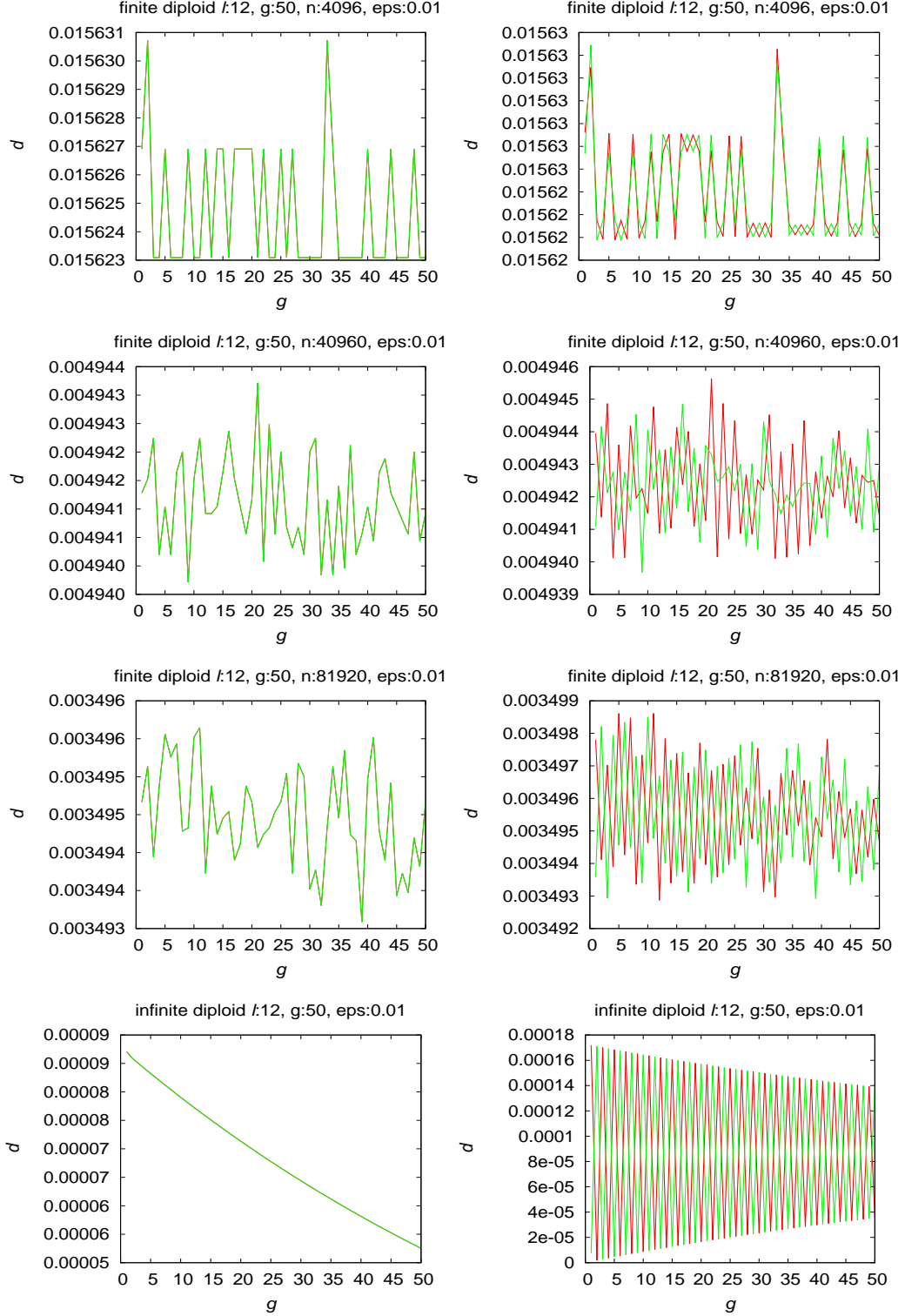


Figure 3.52: Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 12$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

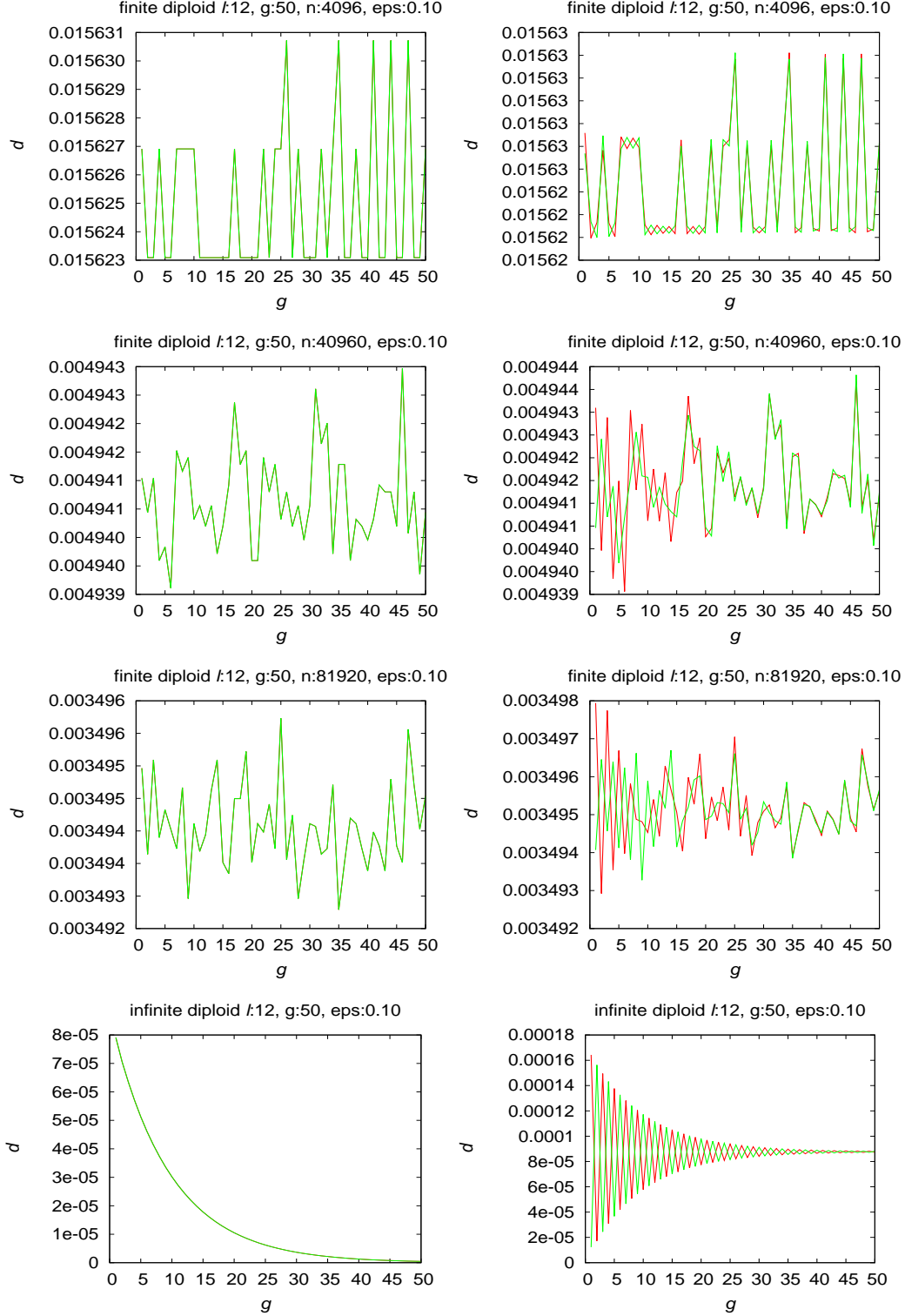


Figure 3.53: Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 12$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

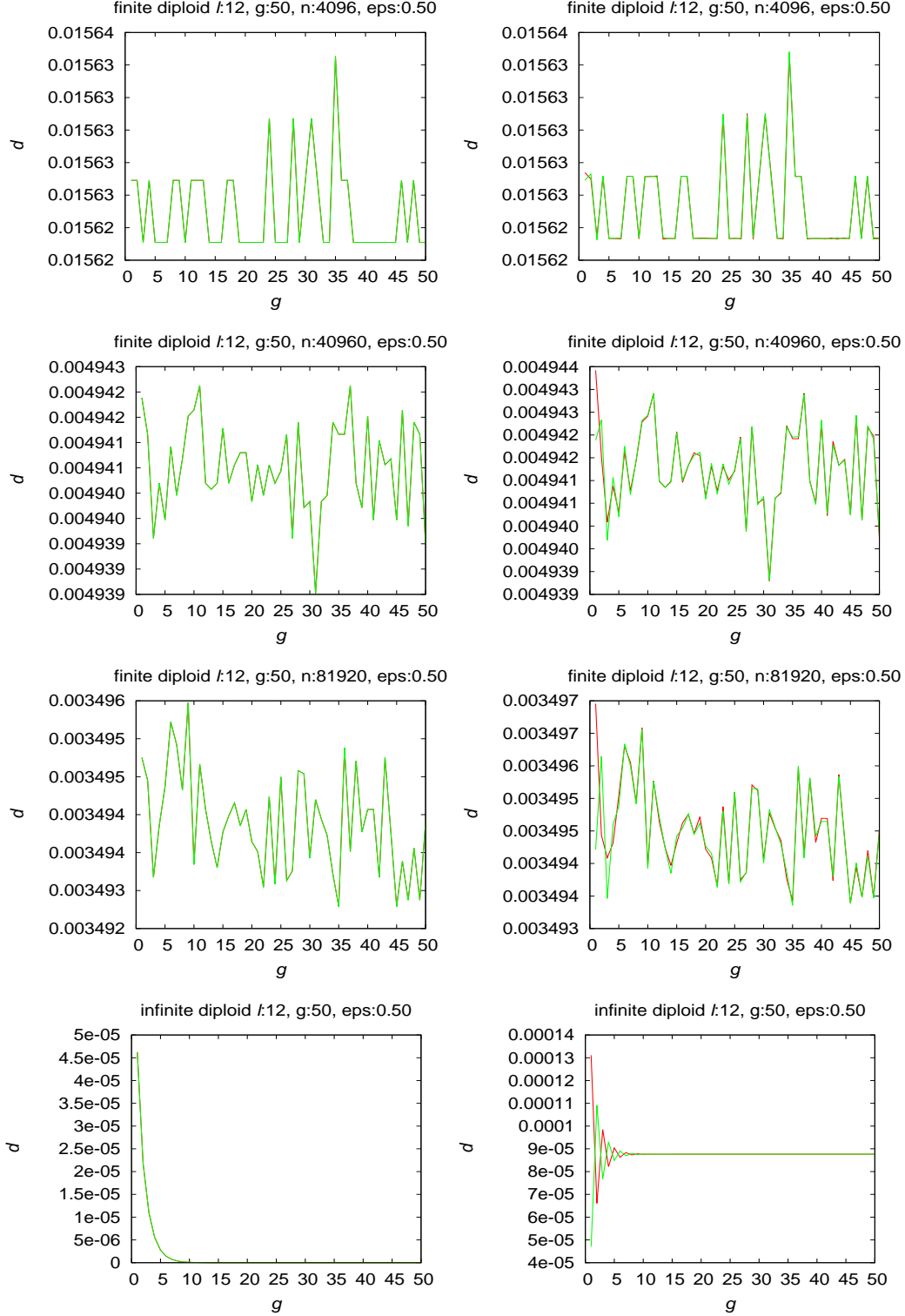


Figure 3.54: Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 12$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

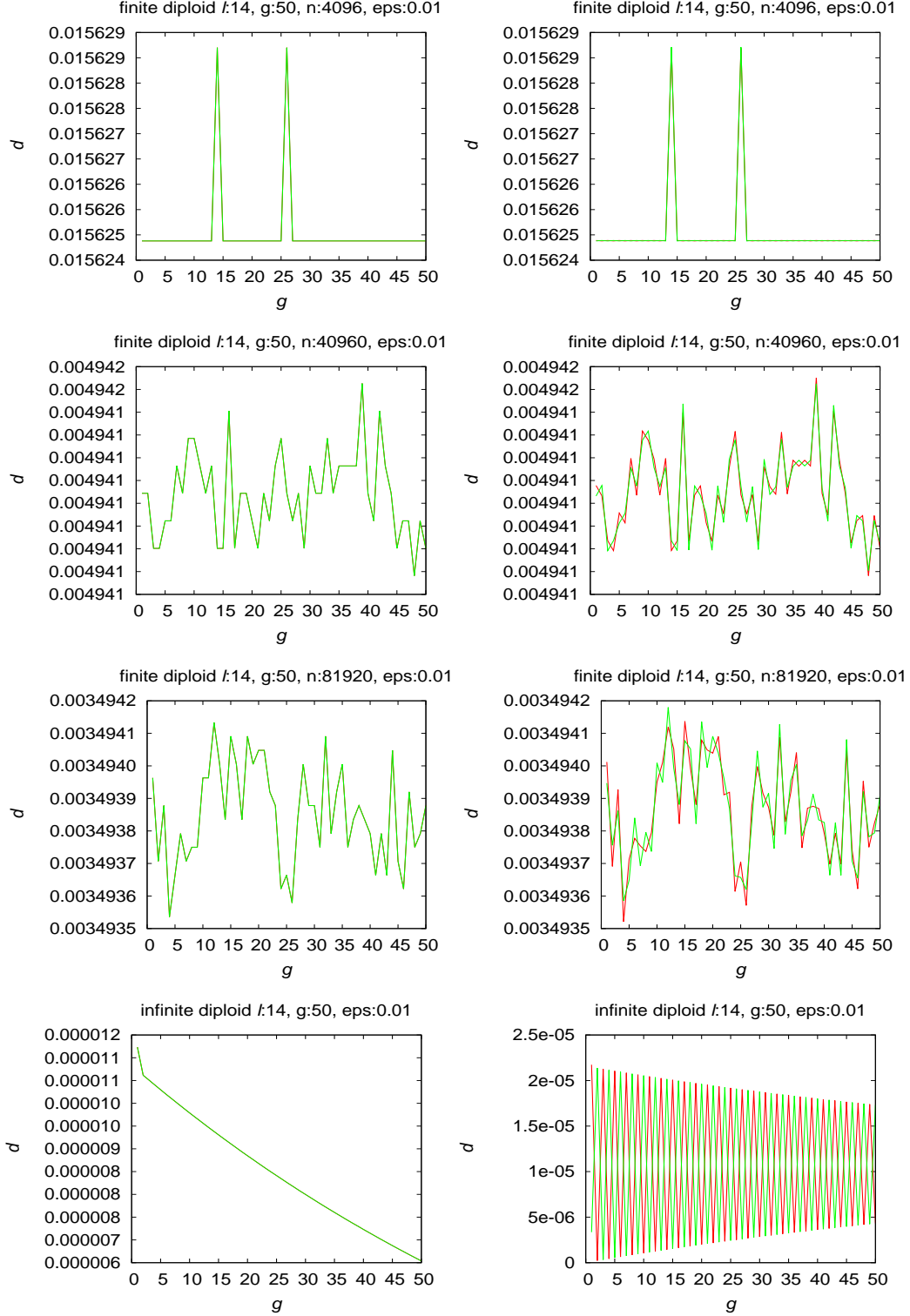


Figure 3.55: Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 14$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

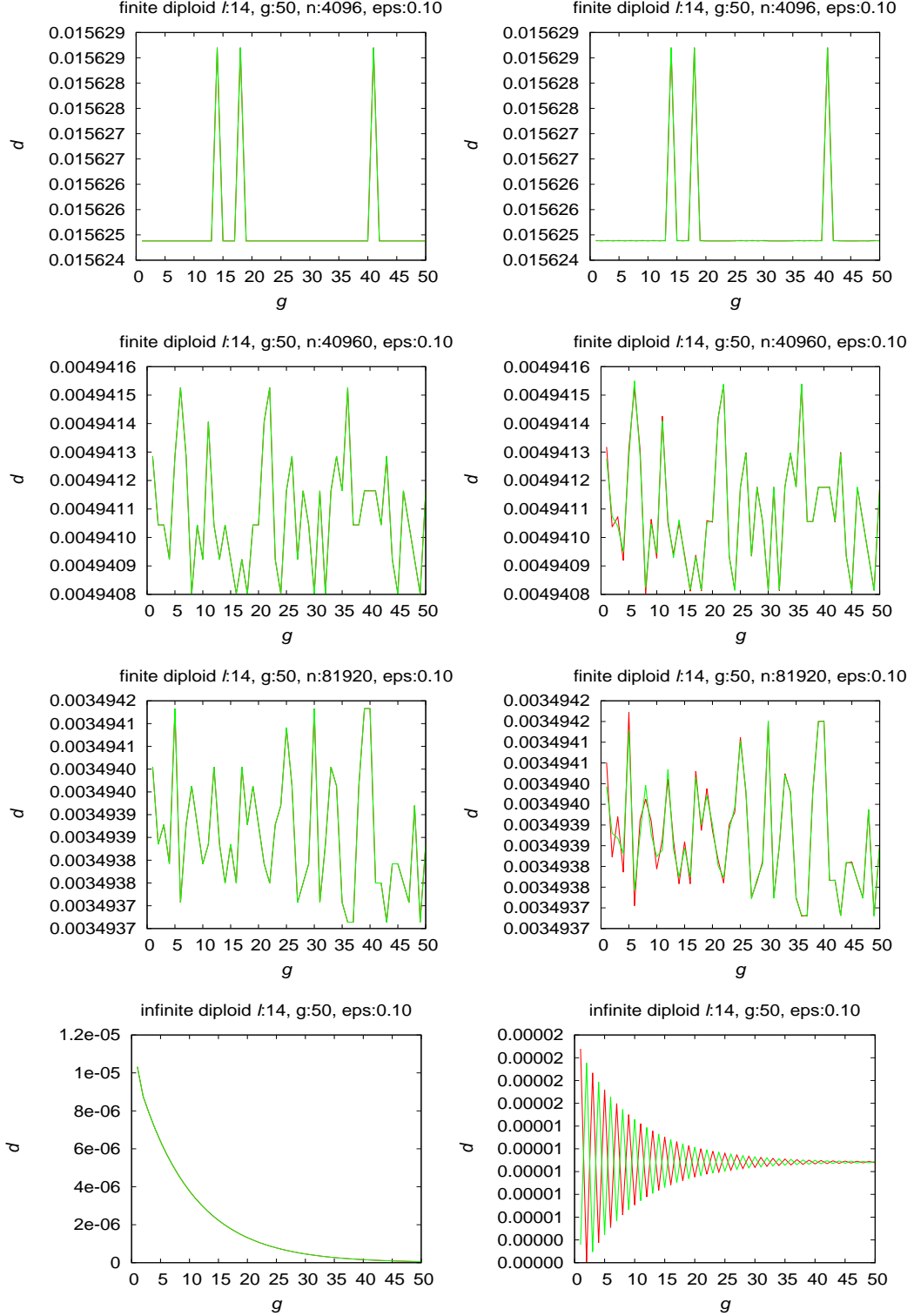


Figure 3.56: Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 14$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

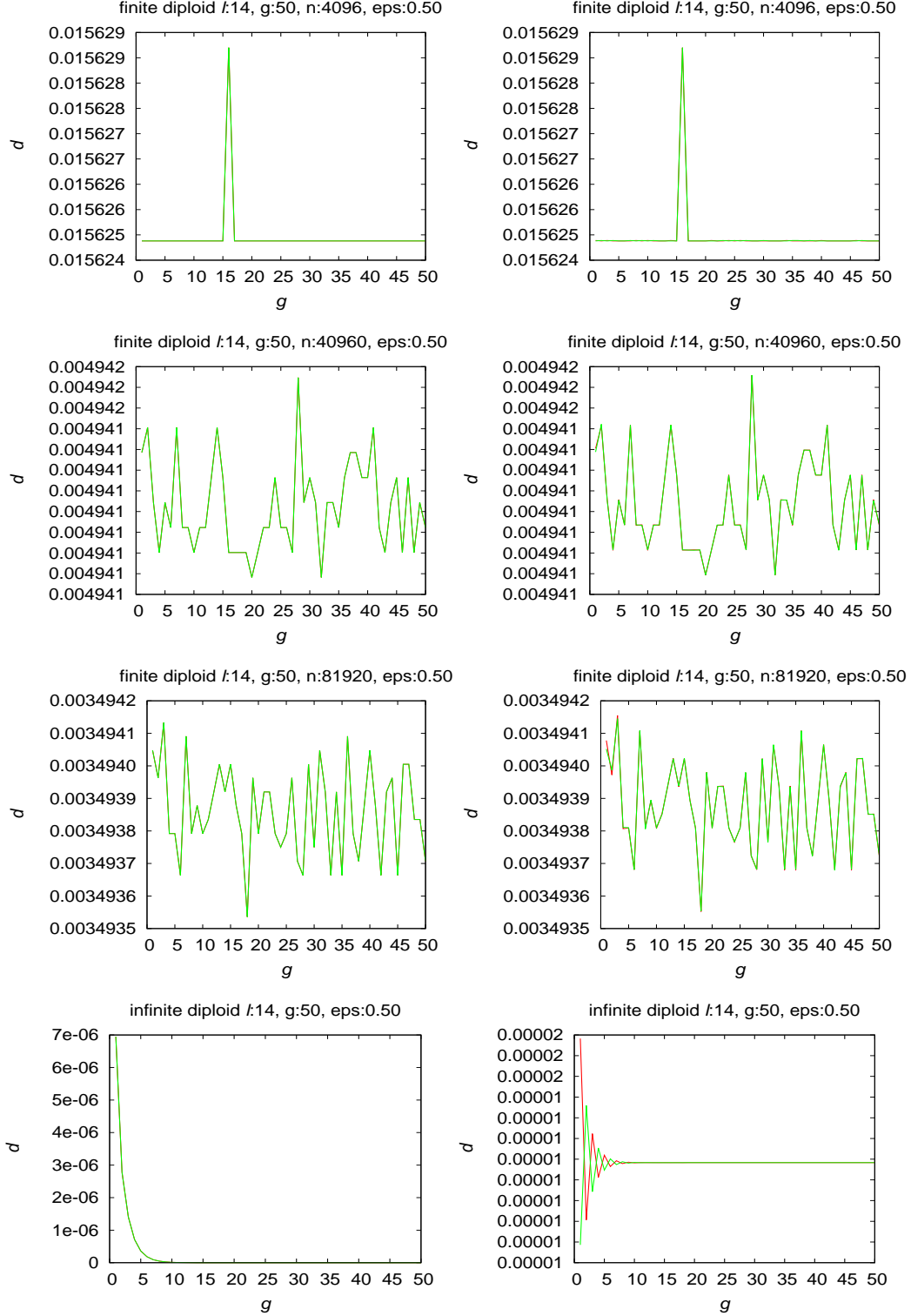


Figure 3.57: Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 14$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

The left column of figures 3.46 through 3.57 shows distance of finite and infinite diploid populations to limit z^* (limit with violation in *chi*). The right column shows distance of finite and infinite diploid populations to non-violation limits p^* and q^* . Graphs in right column give picture of oscillating behavior of diploid population given violation. Graphs in left column in figures 3.46 through 3.57 show distance between finite population and limit z^* (limit with violation in χ distribution) for diploid case.

Likewise in haploid population, both finite and infinite diploid populations oscillate given violation. However, ripples die out quickly, and oscillation ceases for infinite populations. In the finite population case, although ripple amplitudes decrease, oscillation didn't die out completely.

Change in oscillating behavior of population with change in ϵ values $\{0.01, 0.1, 0.5\}$ were also studied. Results show ripples damp out faster with increase ϵ . With smaller values of ϵ , oscillations were sharper and as value of ϵ increased, rate of damping increased. With $\epsilon = 0.01$, oscillation is clearly visible and ripples are sharper; with $\epsilon = 0.1$ oscillation was visible but ripples were damping out quickly; with $\epsilon = 0.5$, oscillation was very minimal or not visible at all.

The distance between finite population and limit z^* (limit with violation in χ distribution) decreases as finite population size increases, and finite population shows behavior similar to infinite population behavior as finite population reach large number. Simulation results show infinite population converges to limit z^* quicker with increase in ϵ .

The distance data for diploid population in case of violation in χ distribution with different values of ϵ for different finite population size N are tabulated in table 3.6.

From table 3.6, average distance calculated for finite population size 4096 is 0.0158, for size 40960 is 0.0055 and for size 81920 is 0.0040. These results show average distance between finite population and limit z^* closely follows expected single step distance between finite and infinite population given in 3.1. The distance decreased as $1/\sqrt{N}$.

Table 3.6: Experimental distance measured for violation in χ : ℓ is genome length, ϵ is error introduced to χ for violation, and average distance between finite and infinite populations is tabulated in last three columns.

ϵ	case	ℓ	N 4096	= N 40960	= N 81920
0.01	diploid	8	0.0156	0.0051	0.0036
		10	0.0156	0.0049	0.0035
		12	0.0156	0.0049	0.0035
		14	0.0156	0.0049	0.0035
0.1	diploid	8	0.0156	0.0050	0.0035
		10	0.0156	0.0049	0.0035
		12	0.0156	0.0049	0.0035
		14	0.0156	0.0049	0.0035
0.5	diploid	8	0.0156	0.0049	0.0035
		10	0.0156	0.0049	0.0035
		12	0.0156	0.0049	0.0035
		14	0.0156	0.0049	0.0035

Violation introduced to the crossover distribution χ creates new masks (different than in case of without violation) to be used in crossover during transmission. With small ϵ , the probability of using the new masks available due to violation is very small and those masks might not be used at all during crossover in finite population, but with higher values of ϵ , those new masks have higher chance of usage during crossover which cause oscillation to damp out quickly or cause no oscillation at all.

From graphs on the right in figures 3.34 through 3.45, and 3.46 through 3.57, haploid population shows sharper and more distinct oscillations for same size of population size than diploid population; diploid population requires larger population size to exhibit sharper oscillations. And from graphs on the left in figures 3.34 through 3.45, and 3.46 through 3.57, diploid population converges to limit z^* more quickly than haploid population.

3.6 Summary

In this chapter, we described limits predicted by Vose for infinite population, and necessary and sufficient condition for convergence to periodic orbits. Through experiment, we showed finite population also oscillate, and converge to infinite population behavior as population size increases. Then we studied through simulations the effect on infinite and finite population behavior of violating the condition for convergence to periodic orbits. Infinite population ceases to oscillate when the condition for convergence to periodic orbits is violated, but finite population continued to approximately oscillate for small values of ϵ . For smaller values of ϵ , finite population does not get aware of violation because the probability of using new masks created in the mutation distribution μ and the crossover distribution μ due to violation is very low, and finite population follows behavior of infinite population without violation in the condition for convergence to periodic orbits.

Chapter 4

Conclusion

This research shows how Vose's haploid model for Genetic Algorithms extends to the diploid case, improving the computation of infinite population evolutionary trajectories by significantly reducing the time and space used. Efficiency is achieved through decoupling haploid evolution from the evolution of infinite diploid populations and employing Walsh transform methods to compute the effects of mask-based crossover and mutation.

Simulations are thereby made feasible which otherwise would require excessive resources, as illustrated through computations exploring the convergence rate of finite population short-term behavior. Results agree with the expected rate of convergence for the single-step haploid case; distance is inversely proportional to square root of population size.

Evolutionary limits predicted by Vose for infinite population were explored and analysed. Simulations showed that when the necessary condition for oscillation in infinite populations is met, finite populations also show oscillating behavior, and approximately converge to infinite population evolutionary limits in the short term. When the condition is violated, infinite populations ceases to oscillate, but finite populations continue to oscillate if the violation is small.

In this research, we did not consider fitness factor for selection for simplicity of model. In future, we plan to extend our work accomodating fitness factor in our model and investigate convergence of short-term finite population behavior to infinite population limits.

Bibliography

Bibliography

- Beauchamp, K. (1975). *Walsh functions and their applications*. Academic Press. [20](#)
- Bethke, A. D. (1980). *Genetic Algorithms As Function Optimizers*. PhD thesis, The University of Michigan. [4](#), [5](#)
- Cooley, J. W. and Tukey, J. W. (1965). An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, 19(90):297–301. [21](#)
- Crow, J. and Kimura, M. (1970). *An introduction to population genetics theory*. New York, Evanston and London: Harper & Row. [28](#)
- Geiringer, H. (1944). On the probability theory of linkage in mendelian heredity. *Ann. Math. Stat.*, 15(1):25–27. [14](#), [18](#)
- Goldberg, D. E. (1987). Simple genetic algorithms and the minimal, deceptive problem. *Genetic algorithms and simulated annealing*, 74:74–88. [4](#)
- Goldberg, D. E. (1989a). Genetic algorithms and walsh functions: Part i, a gentle introduction. *Complex systems*, 3(2):129–152. [5](#)
- Goldberg, D. E. (1989b). Genetic algorithms and walsh functions-partii: Deception and its analysis. *Complex systems*, 3(153–171). [5](#)
- Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science*, 28(706):49–50. [14](#), [15](#)

- Holland, J. H. (1992). *Adaptation in natural and artificial systems*. Cambridge : MIT Press. [3](#), [4](#)
- Koehler, G. J. (1994). A proof of the vose-liepins conjecture. *Annals of Mathematics and Artificial Intelligence*, 10(4):409–422. [5](#)
- Koehler, G. J., Bhattacharyya, S., and Vose, M. D. (1997). General cardinality genetic algorithms. *Evol. Comput.*, 5(4):439–459. [5](#)
- Mendel, G. (1865). Versuche über pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn*, IV:3–47. [15](#)
- Minc, H. (1988). *Nonnegative Matrices*. A Wiley-Interscience Publication. [12](#)
- Mitchell, M. (1999). *An Introduction to Genetic Algorithms*. The MIT Press. [2](#), [4](#)
- Nix, A. E. and Vose, M. D. (1992). Modeling genetic algorithms with markov chains. *Annals of Mathematics and Artificial Intelligence*, 5(1):79–88. [5](#)
- Shanks, J. L. (1969). Computation of the fast walsh-fourier transform. *IEEE Trans. Comput.*, 18(5):457–459. [6](#), [21](#), [24](#)
- Vose, M. and Liepins, G. E. (1991). Punctuated equilibria in genetic search. *Complex systems*, 5(1):31–44. [4](#), [5](#)
- Vose, M. D. (1999). *The simple genetic algorithm: foundations and theory*, volume 12. MIT press. [5](#), [6](#), [7](#), [8](#), [10](#), [11](#), [17](#), [26](#), [30](#), [31](#)
- Vose, M. D. and Wright, A. H. (1998). The simple genetic algorithm and the walsh transform: Part i, theory. *Evol. Comput.*, 6(3):253–273. [6](#), [18](#), [19](#), [21](#), [22](#)
- Wikipedia, C. (2016a). Chebyshev’s inequality. [9](#)
- Wikipedia, C. (2016b). Jensen’s inequality. [10](#)
- Wikipedia, C. (2016c). Markov chain. [12](#), [48](#)

Appendix

Vita

Vita goes here...