To the Graduate Council:
I am submitting herewith a thesis written by Mahendra Duwal Shrestha entitled "Efficient 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

Judy D. Day

Hairong Qi

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 "Efficient 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

Judy D. Day

Hairong Qi

Accepted for the Council:

Dixie Thompson

Vice Provost and Dean of the Graduate School

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

Efficient Simulation Of A Simple Evolutionary System

A Thesis Presented for the

Master of Science

Degree

The University of Tennessee, Knoxville

Mahendra Duwal Shrestha

May 2017

To my loving

Father & Mother,

Whose love, affection, encouragment, and support made me able to stand where I am today

Acknowledgements

I thank Michael D. Vose, Hairong Qi, and Judy D. Day for serving as my committee members. I am especially grateful to my academic advisor Michael D. Vose for suggesting the topic of this research and for giving so generously of his time. I have utterly enjoyed our many hours of discussion related to this thesis and learned a lot from this.

This research would not be complete without support of Sergey Gavrilets from Ecology and Evolutionary Biology, who funded most of my graduate tuition through assistantship, and also provided access to the computing cluster Volos for running simulations. I wish to thank Sergey Gavrilets for his support.

Finally, I wish to thank my family and friends for putting up some encouraging words. Sometimes they had more belief in me than I had in myself.

"Truth is stranger than fiction." - Mark Twain

Efficient Simulation Of A Simple Evolutionary System

Abstract

An infinite population model is considered for diploid evolution under the influence of crossing over and mutation. The evolution equations show how Vose's haploid model for Genetic Algorithms extends to the diploid case, thereby making feasible simulations which otherwise would require excessive resources. This is illustrated through computations confirming the convergence of finite diploid population short-term behaviour to the behaviour predicted by the infinite diploid model. The results show the distance between finite and infinite population evolutionary trajectories can decrease in practice like the reciprocal of the square root of population size.

Under necessary and sufficient conditions (NS) concerning mutation and crossover, infinite populations show oscillating behavior. We explore whether finite populations can also exhibit oscillation or approximate oscillation. Simulation results confirm that approximate finite population oscillation is possible when NS are satisfied.

We also investigate the robustness of finite population oscillation. We show that when the part of NS concerning mutation is violated, the Markov chain which models finite population evolution is regular, and perfect oscillation should not occur. However, our simulation results show finite population approximate oscillation can

occur even though the Markov chain is regular. Finite populations can also exhibit approximate oscillating behavior when the part of NS concerning crossover is violated.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

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

1.1 Notation

Mathematical notations, some standard as well as some non-standard, are introduced here (we borrow from and summarize – with permission – Vose (1999)).

A tuple, which is dennoted by angle brackets $\langle \cdots \rangle$, is to be regarded as a column vector. $\mathbf{1}$ denotes the column vector of all 1s. Superscript T indicates transpose. 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 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 used to specify a closed interval of real numbers as well as to denote *Iverson bracket*. *Iverson bracket* is an indicator function: if *expr* is an expression, then $[expr]$ denotes 1 if *expr* is true, and 0 otherwise.

sup indicates the supremum which is the least upper bound. *inf* indicates the infimum, that is, the greatest lower bound.

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 $x = \langle x_0, x_1, \cdots, x_{\ell-1} \rangle$. The additive identity of $\mathcal{R}$ is $\mathbf{0}$ and the multiplicative identity is $\mathbf{1}$. Let $\overline{g}$ abbreviate $\mathbf{1} + g$. Except when explicitly indicated otherwise, operations acting on elements of $\mathcal{R}$ are as defined in this paragraph. In particular, $g\overline{g} = \mathbf{0} = g + g$, $g^2 = g$, $g + \overline{g} = \mathbf{1}$ for all $g \in \mathcal{R}$.

1.2 Background

1.2.1 Genetic Algorithm

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. 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)). Basic elements of a simple style 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 value (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 exchange 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.

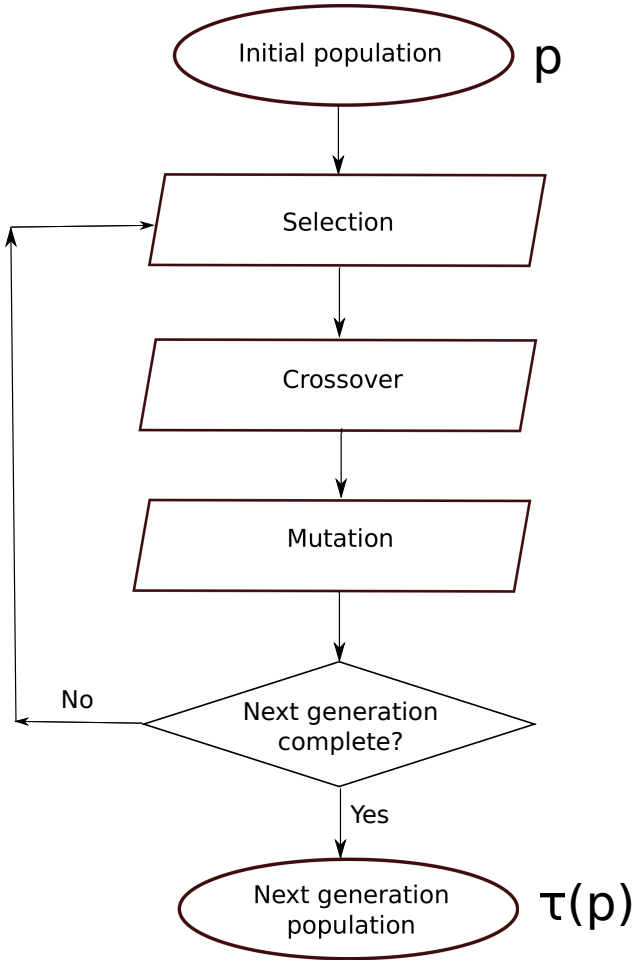


Figure 1.1: Finite GA (Haploid)

A simple Holland style genetic algorithm:

1. Start with some population P containing r binary strings of length ℓ
2. Choose parents u and v from the current population P (using any selection scheme with replacement)
    a. Crossover 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. Repeat step 2 until r offspring are created
4. Replace P by the new generation formed and go to step 2

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

Figure 1.2 illustrates algorithm for finite diploid population genetic algorithm. In case of diploid population GA each parent (u and v) has two haploid components ($\langle u_0, u_1\rangle$ and $\langle v_0, v_1\rangle$ respectively). Instead of crossing over two parent diploids, haploids in each diploid $\langle u_0, u_1\rangle$, and $\langle v_0, v_1\rangle$, crossover and mutate to produce gametes g_0 and g_1. And gametes g_0 and g_1 are fused to form offspring diploid $\langle g_0, g_1\rangle$.

1.2.2 Infinite Population Model

Haldane, in the classic book 'The Causes Of Evolution', presents a summary of the basic models of population genetics by Wright, Fisher, and Haldane (see Haldane (1932)). Holland introduced a population-based algorithm with crossover and mutation, and promoted his schema theorem as a theoretical means by which to analyze genetic algorithm dynamics (see Holland (1992)). 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

*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.

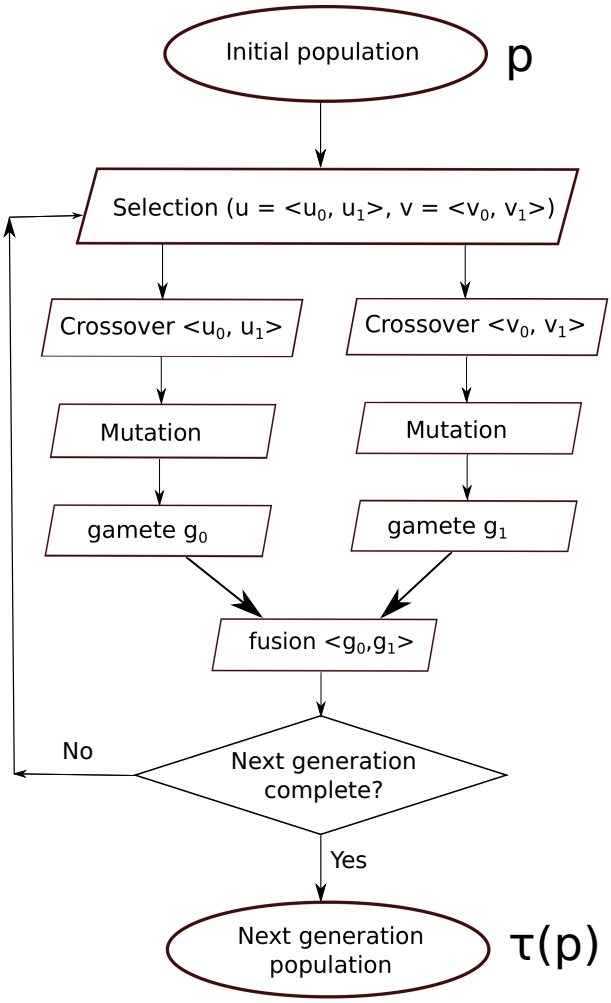


Figure 1.2: Finite GA (Diploid)

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 $\boldsymbol{p}$ (component $\boldsymbol{p}_i$ is the proportion of string i in the *finite population*), the infinite population model computes the expected proportion $\mathcal{G}(\boldsymbol{p})_i$ of string i in the next generation. This is perhaps the most direct connection between the infinite population model and a finite population GA. In the model, $\mathcal{G}$ comprises of fitness matrix F and recombination operator $\mathcal{M}$ that includes application of crossover and mutation.

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

$$\mathcal{G}(\boldsymbol{p})_j \;=\; \text{proportion of } j \text{ in the next generation.}$$

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

$$\boldsymbol{p} \to \mathcal{G}(\boldsymbol{p}) \to \mathcal{G}(\mathcal{G}(\boldsymbol{p})) \to \cdots$$

1.2.3 Finite Population Model

The infinite population model simplifies analysis of GA. However, finite populations can behave differently than infinite populations due to stochasticity involved with selection, crossover and mutaiton. Nix and Vose (see Nix and Vose (1992)) explored issues regarding the relationship between the finite population GA and the infinite population model. In particular, for a mutation rate μ between 0 and 0.5, a finite population GA will form an ergodic Markov chain, visiting every state infinitely often in the long run. Moreover, the short term trajectory followed by a finite population is related to the evolutionary path determined by the infinite population model, and for large populations, the short term trajectory follows closely and with large probability, that path predicted by the infinite population model.

Vose later generalized both infinite and finite population models as special case of a general abstract search framework called Random Heuristic Search (RHS) (described more in section 1.3).

1.2.4 Walsh Transform

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 calculations related to 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 analyze 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 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 $\mathcal{G}(\boldsymbol{p})$ 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.

[‡]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 $(M^*)_{i,j} = M_{i+j,i}$.

1.3 Random Heuristic Search

This section borrows from and summarizes – with permission – Vose (1999). 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 set of population descriptors is the *simplex* :

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

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

Given current population vector $\boldsymbol{p}$, the next population vector $\tau(\boldsymbol{p})$ cannot be predicted with certainty because τ is stochastic; it results from r independent, identically distributed random choices. Let $\mathcal{G} : \Lambda \to \Lambda$ be a function that maps current population vector $\boldsymbol{p}$ to a vector whose ith component is the probability that the ith element of Ω is chosen. Thus, $\mathcal{G}(\boldsymbol{p})$ specifies the distribution from which the aggregate of r choices forms the subsequent generation. The probability that population $\boldsymbol{q}$ is the next population vector given current population (vector) $\boldsymbol{p}$ is (see

Vose (1999))

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

where summation is restricted to indices for which $\boldsymbol{q}_j > 0$ and θ is a function such that $0 < \theta < 1$. Each random vector in the sequence $\boldsymbol{p}, \tau(\boldsymbol{p}), \tau^2(\boldsymbol{p}), ...$ 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 $\boldsymbol{p}$, produce $\boldsymbol{q}$ with probability $Q_{\boldsymbol{p},\boldsymbol{q}}$. The expected next generation $\mathcal{E}(\tau(\boldsymbol{p}))$ is $\mathcal{G}(\boldsymbol{p})$ (see Vose (1999)). The expression

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

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

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

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

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

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

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

indicates the probability that $\boldsymbol{q}$ is the next generation decays exponentially with increasing dispersion. As Vose stated in his book (see Vose (1999)):

> The combined effect of the two influences of discrepancy and dispersion is that random heuristic search favors a less disperse population near the expected next generation. In particular, if the current population is near the expected next generation, then the first factor does not contribute a strong bias for change. When $\mathcal{G}(\boldsymbol{p})$ is nearly the initial population $\boldsymbol{p}$, the influence of discrepancy favors $\boldsymbol{p}$ as the next generation since the alternatives, being lattice points, are constrained to be some distance away from the expected next generation. This phenomenon is expressed quantitivelty by theorem 3.4. Moreover, the second factor may exert a stabilizing effect provided the current population has low dispersion compared to the alternatives.

Figure 1.3 illustrates population points in a simplex for $\ell = 2, r = 4$. Finite

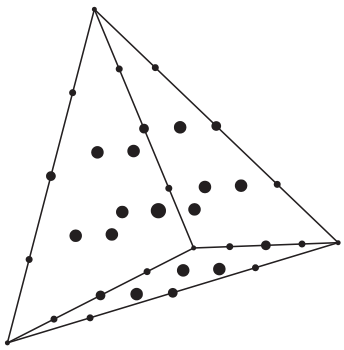


Figure 1.3: Population points

populations are represented by dots, where smaller dots have lower dispersion and are more likely points whereas larger dots have higher dispersion and are less likely

points. The diagram also illuminates that finite populations are constrained to occupy lattice points within Λ. As population size $r \to \infty$, 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

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

1.4 Research Problems

- Following Chebyshev's inequality (see Wikipedia (2016a)) equation 1.2 becomes

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

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

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

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

  and

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

  If

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

  then (1.3) becomes

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

  Therefore, $\tau(\boldsymbol{p})$ converges in probability to $\mathcal{G}(\boldsymbol{p})$ 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 1.3, 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 (see Vose (1999)):

> *If* $\boldsymbol{p}, \boldsymbol{q} \in \Lambda$ *are arbitrary population vectors for population size* r*, and* $\boldsymbol{\xi}$ *denotes an arbitrary element of* Λ*, then*
>
> $$\inf_{\boldsymbol{p}\neq\boldsymbol{q}} \|\boldsymbol{p}-\boldsymbol{q}\| = \sqrt{2}/r \tag{1.5}$$
> $$\sup_{\boldsymbol{\xi}} \inf_{\boldsymbol{p}} \|\boldsymbol{\xi}-\boldsymbol{p}\| = O(1/\sqrt{r}) \tag{1.6}$$
>
> *where the constant (in the "big oh") is independent of the dimension* n *of* Λ.

From 1.6, the distance between an infinite population $\boldsymbol{\xi}$ and finite population p is $O(1/\sqrt{r})$. This suggests that the distance between $\tau(\boldsymbol{p})$ and $\mathcal{G}(\boldsymbol{p})$ might decrease as $1/\sqrt{r}$.

Let η be the random variable $\|\tau(\boldsymbol{p}) - \mathcal{G}(\boldsymbol{p})\|$, 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,

$$\mathcal{E}(\|\tau(\boldsymbol{p}) - \mathcal{G}(\boldsymbol{p})\|) = \mathcal{E}(\eta) \leq \sqrt{\mathcal{E}(\eta^2)} = \frac{\sqrt{1 - \|\mathcal{G}(\boldsymbol{p})\|^2}}{\sqrt{\boldsymbol{r}}} \tag{1.7}$$

This suggests that the distance between $\tau(\boldsymbol{p})$ and $\mathcal{G}(\boldsymbol{p})$ might decrease as $1/\sqrt{r}$.

Equations 1.4, 1.6, and 1.7 all suggest that the distance between $\tau(\boldsymbol{p})$ and $\mathcal{G}(\boldsymbol{p})$ might decrease as $1/\sqrt{r}$. All three of them are inequalities. The distance may

decrease much faster than as $1/\sqrt{r}$ in reality. The first research question to consider is whether that rate of decrease can be exhibited in practice. We investigate the rate of decrease with experiments in Chapter 2.

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

$$\boldsymbol{p}, \mathcal{G}(\boldsymbol{p}), \mathcal{G}^2(\boldsymbol{p}), ...$$

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 fixed point ω

$$\mathcal{G}(\omega) = \lim_{n \to \infty} \mathcal{G}^n(\boldsymbol{p}) = \omega.$$

When specialized to a simple GA (the details are explained in Chapter 2), it turns out that $\mathcal{G}$ is focused under certain conditions, but under other conditions the sequence $\boldsymbol{p}, \mathcal{G}(\boldsymbol{p}), \mathcal{G}^2(\boldsymbol{p}), ...$ 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 finite populations might also show oscillating behavior. Akin analytically proves the existence of cycling for a continuous-time diploid two loci, two allele model (see Akin (1982)). In contrast, we consider a discrete-time model with more than two loci. Hastings used a numerical approach to study the behavior of cycling populations with the infinite diploid population model (see Hastings (1981)). His model includes crossover but not mutation. Moreover, the study was limited to two loci and two alleles. In contrast, we consider more than two loci, and include mutation. Wright and Bidwell provided examples when cycles in an infinite population model occur with mutation and crossover for 3 and 4 bit populations (see Wright and Bidwell (1997)). Different behavior cases were observed. For a 3 bit example, both approximate period 2 cycling and long period cycling were observed. For a 4 bit example, long

period cycling was observed. The examples provided were for specific parameter values. Their examples are based on computing a specific fitness function and a specific initial population from randomly generated mutation and crossover distributions in an attemp to find cyclic behavior anywhere within the parameter space of fitness, crossover and mutation. In contrast, we investigate cyclic behavior within a slice of the parameter space corresponding to fixed fitness, and consider randomly generated initial populations, as well as randomly generated crossover and mutation distributions.

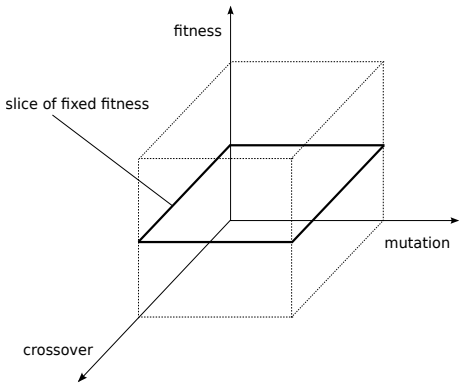


Figure 1.4: Parameter space of crossover, mutation and fitness

Similar in some respects to our approach, Wright and Agapie describe cycling behavior in one slice of the parameter space of fitness, crossover and mutation (see Wright and Agapie (2001)). They fix the fitness function to be one plus the integer value (of the population member). They showed results for 1-bit to 4-bit infinite population evolutions, and observed cyclic behavior of periods 2, 3, 4, 8 and 10. They also present data for finite populations exhibiting cyclic behavior. A significant difference between their investigation and ours is that their mutation is dynamic; the manner in which a population member mutates is dependent upon where the population is located in the state space Λ. In contrast, we consider static mutation which mutates population members uniformly irrespective of where the population is located in Λ. Moreover, works

of Wright and Bidwell, and Wright and Agapie focus only on haploid population evolution whereas we consider both haploid and diploid population evolution. Our research also studies cyclic behavior for a different slice of fitness than Wright and Agapie used in their work. We consider uniform fitness where every population member has the same fitness. We find cycles in both finite and infinite population evolution, and provide visualization of oscillation related to infinite population fixed points. In Chapter 3, we investigate the second research question: Do finite populations exhibit oscillation in practice when infinite population oscillates?

- The third research question concerns the robustness of finite population oscillation. Consider the lattice points in the simplex Λ which represent finite populations (for some fixed population size r) and let $\boldsymbol{P}_j$ denote the jth population represented by the jth lattice point. Let $\boldsymbol{\pi}^k$ be the probability vector having as jth component the probability that $\boldsymbol{P}_j$ is the kth generation. If $\boldsymbol{\pi}^0$ is the initial population distribution, the steady state distribution $\boldsymbol{\pi}$ is given by (see Häggström (2002))

$$\boldsymbol{\pi} \;=\; \lim_{k\to\infty} \boldsymbol{\pi}^k \;=\; \lim_{k\to\infty} \boldsymbol{\pi}^0 Q^k \tag{1.8}$$

assuming the limit exists. The jth component $\boldsymbol{\pi}_j$ can be interpreted as the proporiton of time that a GA spends in population $\boldsymbol{P}_j$. If transition matrix Q is irreducible[§] and aperiodic[¶], then the Markov chain is regular (see Iosifescu (1980)), the steady state distribution $\boldsymbol{\pi}$ exists, and it has positive components (see Minc (1988)). The solution to equation 1.8 satisfies

$$\boldsymbol{\pi} \;=\; \boldsymbol{\pi} Q \tag{1.9}$$

[§]A Markov chain is said to be *irreducible* if it is possible to get to any state from any state.
[¶]A Markov chain is *aperiodic* if it can return to state i at irregular times.

where $\boldsymbol{\pi}$ is normalized so that its components sum to one. If GA were to perfectly oscillate between two populations $\boldsymbol{P}_i$ and $\boldsymbol{P}_j$, then $\boldsymbol{\pi}_i^k = 1$ (other components are 0) when k is odd, and $\boldsymbol{\pi}_j^k = 1$ (other components are 0) when k is even. Therefore, perfect oscillation should not occur. In Chapter 4, we investigate oscillation behavior of finite populations when the Markov chain is regular. The third research question concerns whether finite population approximate oscillation can be exhibited in practice when the Markov chain is regular and infinite population trajectories have no periodic orbit.

- In their work on cyclic behavior of populations, Wright and Agapie point out that the presence or absence of crossover did not affect cyclic behavior (see Wright and Agapie (2001)). But in our work, the condition for infinite population evolution to converge to periodic orbits depends upon crossover and if the crossover distribution condition is violated, infinite populations will not have periodic orbits (see Vose (1999)). We investigate the robustness of finite population oscillation. The fourth research question is: Can finite population approximate oscillation be exhibited in practice when infinite population trajectories have no periodic orbit due to the crossover distribution violating the condition required for infinite population oscillation?

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 the next generation is obtained from the 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 simulations to study the convergence of finite population short-term behavior to behavior predicted by the infinite population model. The results confirm that the distance between the short-term evolutionary trajectory of finite diploid populations and the evolutionary trajectory of infinite diploid populations can in practice decrease like the inverse of the square root of population size. Our first research question is thereby answered affirmatively.

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 $\ell_1, \ldots, \ell_h$; thus $\ell = \ell_1 + \cdots + \ell_h$).

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^n_α 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^n_α — obtain gamete γ_0
- From parent β — selected with probability q^n_β — obtain gamete γ_1

[‡]The representation here is the conceptual equivalent of Hardy's model.

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

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

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

$$q_{\langle\gamma_0,\gamma_1\rangle}^{n+1} \;=\; p_{\gamma_0}^{n+1} \, p_{\gamma_1}^{n+1} \tag{2.2}$$

[§]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,

$$p^{n+1}_{\gamma_0} \;=\; \sum_{\alpha} q^n_\alpha \, t_\alpha(\gamma_0) \tag{2.3}$$

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

$$\begin{aligned} p^{n+1}_{\gamma_0} \;&=\; \sum_{\alpha_0,\,\alpha_1} q^n_{\langle \alpha_0,\,\alpha_1\rangle} \, t_{\langle \alpha_0,\,\alpha_1\rangle}(\gamma_0) \\ &=\; \sum_{\alpha_0,\,\alpha_1} p^n_{\alpha_0} \, p^n_{\alpha_1} \, t_{\langle \alpha_0,\,\alpha_1\rangle}(\gamma_0) \end{aligned} \tag{2.4}$$

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

$$1 \;=\; \sum_{\alpha} q^n_\alpha \;=\; \sum_{\alpha_0,\,\alpha_1} p^n_{\alpha_0} \, p^n_{\alpha_1} \;=\; \Big(\sum_{\alpha_0} p^n_{\alpha_0} \Big)^2$$

The weighted count of haploid g in generation n is

$$\sum_{\alpha_0,\,\alpha_1} q^n_{\langle \alpha_0,\alpha_1\rangle} ([g=\alpha_0] + [g=\alpha_1]) \tag{2.5}$$

$$= \sum_{\alpha_0,\,\alpha_1} p^n_{\alpha_0} \, p^n_{\alpha_1} [g=\alpha_0] + \sum_{\alpha_0,\,\alpha_1} p^n_{\alpha_0} \, p^n_{\alpha_1} [g=\alpha_1] \tag{2.6}$$

$$= 2p^n_g \tag{2.7}$$

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

$$\psi : \boldsymbol{q}^n \;\longmapsto\; \boldsymbol{p}^n$$

Evolution equation (2.4) in matrix form is

$$p'_g \;=\; p^T M_g \, p \tag{2.8}$$

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

$$\Big(M_g\Big)_{u,v} \;=\; t_{\langle u,v\rangle}(g) \tag{2.9}$$

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

2.3 Specialization

This section borrows from and summarizes (with permission) the development in 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 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 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, $\mu \in [0, 0.5)$ is regarded as a *mutation rate* which implicitly specifies distribution $\boldsymbol{\mu}$ according to the rule (see Vose and Wright

(1998))

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

2.3.2 Crossover

Crossover refers to crossing over (also termed recombination) between two chromosomes (strings in our case). Crossover like mutation also provides a mechanism for injection of new strings into the next generation population. Geiringer (see Geiringer (1944)) used crossover masks to implement recombination. 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, χ can be considered as a *crossover rate* that specifies the distribution $\boldsymbol{\chi}$ given by the rule (see Vose and Wright (1998))

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

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 $\mathcal{R}$

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

Crossover and mutation are defined in a manner respecting arbitrary partioning 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 disscussed in above sections. For mutation distribution $\boldsymbol{\mu}$ and crossover distribution $\boldsymbol{\chi}$, the transmission function can be expressed as (see Vose and Wright (1998))

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

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

$$p'_g \;=\; (\sigma_g p)^T M \, (\sigma_g p) \tag{2.11}$$

where the permutation matrix σ_g is defined by component equations

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

2.4 Walsh Transorm

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

$$W_{n,t} = N^{-1/2}(-1)^{n^T t} \tag{2.12}$$

where $N^{-1/2}$ can be thought of as a normalization factor. The matrix is symmetric, i.e.,

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

and it has entries satisfying

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

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

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

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

2.4.1 Fast Walsh Transform

Computation of the Walsh transform given by equation (2.12) might take n^2 operations if implemented naively. An algorithm using $O(n \log_2 n)$ operations 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 FWT algorithm can be translated into pseudocode as:

```
1:  procedure FWT
2:      n = 2^d ← size of array X where d is positive integer
3:      for i = 1 to d  do
4:          m = 2^i
5:          z = m/2
6:          for k = 0 to z − 1 do
7:              for j = 0 to n − 1 step m do
8:                  t1 = j + k
9:                  t2 = t1 + z
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
```

Algorithm 1: FWT pseudocode

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, making feasible computation-based 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 simplified by changing basis to diagonalize the σ_g. Columns of the Walsh matrix 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

$$\widehat{M}_{u,v} \;=\; 2^{\ell-1}\,[uv=\mathbf{0}]\,\widehat{\boldsymbol{\mu}}_u\widehat{\boldsymbol{\mu}}_v \sum_{k\in\overline{u+v}\mathcal{R}} \boldsymbol{\chi}_{k+u}+\boldsymbol{\chi}_{k+v} \tag{2.13}$$

and equation (2.11) takes the form

$$\widehat{p}'_g \;=\; 2^{\ell/2} \sum_{i \in g\mathcal{R}} \widehat{p}_i \, \widehat{p}_{i+g} \, \widehat{M}_{i,i+g} \tag{2.14}$$

where $g\mathcal{R} = \{gi \,|\, i \in \mathcal{R}\}$ (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 (2.14) in terms of the mixing matrix (2.13), is Markovian; the next state p' depends only upon the current state p. Let $\mathcal{M}$ represent the mixing transformation,

$$p' \;=\; \mathcal{M}(p) \tag{2.15}$$

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 $p^1 \;=\; \psi(q^1)$. 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; (2.15) can be computed in Walsh coordinates relatively efficiently via (2.13) and (2.14).

2.5 Distance

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

$$S_{\boldsymbol{f}} \;=\; \{\alpha \,|\, \boldsymbol{f}_\alpha > 0\}$$

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

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

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

$$\|\boldsymbol{f} - \boldsymbol{q}\|^2 \;=\; \sum_{\alpha \notin S_{\boldsymbol{f}}} (\boldsymbol{f}_\alpha - \boldsymbol{q}_\alpha)^2 + \sum_{\alpha \in S_{\boldsymbol{f}}} (\boldsymbol{f}_\alpha - \boldsymbol{q}_\alpha)^2 \tag{2.16}$$

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

$$\begin{aligned} \sum_{\langle \alpha_0, \alpha_1 \rangle \notin S_{\boldsymbol{f}}} (\boldsymbol{p}_{\alpha_0} \boldsymbol{p}_{\alpha_1})^2 \;&=\; \sum_{\langle \alpha_0, \alpha_1 \rangle} (\boldsymbol{p}_{\alpha_0})^2 (\boldsymbol{p}_{\alpha_1})^2 \;-\; \sum_{\langle \alpha_0, \alpha_1 \rangle \in S_{\boldsymbol{f}}} \left(\boldsymbol{p}_{\alpha_0} \boldsymbol{p}_{\alpha_1}\right)^2 \\ &=\; \sum_{g}^{2} (\boldsymbol{p}_g)^2 \;-\; \sum_{\alpha \in S_{\boldsymbol{f}}} (\boldsymbol{q}_\alpha)^2 \end{aligned} \tag{2.17}$$

It follows from (2.16) and (2.17) that

$$\begin{aligned} \|\boldsymbol{f} - \boldsymbol{q}\|^2 \;&=\; \sum_{g}^{2} (\boldsymbol{p}_g)^2 \;+\; \sum_{\alpha \in S_{\boldsymbol{f}}} (\boldsymbol{f}_\alpha - \boldsymbol{q}_\alpha)^2 - \sum_{\alpha \in S_{\boldsymbol{f}}} (\boldsymbol{q}_\alpha)^2 \\ &=\; \sum_{g}^{2} (\boldsymbol{p}_g)^2 \;+\; \sum_{\alpha \in S_{\boldsymbol{f}}} \boldsymbol{f}_\alpha (\boldsymbol{f}_\alpha - 2\boldsymbol{q}_\alpha) \end{aligned} \tag{2.18}$$

which involves $2^\ell + |S_{\boldsymbol{f}}|$ terms, assuming that $S_{\boldsymbol{f}}$ is known as a byproduct of computing $\boldsymbol{f}$. Therefore, (2.18) computes distance between finite and infinite population efficiently.

2.6 Simplification

Computations in the haploid case are simplified by equations (2.13) and (2.14) 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 $n \log n$ (see Shanks (1969)).

Only one mixing matrix as opposed to 2^ℓ matrices is needed to compute the next generation; evolution equation (2.14) 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 (2.13) with equation (2.10). Also, the relevant quadratic form is computed with a single sum as opposed to a double sum; computing via (2.14) 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 behavior 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), (2.13), (2.14), (2.18) were employed to efficiently compute the distance

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

where $\boldsymbol{f}^n$ and $\boldsymbol{q}^n$ 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 ($\boldsymbol{f}^0 = \boldsymbol{q}^0$). Genome lengths $\ell \in \{4, 6, 8, 10, 12, 14\}$ and population sizes $N = 2^i$ for integer $0 \le i \le 20$ were considered. The crossover distribution $\boldsymbol{\chi}$ corresponds to independent assortment of bits, and the mutation distribution $\boldsymbol{\mu}$ corresponds to independent bit mutation probability 0.001,

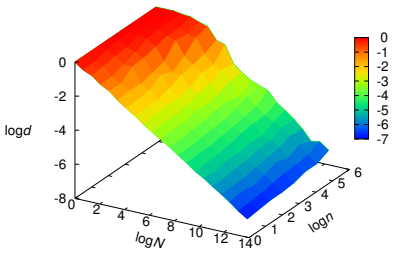
$$\boldsymbol{\chi}_m \;=\; 2^{-\ell}, \qquad \boldsymbol{\mu}_g \;=\; (0.001)^{\mathbf{1}^{\mathrm{T}} g}(0.999)^{\ell - \mathbf{1}^{\mathrm{T}} g}$$

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

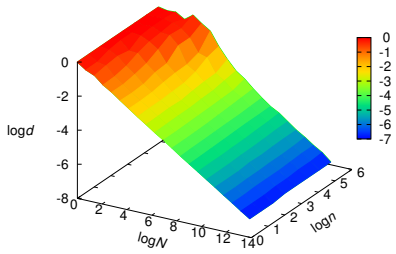
The data, presented in six surface graphs in figure 2.1 and organized by genome length, shows a near linear dependence of $\log d$ on $\log N$. As expected, the graphs show smoothing with increasing genome length (the computation of d involves averaging over ℓ components), and also with increased population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

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

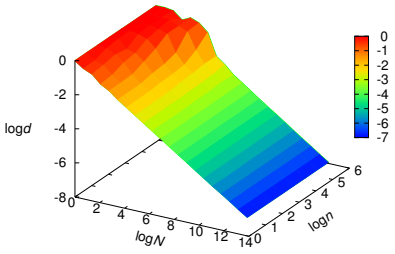
$$\log d = m \log N + b \tag{2.19}$$



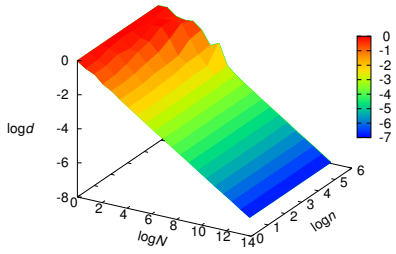
(a) $\ell = 4$.



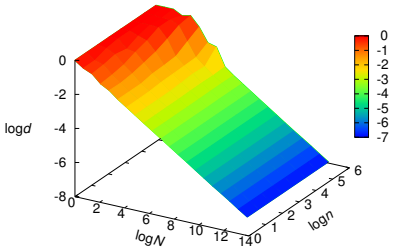
(b) $\ell = 6$.



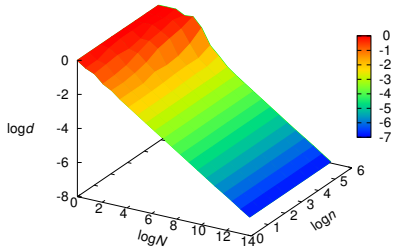
(c) $\ell = 8$.



(d) $\ell = 10$.



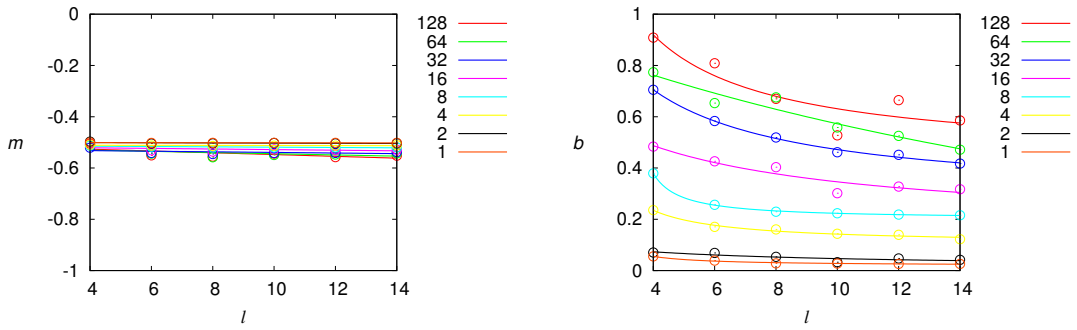
(e) $\ell = 12$.



(f) $\ell = 14$.

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

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



(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 $n \in \{1, 2, 4, 8, 16, 32, 64, 128\}$

Taking the exponential of the regression line (2.19) yields the estimate $d \approx N^m e^b$. Slopes of the regression lines shown in figure 2.2 are approximately -0.5, indicating

$$d \approx k/\sqrt{N}. \tag{2.20}$$

Equation 2.20 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 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 2.2b 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 partly from genetic drift experienced by

finite populations (see Crow and Kimura (1970)), and partly because as generations increase, differences between actual finite and infinite populations may accumulate which can be understood from figure 2.3.

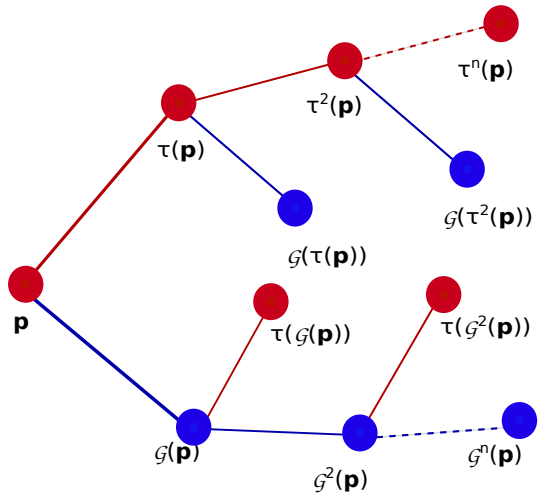


Figure 2.3: Non linearity in distance as generation increases: Red nodes represent finite populations, blue node represents infinite populations. Edges connect one generation to the next (red for finite population, blue for infinite population)

In figure 2.3, distance between the two immediate children of any node is aproximately $1/\sqrt{N}$. But the distance between descendents k generations later may accumulate to be like $k/\sqrt{N}$.

2.8 Summary

We began with a description of a simple diploid Markov model under mutation and crossover with no selective pressure. The model was reduced to the haploid case and specialized using mask-based recombination operators to extend Vose's infinite population model to the diploid case. Using computational benefits of this reduction, we showed via experiment and regression of the resulting data that distance between finite diploid population and infinite diploid population can indeed decrease like

$1/\sqrt{N}$ in practice. That rate of decrease is consistent with the single-step convergence bounds predicted by Vose's infinite population model for the haploid case.

Chapter 3

Oscillation

This chapter investigates the qualitative similarity between finite population short-term behavior and infinite population evolutionary limits predicted by Vose. It uses computation to verify predicted infinite population limits and presents necessary and sufficient conditions for convergence to periodic orbits. We compute mutation distribution $\boldsymbol{\mu}$ and crossover distribution $\boldsymbol{\chi}$ to satisfy those conditions. Through experiments, we explore our second research question: can finite populations exhibit oscillation behavior in practice?

3.1 Limits

Vose states that under mild assumptions on mutation and crossover (explained later), infinite populations converge under repeated application of $\mathcal{M}$ in the absense of selective pressure. Vose mentions that periodic orbits are possible, but populations converge under repeated application of $\mathcal{M}^2$ and the limits $\boldsymbol{p}^* = \lim_{n\to\infty} \mathcal{M}^{2n}(\boldsymbol{p})$ and $\boldsymbol{q}^* = \lim_{n\to\infty} \mathcal{M}^{2n+1}(\boldsymbol{q})$ exist (see Vose (1999)).

Following Vose (see Vose (1999)), let $S_g = g\mathcal{R} \setminus \{\mathbf{0}, g\}$, and let $|g|$ be the number of non zero bits in g. If $\widehat{\boldsymbol{p}}$ represents the current population in Walsh coordinates,

then the next generation $\widehat{\boldsymbol{p}}'_g$ (expressed in Walsh coordinates) is

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

where

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

Moreover,

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

With above notations, the limits can be expressed in the Walsh basis by recursive equations (see Vose (1999))

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

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

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

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

$$\boldsymbol{p}, \mathcal{M}(\boldsymbol{p}), \mathcal{M}^2(\boldsymbol{p}), \cdots$$

to converge to a periodic orbit is that for some g, $g \neq 0$

$$-1 = \sum_{j} (-1)^{g^T j} \boldsymbol{\mu}_j = -\sum_{k \in \bar{g}\mathcal{R}} \boldsymbol{\chi}_{k+g} + \boldsymbol{\chi}_k \tag{3.3}$$

3.2 Mutation and Crossover Distributions

The following describes the generation of mutation and crossover distributions that satisfy equation 3.3 for evolution to converge to a periodic orbit. Let $\boldsymbol{\mu}$ and $\boldsymbol{\chi}$ represent mutation and crossover distributions (respectively), and let $U01()$ return a random number between 0 and 1. For some $g \in \mathcal{R}$, $g \neq 0$, and for all $j \in \mathcal{R}$,

$$\boldsymbol{\mu}_j = \begin{cases} U01() & \text{if } g^T j \text{ is odd} \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

Normalization yields $\boldsymbol{\mu}$ (the mutation distribution),

$$\boldsymbol{\mu}_j := \boldsymbol{\mu}_j / \sum_{j \in \mathcal{R}} \boldsymbol{\mu}_j.$$

Moreover, $\boldsymbol{\mu}$ satisfies condition 3.3.

Condition $k \in \bar{g}\mathcal{R}$ in equation 3.3 is

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

Multiplying through by $\bar{g}$ yields

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

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

$$\begin{aligned} \boldsymbol{\chi}_k &= U01() \text{ if } \bar{g}k = k \\ \boldsymbol{\chi}_{k+g} &= U01() \text{ if } \bar{g}k = k \\ \boldsymbol{\chi}_k &= 0 \ otherwise. \end{aligned} \tag{3.5}$$

Normalization yields $\boldsymbol{\chi}$ (the crossover distribution),

$$\boldsymbol{\chi}_k := \boldsymbol{\chi}_k / \sum_{k \in \mathcal{R}} \boldsymbol{\chi}_k.$$

Moreover, $\boldsymbol{\chi}_k$ satisfies condition 3.3.

3.3 Initial Population

To investigate oscillation in infinite population and finite population behavior, it is desirable to have the same or corresponding initial populations.

For string length ℓ, the number of possible haploids is $x = 2^\ell$. Let array $\boldsymbol{t}$ represent a population of size N as follows: $\boldsymbol{t}_j$ is the jth population member (some element of $\{0, .., x-1\}$ where elements are base 2 length ℓ binary strings). Array $\boldsymbol{t}$ is generated from a random vector $\boldsymbol{u}$ of size x as follows.

$$\begin{aligned} \boldsymbol{u}_i &= U01(); \qquad i = 0, .., x-1 \\ \boldsymbol{t}_j &= randp(\boldsymbol{u}); \qquad j = 0, .., N-1 \end{aligned}$$

where $randp(\boldsymbol{u})$ returns random index i into array $\boldsymbol{u}$ with probability $\boldsymbol{u}_i$.

Let $\boldsymbol{c}_i$ be the count of haploid member i in population $\boldsymbol{t}$,

$$\boldsymbol{c}_i = \sum_{j=0}^{N-1} [\boldsymbol{t}_j = i]; \qquad i = 0, .., x-1$$

The infinite population vector $\boldsymbol{p}$ has ith component

$$\boldsymbol{p}_i = \frac{\boldsymbol{c}_i}{N}.$$

This randomly generated infinite haploid population vector $\boldsymbol{p}$ is used to obtain a diploid infinite population vector $\boldsymbol{q}$, and finite population vectors $\boldsymbol{s}$ and $\boldsymbol{f}$ as follows.

Infinite diploid population $\boldsymbol{q}$ is calculated corresponding to initial haploid population $\boldsymbol{p}$ as

$$\boldsymbol{q}_{i,j} = \boldsymbol{p}_i \boldsymbol{p}_j; \qquad (0 \leq i, j < x)$$

The finite haploid population members are the elements of array $\boldsymbol{t}$, the corresponding finite haploid population vector $\boldsymbol{s}$ is identical to $\boldsymbol{p}$. Let $\boldsymbol{v}$ be a finite diploid population member array of dimension two and of size N^2 whose diploid member $\boldsymbol{v}[i][j]$ at index $[i][j]$ is

$$\boldsymbol{v}[i][j] \;=\; \langle \boldsymbol{t}_i, \boldsymbol{t}_j \rangle \qquad 0 \leq i, j < N$$

The finite diploid population (proportion) vector $\boldsymbol{f}$ corresponding to finite diploid population member array $\boldsymbol{v}$ is identical to $\boldsymbol{q}$.

Thus, initial infinite haploid population vector $\boldsymbol{p}$ corresponds to initial infinite diploid population vector $\boldsymbol{q}$, and to initial finite haploid population vector $\boldsymbol{s}$ with population size N and population member array $\boldsymbol{t}$, and to initial finite diploid population vector $\boldsymbol{f}$ with population size N^2 and population member array $\boldsymbol{v}$.

3.4 Oscillation

Crossover distributions $\boldsymbol{\chi}$ and mutation distributions $\boldsymbol{\mu}$ satisfying condition (3.3) are considered to investigate oscillating behavior in terms of predicted infinite population evolutionary limits.

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

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

where $\gamma = \langle\gamma_0, \gamma_1\rangle$ is a 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 conisdered. 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 $N \in \{N_0^2, 10N_0^2, 20N_0^2\}$. To study oscillation in finite populations, the distances of $\boldsymbol{p}^n$ and $\boldsymbol{s}^n$ to haploid evolutionary limits $\boldsymbol{p}_h^*$ and $\boldsymbol{q}_h^*$ were plotted and the distances of $\boldsymbol{q}^n$ and $\boldsymbol{f}^n$ to diploid evolutionary limits $\boldsymbol{p}_d^*$ and $\boldsymbol{q}_d^*$ were plotted.

According to the results and conclusions from chapter 2, the expected distance d between finite population of size N and infinite population is

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

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

N	4096	40960	81920
d	0.0156	0.0049	0.0035

The distance between finite population and infinite population, for both haploid and diploid cases, were also plotted.

3.4.1 Haploid Population

Figures 3.1, 3.2, 3.3 and 3.4 show oscillations in finite haploid populations, and distances between finite and infinite haploid populations arranged by genome length ℓ. In each figure, sub-figures are arranged by population size N. The first three rows of sub-figures in the left column show distance d' of finite population to limits, the sub-figure in fourth row of the left column shows distance d' of infinite population to limits. These sub-figures depict oscillating behavior of both infinite and finite haploid populations when condition 3.3 is met. As population size increases, oscillation approaches the behavior exhibited by infinite population.

In each figure (3.1, 3.2, 3.3, and 3.4), the first three graphs in the right column show distance variation (difference of distance d and average distance d_{avg}) where d is distance between haploid finite and infinite populations and d_{avg} is average value of d. The graph in the fourth row shows distance between finite and infinite populations decreases as population size increases, consistent with results from section 2.1. The graphs of $d - d_{avg}$ decrease in amplitude as population size increases. As ℓ increases, the distance graphs become smoother, and amplitude of oscillations decrease.

Distance data obtained from simulations for haploid populations are summarized in table 3.2, which tabulates average distance between finite and infinite populations.

Table 3.2: Distance measured for haploid population: N is population size, ℓ is genome length, average distance between finite and infinite population is tabulated in the last three columns, and last row is expected single step distance.

ℓ	$N = 4096$	$N = 40960$	$N = 81920$
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
$1/\sqrt{N}$	0.0156	0.0049	0.0035

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

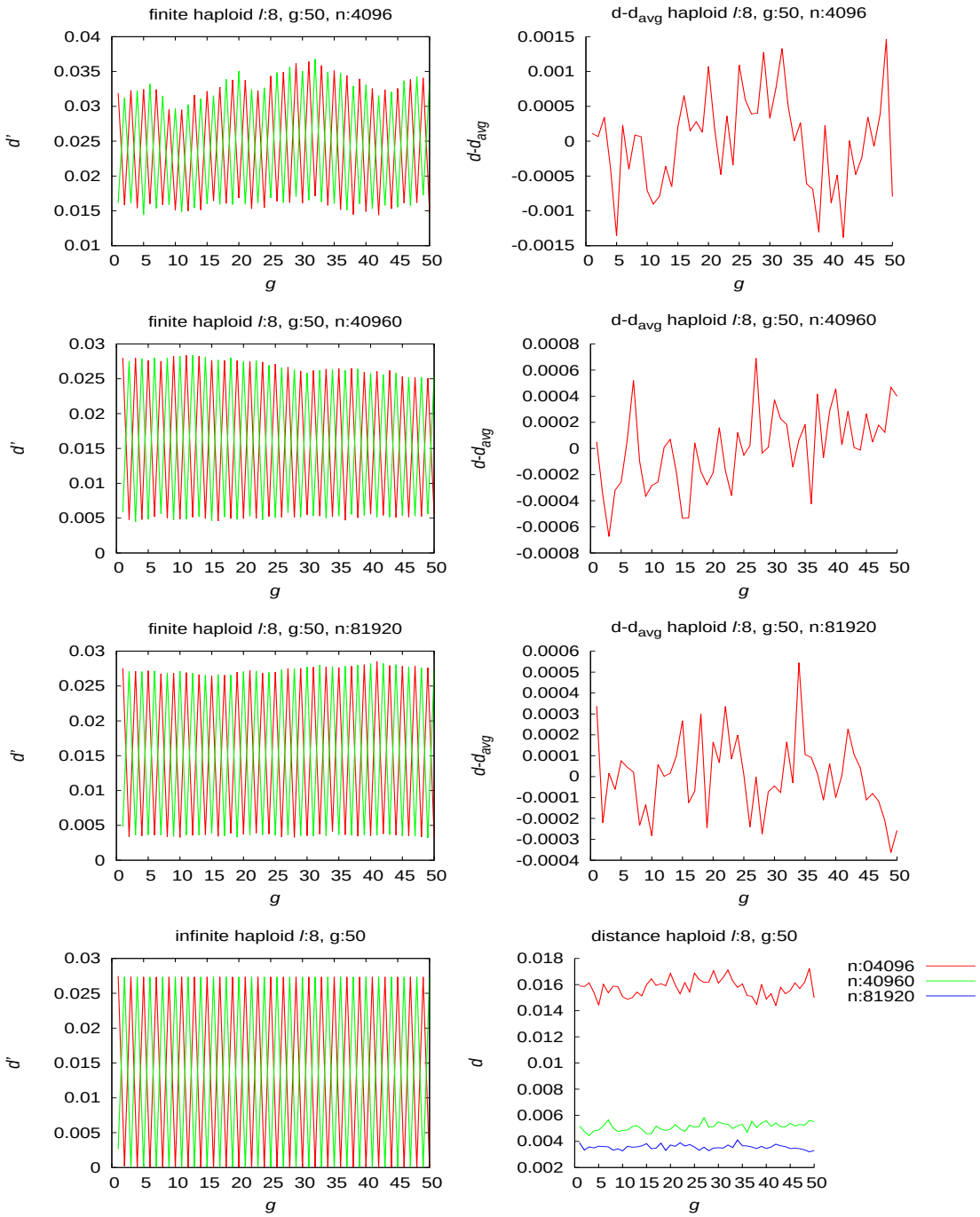


Figure 3.1: Infinite and finite haploid population behavior for genome length $\ell = 8$: In left column, d' is distance of finite population of size n or infinite population to limits for g generations. Green line is distance to $\boldsymbol{p}*$ and red line is distance to $\boldsymbol{q}*$. In right column, d is distance of finite population to infinite population for g generations and d_{avg} is average distance.

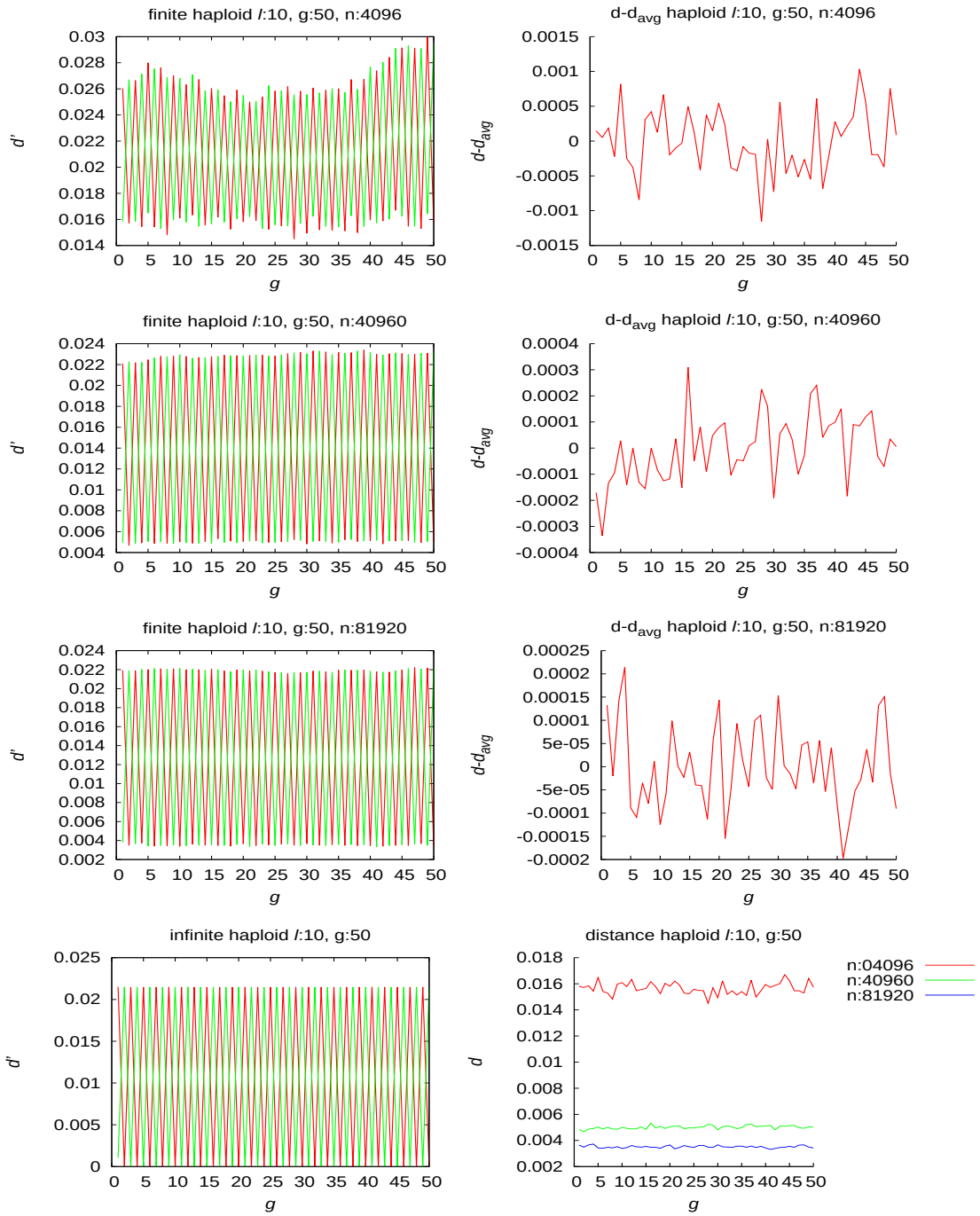


Figure 3.2: Infinite and finite haploid population oscillation behavior for genome length $\ell = 10$: In left column, d' is distance of finite population of size n or infinite population to limits for g generations. Green line is distance to $\boldsymbol{p}*$ and red line is distance to $\boldsymbol{q}*$. In right column, d is distance of finite population to infinite population for g generations and d_{avg} is average distance.

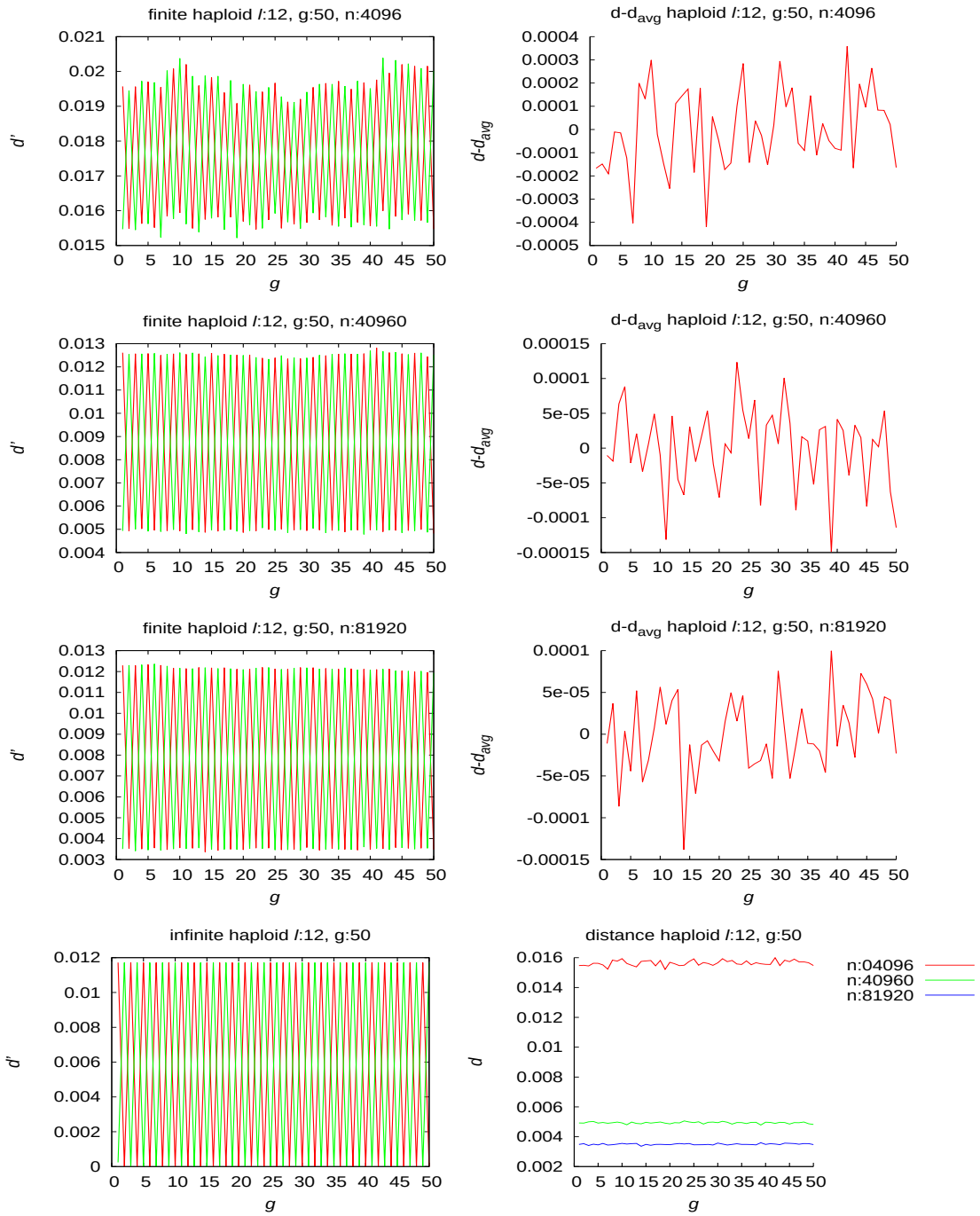


Figure 3.3: Infinite and finite haploid population oscillation behavior for genome length $\ell = 12$: In left column, d' is distance of finite population of size n or infinite population to limits for g generations. Green line is distance to $\boldsymbol{p}*$ and red line is distance to $\boldsymbol{q}*$. In right column, d is distance of finite population to infinite population for g generations and d_{avg} is average distance.

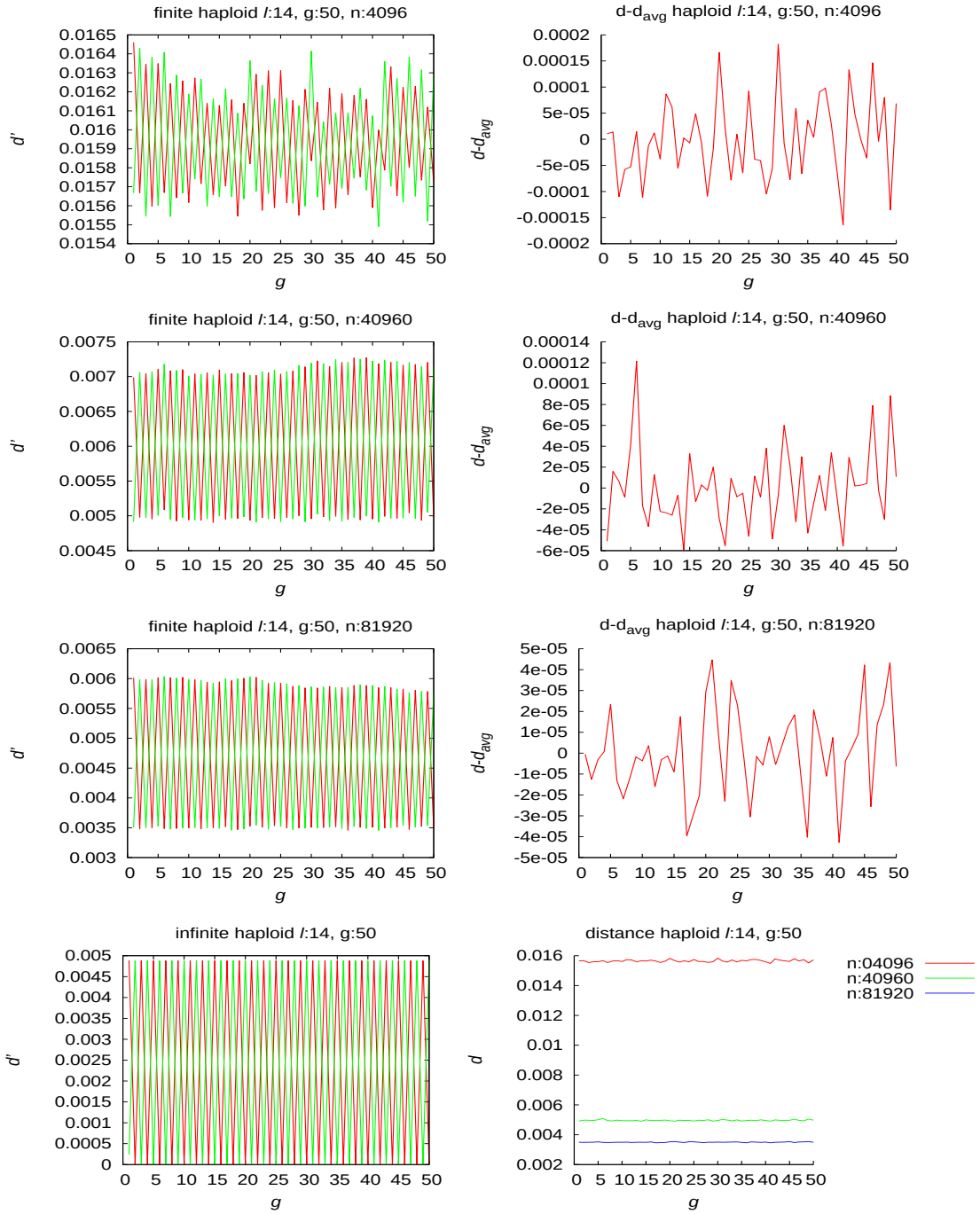


Figure 3.4: Infinite and finite haploid population oscillation behavior for genome length $\ell = 14$: In left column, d' is distance of finite population of size n or infinite population to limits for g generations. Green line is distance to $\boldsymbol{p}*$ and red line is distance to $\boldsymbol{q}*$. In right column, d is distance of finite population to infinite population for g generations and d_{avg} is average distance.

3.4.2 Diploid Population

Figures 3.5, 3.6, 3.7 and 3.8 show oscillations in finite diploid populations, and distances between finite diploid populations and infinite diploid populations arranged by genome length ℓ in ascending order. In each figure for unique genome length ℓ, sub-figures are arranged by population size N. In each figure, the first three rows of sub-figures in the left column show distance d' of finite population to limits, the sub-figure in fourth row of the left column shows distance d' of infinite population to limits. These sub-figures depict oscillating behavior of both infinite and finite diploid populations when condition 3.3 is met. Like in haploid population case, as population size increases, oscillation approaches the behavior exhibited by infinite population.

In each figure (3.5, 3.6, 3.7, and 3.8), the first three graphs in the right column show distance variation (difference of distance d and average distance d_{avg}) where d is distance between diploid finite and infinite populations and d_{avg} is average value of d. The graph in the fourth row of the right column combines distance plots between finite and infinite populations for sizes ($N = N_0^2, 10N_0^2, 20N_0^2$). The graphs show distance decreases as population size increases, consistent with results from section 2.1. The graphs of $d - d_{avg}$ decrease in amplitude as population size increases. For fixed finite population size, as ℓ increases, the distance graphs become smoother, and amplitude of oscillations decrease.

Distance data obtained from simulations for diploid populations are summarized in table 3.3, which tabulates average distance between finite and infinite populations.

Results from table 3.3 show average distance between finite and infinite population closely follows the expected single step distance. The distance decreases as $1/\sqrt{N}$.

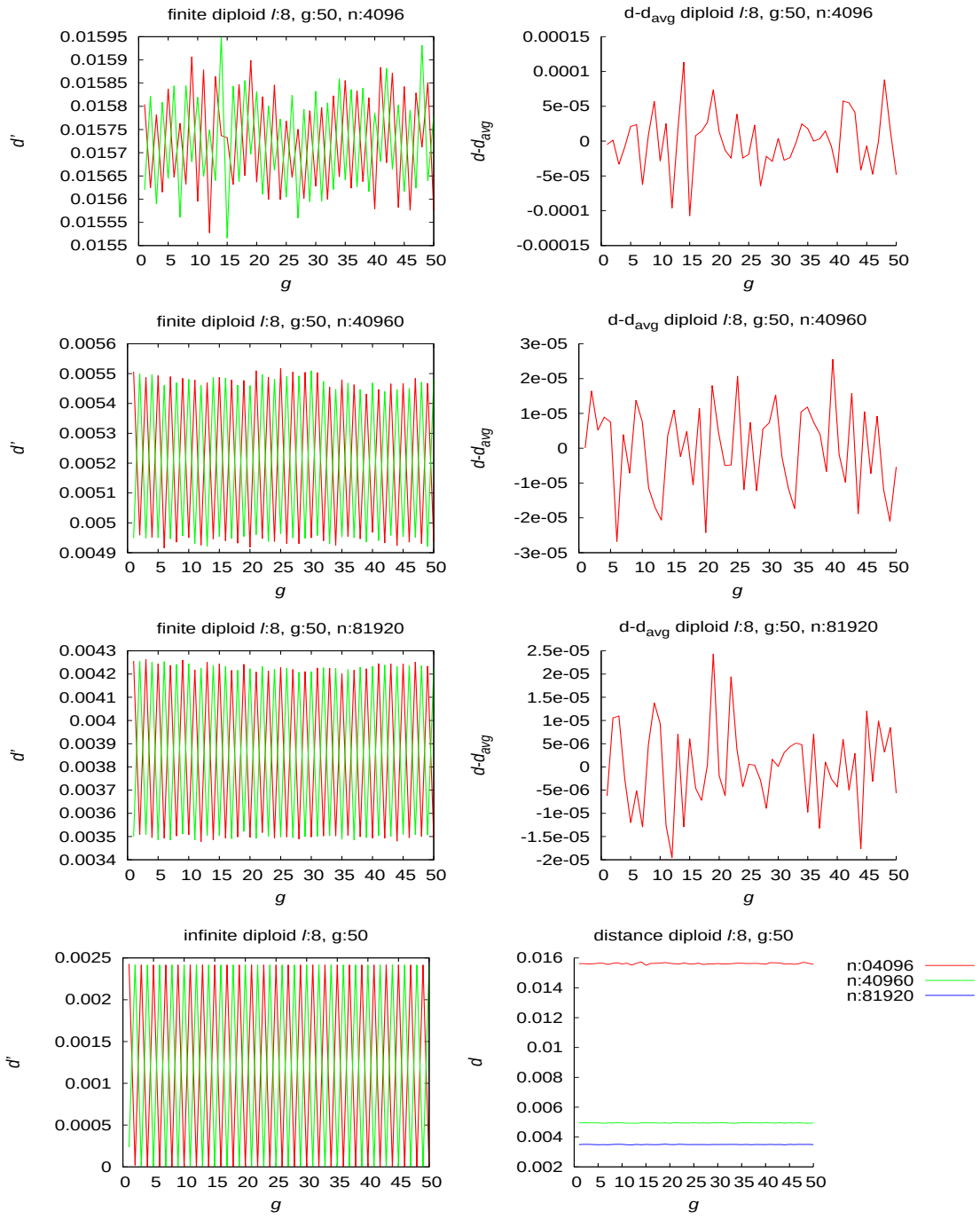


Figure 3.5: Infinite and finite diploid population oscillation behavior for genome length $\ell = 8$: In left column, d' is distance of finite population of size n or infinite population to limits for g generations. Green line is distance to $\boldsymbol{p}*$ and red line is distance to $\boldsymbol{q}*$. In right column, d is distance of finite population to infinite population for g generations and d_{avg} is average distance.

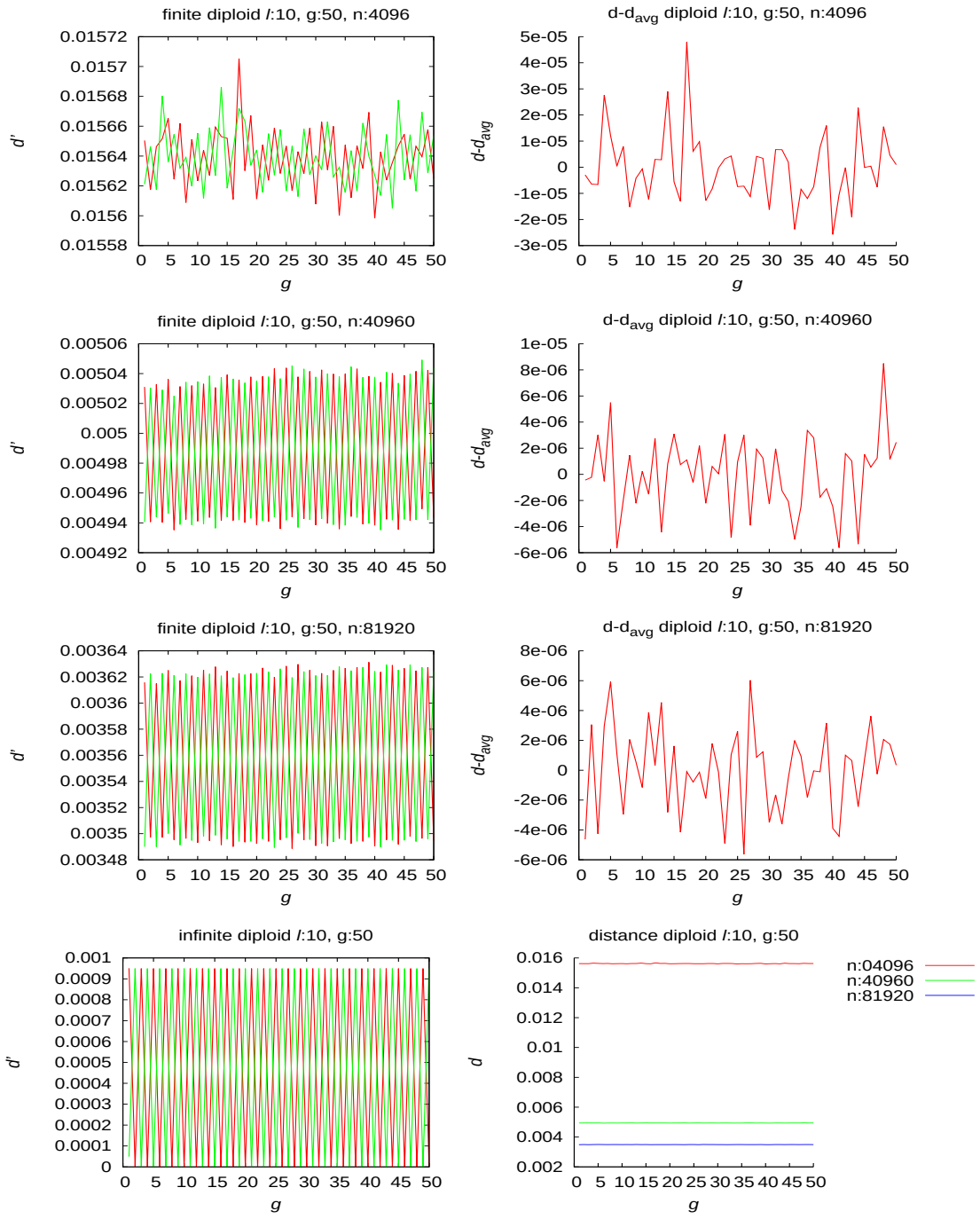


Figure 3.6: Infinite and finite diploid population oscillation behavior for genome length $\ell = 10$: In left column, d' is distance of finite population of size n or infinite population to limits for g generations. Green line is distance to $\boldsymbol{p}*$ and red line is distance to $\boldsymbol{q}*$. In right column, d is distance of finite population to infinite population for g generations and d_{avg} is average distance.

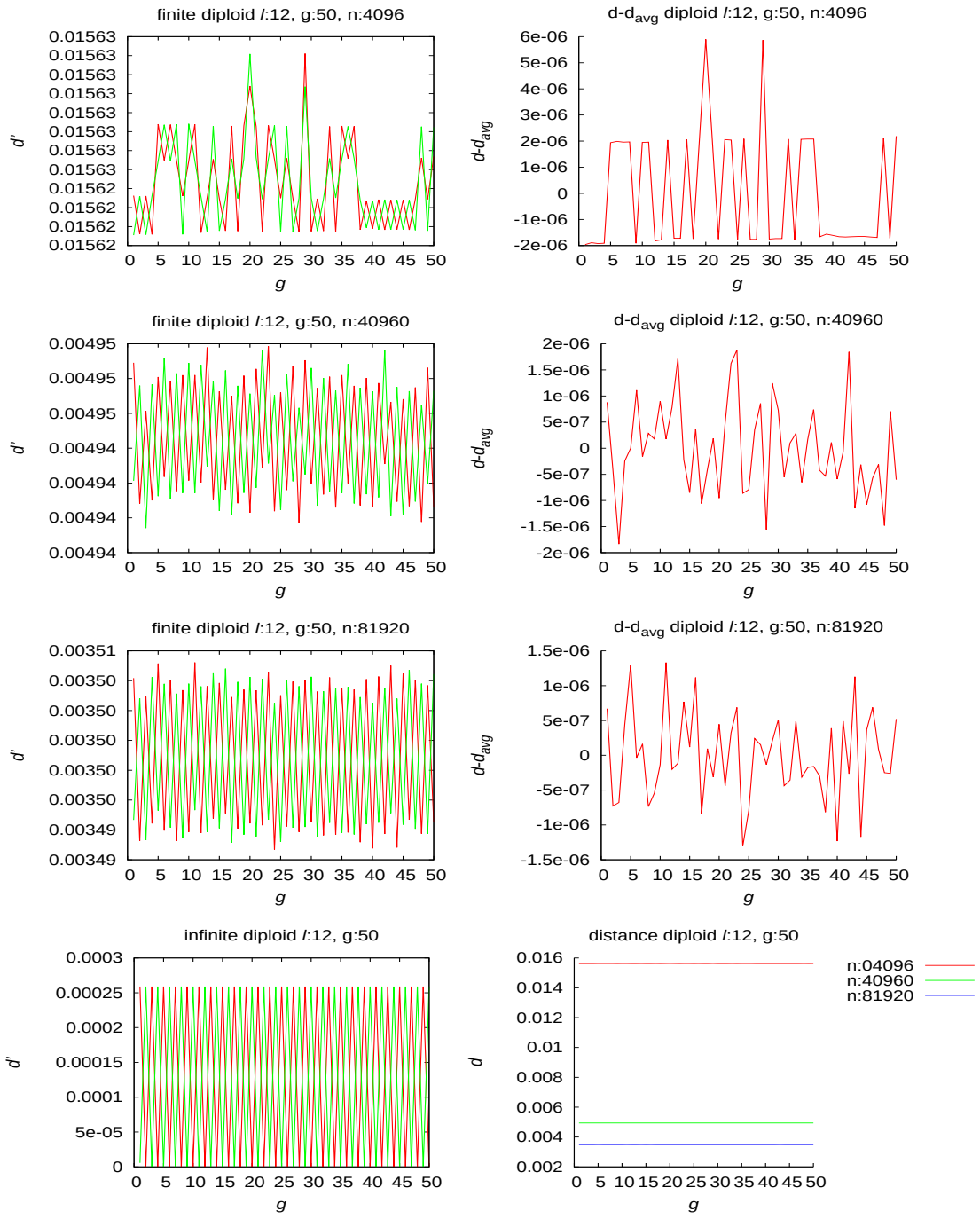


Figure 3.7: Infinite and finite diploid population oscillation behavior for genome length $\ell = 12$: In left column, d' is distance of finite population of size n or infinite population to limits for g generations. In right column, d is distance of finite population to infinite population for g generations and d_{avg} is averag Green line is distance to $\boldsymbol{p}*$ and red line is distance to $\boldsymbol{q}*$.e distance.

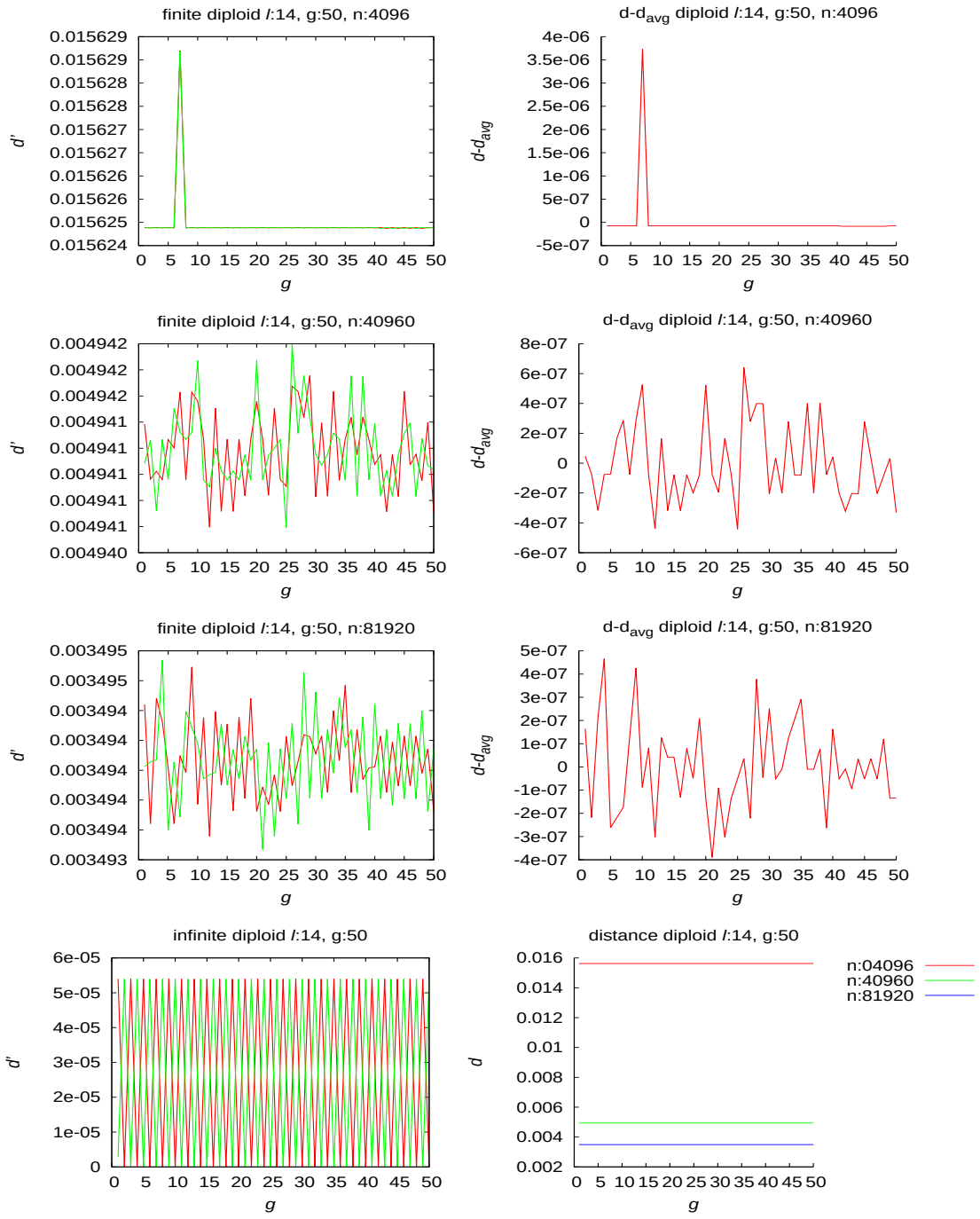


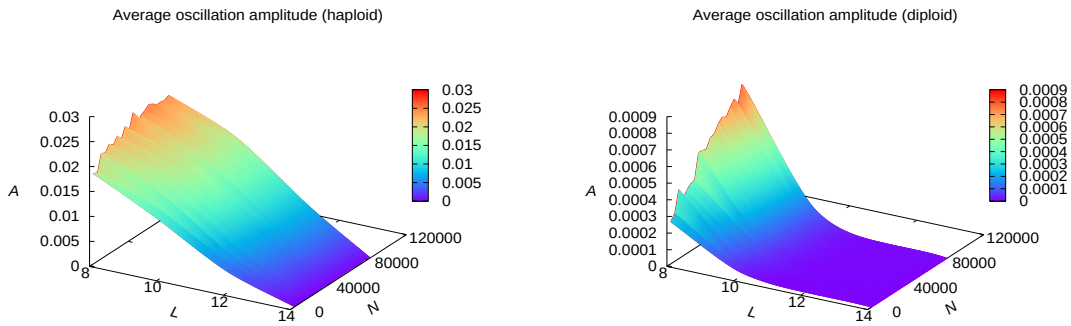
Figure 3.8: Infinite and finite diploid population oscillation behavior for genome length $\ell = 14$: In left column, d' is distance of finite population of size n or infinite population to limits for g generations. Green line is distance to $\boldsymbol{p}*$ and red line is distance to $\boldsymbol{q}*$. In right column, d is distance of finite population to infinite population for g generations and d_{avg} is average distance.

Table 3.3: Distance measured for diploid population: N is population size, ℓ is genome length, average distance between finite and infinite population is tabulated in the last three columns, and last row is expected single step distance.

ℓ	$N = 4096$	$N = 40960$	$N = 81920$
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
$1/\sqrt{N}$	0.0156	0.0049	0.0035

3.5 Discussion

For same genome length ℓ and same size finite populations, graph showing distance between finite diploid population and infinite population is smoother than graph showing distance between finite haploid population and infinite population. Average oscillation amplitude is plotted for both haploid and diploid populations as surface graphs in figures 3.9a and 3.9b.

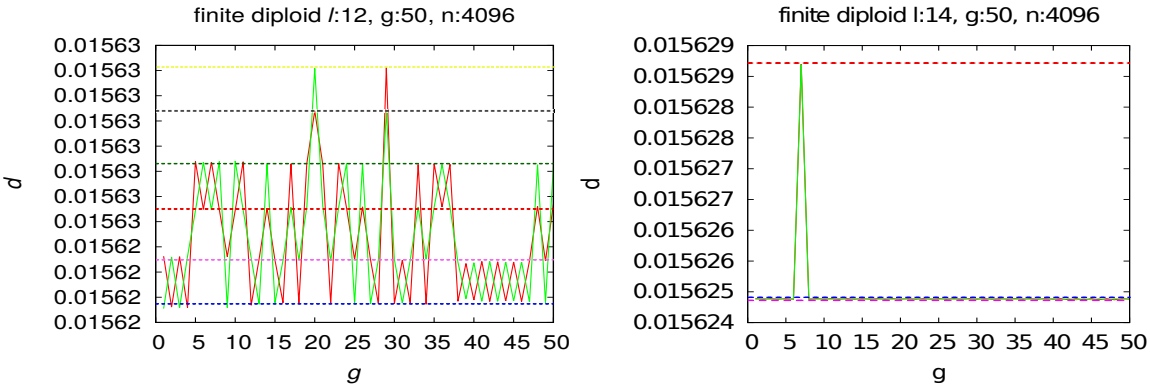


(a) **(b)**

Figure 3.9: Average oscillation amplitude: A is average amplitude of oscillation, L is genome length ℓ, and N is population size

Oscillation amplitude increases with increase in population size for both haploid and diploid populations, and better oscillations are observed with larger population

size for a given ℓ. Also amplitude of oscillation decreases with increase in ℓ value for same population size, and since total genome length of diploid population is twice that of haploid population, amplitude of oscillation of diploid population is smaller than haploid population of same population size and same value of ℓ. So for longer genome length ℓ, larger population size is needed to observe clear oscillations.



(a) (b)

Figure 3.10: Finite diploid population oscillation for $\ell = 12$ & 14 and $N = 4096$

For the diploid case, when genome length ℓ is longer ($\ell = 12$ and 14 in our simulation), and population size is small (like N = 4096 in our simulation), finite population has tendency to oscillate between different levels. Figures 3.10a and 3.10b show such tendency for $\ell = 12$ and $\ell = 14$. Very good oscillations with small amplitude were observed in temporarily stable states in these cases. Figure 3.11 shows magnified scale oscillations for $\ell = 14$ when high peak is omitted from the plot of 3.10b. As string length increases, the number of fixed points for other crossover and mutation distributions around the vicinity of finite population path also increases, and finite populations may get attracted to those near by fixed points(see Vose (1999)). With many fixed points available, there are several regions for finite populations to prefer. But when population size is large, finite populations intend to follow the infinite population, and infinite population tends to converge to a single fixed point or oscillate

between two fixed points, hence larger populations have lower tendency to jump between different levels.

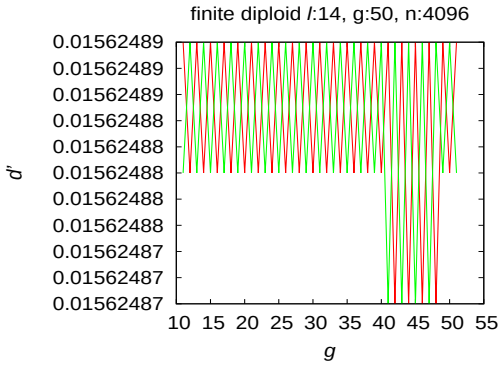


Figure 3.11: Finite diploid population oscillation for $\ell = 14$ and $N = 4096$ from 10 to 50 generations

3.6 Summary

In this chapter, we described infinite population limits predicted by Vose, and conditions for convergence to a periodic orbit. Mutation and crossover distributions were computed to satisfy the conditions for infinite populations to converge to a periodic orbit. Through experiment, we showed finite populations can also exhibit approximate oscillation. We found amplitude of oscillation is affected by string length and population size. As string length increases, oscillation amplitude decreases. Oscillation degrades as string length increases. As population size increases, oscillation amplitude increases, and also randomness in oscillation decreases. Simulations show finite populations with smaller population size and higher string length may oscillate between different pairs of points, which in our simulations occurred only in diploid populations. Moreover, the distance between finite populations and infinite populations can in practice decrease as $1/\sqrt{N}$ as the populations size increases, which agrees with previous results from chapter 2.

Chapter 4

Violation in Mutation Distribution

The results from chapter 3 show that oscillation occurs when the crossover distribution $\boldsymbol{\chi}$ and the mutation distribution $\boldsymbol{\mu}$ satisfy condition 3.3. This chapter explores the robustness of finite population oscillation when condition 3.3 is violated for $\boldsymbol{\mu}$. The violation of the condition 3.3 prevents infinite population convergence to a periodic orbit. Violation of the condition 3.3 for $\boldsymbol{\mu}$, mutation-violation as we call it, is expressed as:

$$\text{For all } g \text{ , } g \neq 0, \qquad -1 \neq \sum_{j} (-1)^{g^T j} \boldsymbol{\mu}_j \tag{4.1}$$

Mutation-violation also makes the Markov chain representing finite population evolution regular (sometimes called ergodic). If the Markov chain is regular, then positive steady state distribution exists for the Markov chain, no finite population periodic orbit exists, and perfect finite population oscillation can not occur. The question explored in this chapter is: Can finite populations exhibit approximate oscillation when the Markov chain is regular and infinite population trajectories have no periodic orbit?

Error $\boldsymbol{\epsilon}$ is introduced into the mutation distribution $\boldsymbol{\mu}$ so as to violate condition 3.3; this guarantees that infinite population trajectories have no periodic orbit. Consequently, $\boldsymbol{p}^* = \boldsymbol{q}^* = \boldsymbol{z}^*$. Going forward, we use 'limit $\boldsymbol{z}^*$' to denote

evolutionary limit when mutation distribution $\boldsymbol{\mu}$ violates condition 3.3, and 'non-violation limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$' to denote limits without violation (i.e., $\boldsymbol{\epsilon} = 0$).

4.1 Violation

The mutation distribution $\boldsymbol{\mu}$ is modified as follows

$$\boldsymbol{\mu}_i := (1 - \boldsymbol{\epsilon})\boldsymbol{\mu}_i \; ; \qquad i = \{0, 1, 2, .., 2^\ell - 1\}$$

Thus summing components of $\boldsymbol{\mu}$ distribution yields,

$$1 - \boldsymbol{\epsilon} = \sum_{i=0}^{2^\ell - 1} \boldsymbol{\mu}_i$$

Then set

$$\boldsymbol{\mu}_0 = \boldsymbol{\epsilon}$$

The modified mutation distribution $\boldsymbol{\mu}$ is normalized such that $\sum \boldsymbol{\mu}_i = 1$. The new $\boldsymbol{\mu}$ satisfies condition 4.1. Moreover, the no mutation event (using mask 0) has positive probability ($\boldsymbol{\mu}_0 = \boldsymbol{\epsilon} > 0$).

The modification described above makes it possible for any population member to mutate to any other population member. Let us exlore for two cases of g in 3.3:

1. When g is all 1s:

Any mask with a 1 at position k ($0 \leq k < \ell$) and 0 at all other positions can mutate the kth bit, and since the all 0s mask has positive probability, strings have an option to not mutate. This makes possibile for any string to mutate to any other string. Let us take an example with $\ell = 8$. Let $g = 11111111$. Then, mask $i = 00000100$ will have positive probability according to condition 3.3. Mask i can be used to mutate the sixth bit of a population member. More generally, any bit has the option of mutating or not, so any string can mutate to any other.

2. When g has at least one 0:

Any mask with a 1 at position k and 0 at all other positions will have positive probability if g also is 1 at position k. Thus, any bit where g is 1 has the option of mutating or not. Any mask with 1 in just one of the positions where g has 1s and also 1 in just one of the positions where g has 0s can be used to mutate a bit where g is 0. Let us take an example with $\ell = 8$. Let $g = 11001111$. Then, mask $i = 00000100$ will have positive probability according to condition 3.3. Also mask $j = 00010100$ will have positive probability. Mask i can be used to mutate the sixth bit, and mutation with mask i followed by mutation with mask j will result in mutating the fourth bit. More generally, any bit has the option of mutating or not, so any string can mutate to any other. Since any population can therefore mutate to any other population (this may involve many generations because there are many population members which may need to be mutated), the Markov chain is irreducible.

The Markov chain is also aperiodic. We prove this by simple induction. Let $S(n)$ be the assertion that population P can be returned to in n generations. Our base case is $n = 1$. The GA can stay in its original state P if no mutation or crossover events occur. Population P has option to not mutate to any other population, since all 0s mutation mask has positive probability. So $S(n)$ is true. Now assume $S(k)$ is true, population P can be returned to in $n = k$ generations. In the $k+1$th generation, population P has the option to stay in state P. So $S(k+1)$ is also trueand that completes the inductive proof. Since any population state can be returned to in any period of time, the Markov chain is aperiodic.

Because the Markov chain formed by GA after violation in $\boldsymbol{\mu}$ is irreducible and aperiodic, the Markov chain is regular (see Iosifescu (1980)), and a steady state distribution with positive components exists for the GA (see Minc (1988)).

Simulations were repeated with mutation-violation described above. The initial population is computed using same procedure as described in section 3.3. To explore the effects of the degree of violation, different values of $\boldsymbol{\epsilon} \in \{0.01, 0.1, 0.5\}$ are used in experiments. String lengths $\ell \in \{8, 10, 12, 14\}$ are considered for simulation.

The distances of both infinite and finite populations to limit $\boldsymbol{z}^*$ were plotted. The distances of both infinite and finite populations to non-violation limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$ (i.e. $\boldsymbol{\epsilon} = 0$) were also plotted.

4.1.1 Haploid Population $\sim \epsilon : 0.01$

The right column in figures 4.1 through 4.4 shows distance of finite and infinite haploid populations to non-violation limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$ with $\boldsymbol{\epsilon} = 0.01$. Those graphs indicate oscillating behavior of finite haploid populations given violation. Infinite populations initially oscillate given violation but the oscillation dies out. Since the value of $\boldsymbol{\epsilon}$ is small, damping of ripples is slow. The all zeros mask created in mutation distribution with $\boldsymbol{\epsilon} = 0.01$ is unlikely to be used during mutation, and when it is not used, behavior should be consistent with the behavior without violation. Moreover, $\boldsymbol{\epsilon}$ is small enough so that infinite population oscillation does not die out completely in 50 generations, even though oscillation will eventually die out completely. That is not the case for finite populations; if oscillation were to die out, it must reappear infinitely often because the Markov chain is regular.

The left column of figures 4.1 through 4.4 shows distance of finite and infinite haploid populations to limit $\boldsymbol{z}^*$ (limit with violation in mutation distribution $\boldsymbol{\mu}$) when $\boldsymbol{\epsilon} = 0.01$. The distance between finite population and limit $\boldsymbol{z}^*$ decreases as finite population size N increases, and finite population shows behavior similar to infinite population as population size increases. Average distance data of haploid population for $\boldsymbol{\mu}$ violation with $\boldsymbol{\epsilon} = 0.01$ are tabulated in table 4.1.

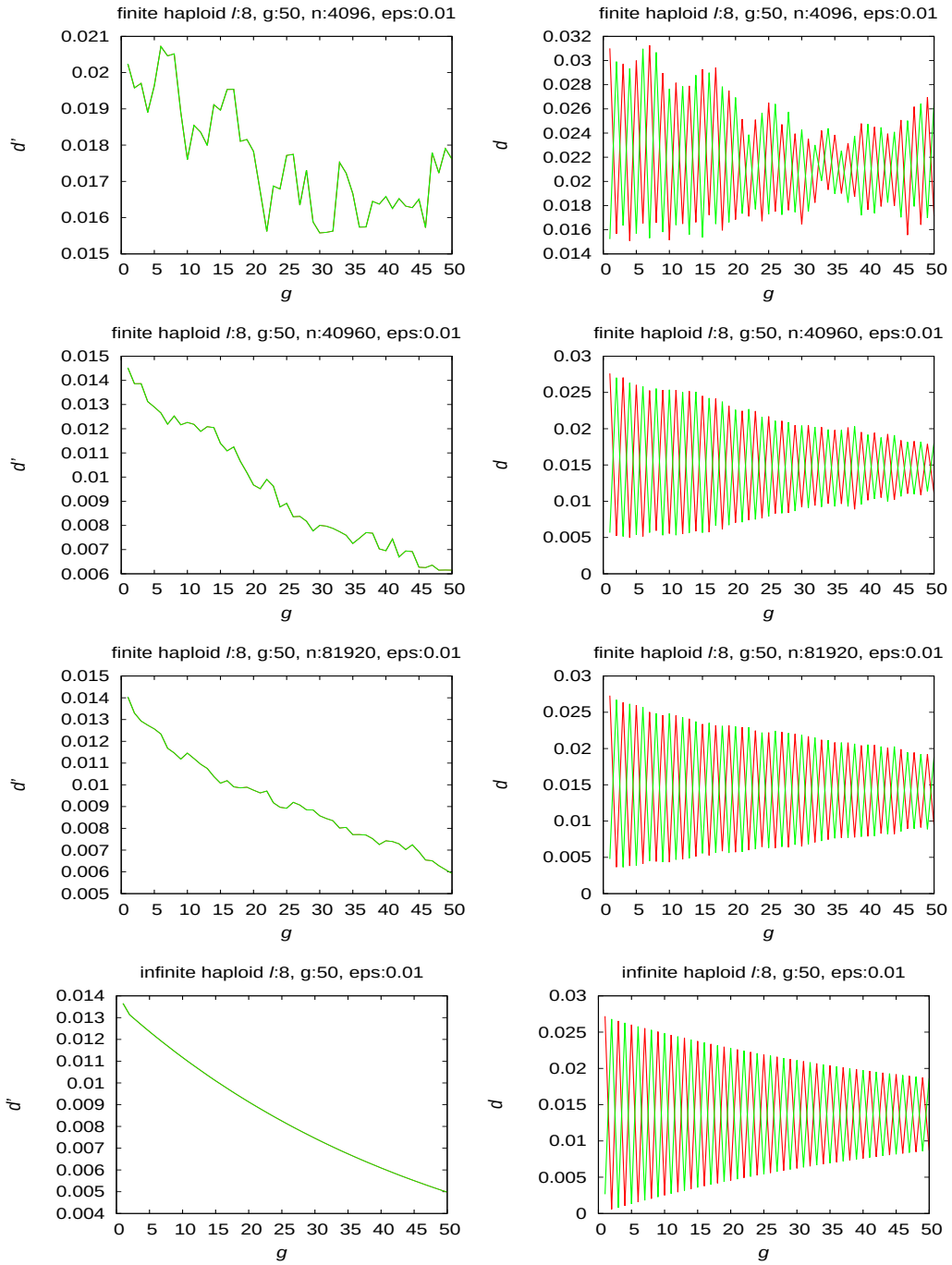


Figure 4.1: Infinite and finite haploid populations behavior for μ violation and $\ell = 8$ and $\epsilon = 0.01$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

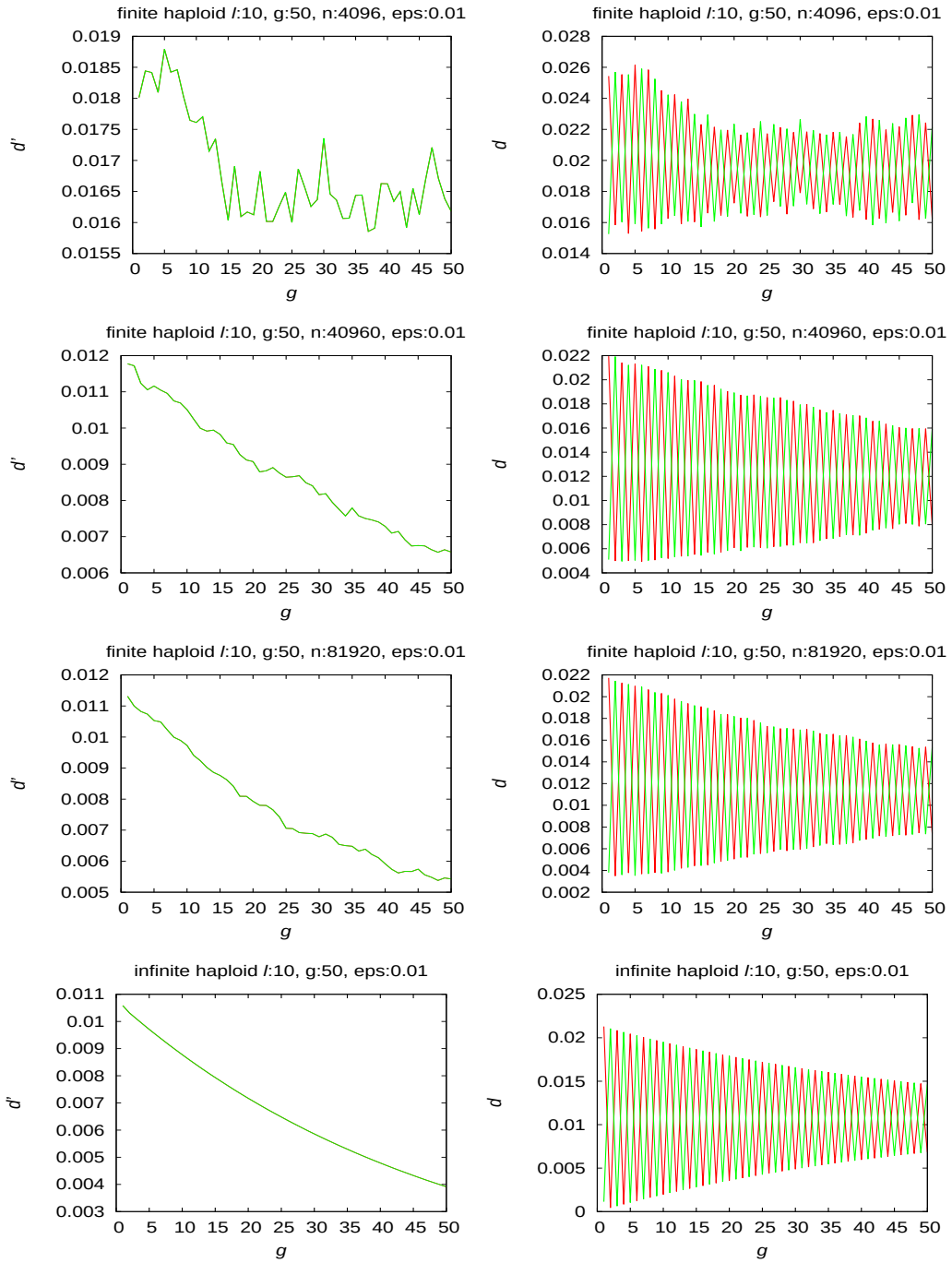


Figure 4.2: Infinite and finite haploid population behavior for μ violation, genome length $\ell = 10$ and $\epsilon = 0.01$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

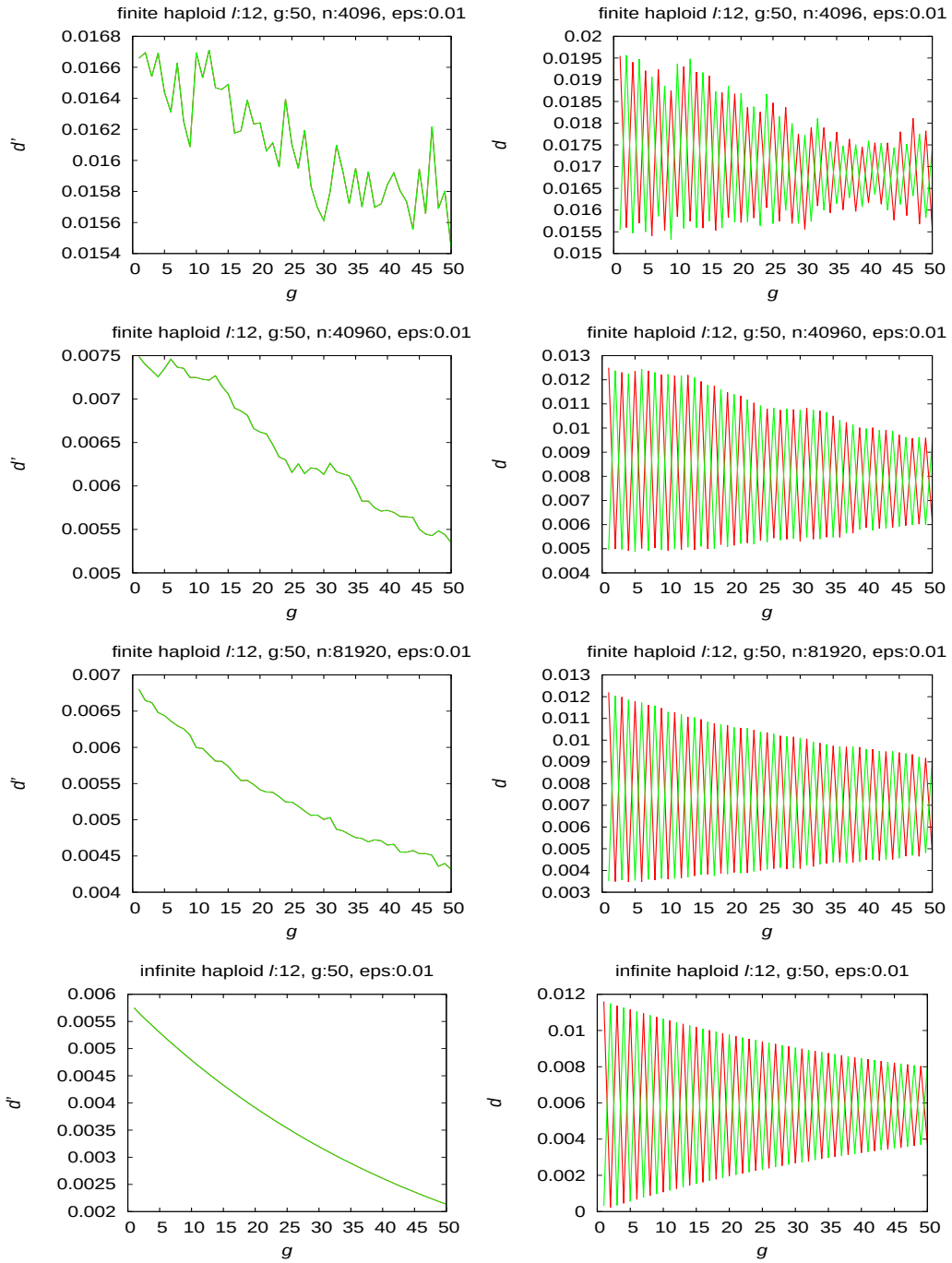


Figure 4.3: Infinite and finite haploid population behavior for μ violation, genome length $\ell = 12$ and $\epsilon = 0.01$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

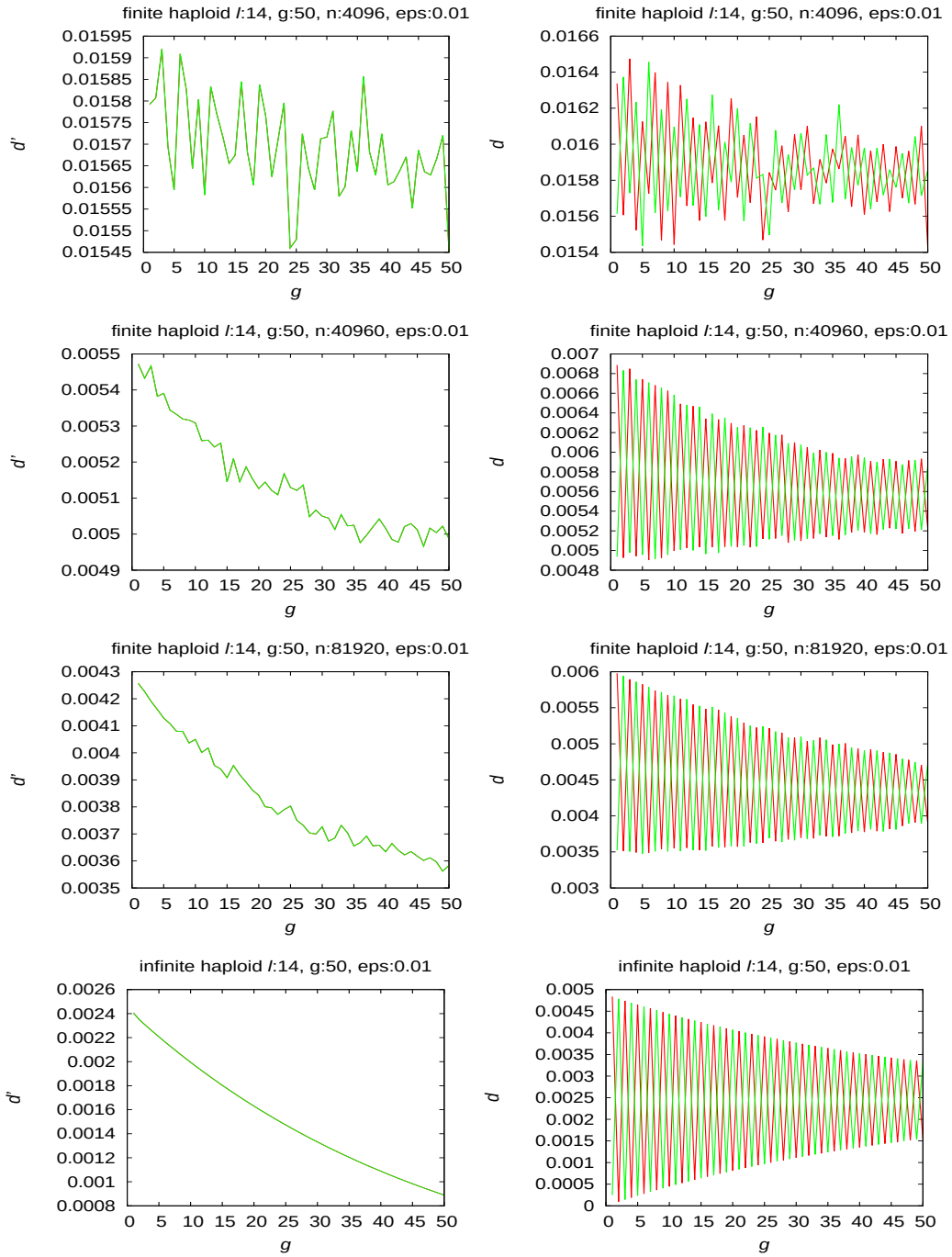


Figure 4.4: Infinite and finite haploid population behavior for μ violation, genome length $\ell = 14$ and $\epsilon = 0.01$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

Table 4.1: Distance measured for violation in $\boldsymbol{\mu}$ with $\boldsymbol{\epsilon} = 0.01$ for haploids: ℓ is genome length, average distance between finite and infinite population is tabulated in the last three columns, and last row is expected single step distance.

ℓ	$N = 4096$	$N = 40960$	$N = 81920$
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
$1/\sqrt{N}$	0.0156	0.0049	0.0035

Table 4.1 shows the average distance between finite and infinite population decreases with increasing string length, approaching the expected single step distance $1/\sqrt{N}$.

4.1.2 Haploid Population $\sim \epsilon : 0.1$

The right column in figures 4.5 through 4.8 shows distance of finite and infinite haploid populations with $\boldsymbol{\epsilon} = 0.1$ to non-violation limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Those graphs indicate oscillating behavior of finite haploid population given violation. Infinite populations initially oscillate given violation, and oscillation amplitude decreases with generation. Rate of damping of ripples with $\boldsymbol{\epsilon} = 0.1$ is larger than with $\boldsymbol{\epsilon} = 0.01$. For $\boldsymbol{\epsilon} = 0.1$, oscillations in infinite populations die out quickly, but oscillations in finite populations do not die out completely (even though it appears to be dying out) because the Markov chain is regular. Since the Markov chain is regular, finite population must visit every population state infinitely often. So, if finite population oscillation were to die out, it must reappear infinitely often.

The left column of figures 4.5 through 4.8 shows distance of finite and infinite haploid populations to limit $\boldsymbol{z}^*$ (limit with violation in $\boldsymbol{\mu}$) when $\boldsymbol{\epsilon} = 0.1$. The distance decreases as finite population size increases, and finite population shows behavior similar to infinite population behavior as population size grows.

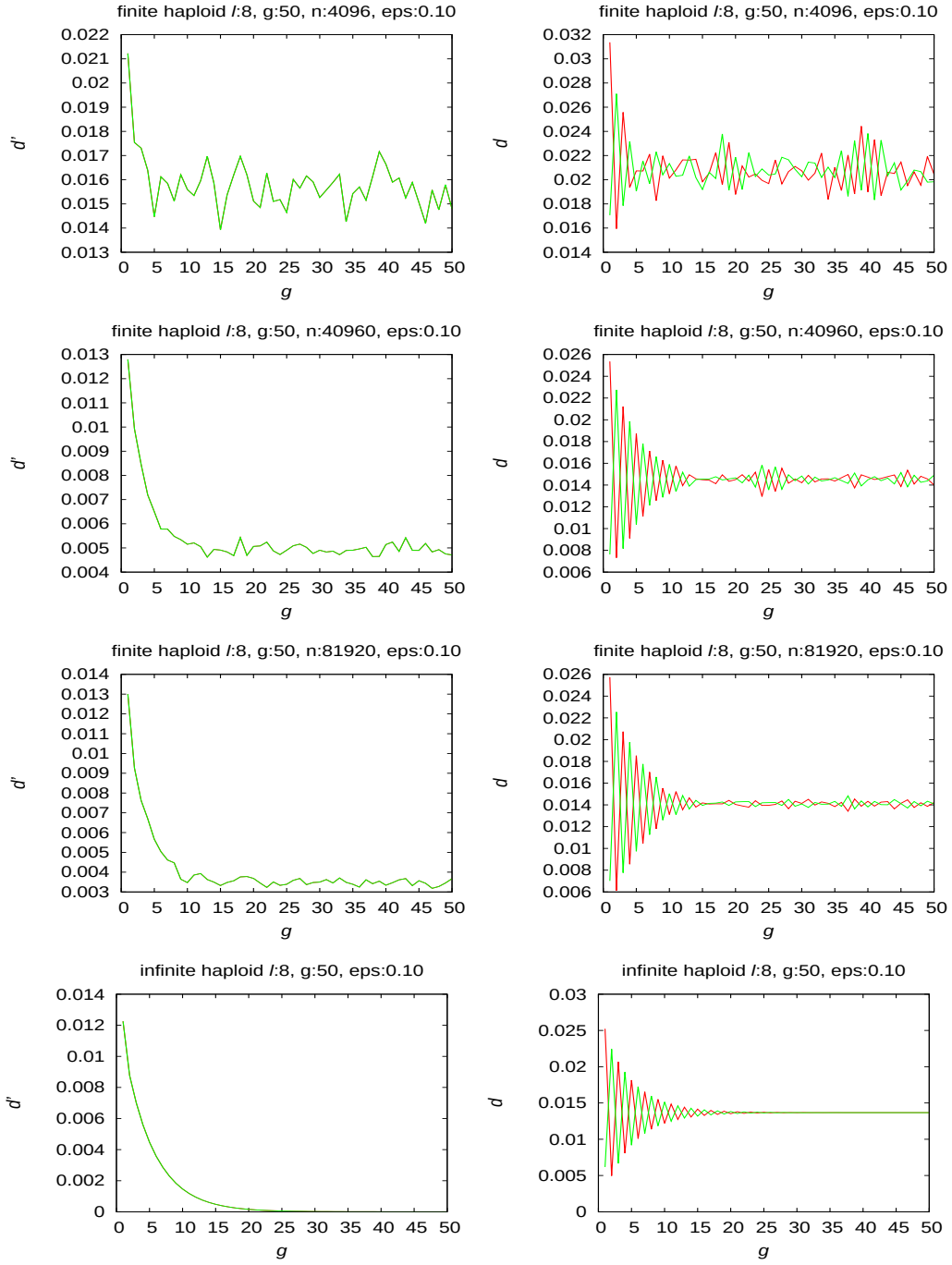


Figure 4.5: Infinite and finite haploid population behavior for μ violation and $\ell = 8$ and $\epsilon = 0.1$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

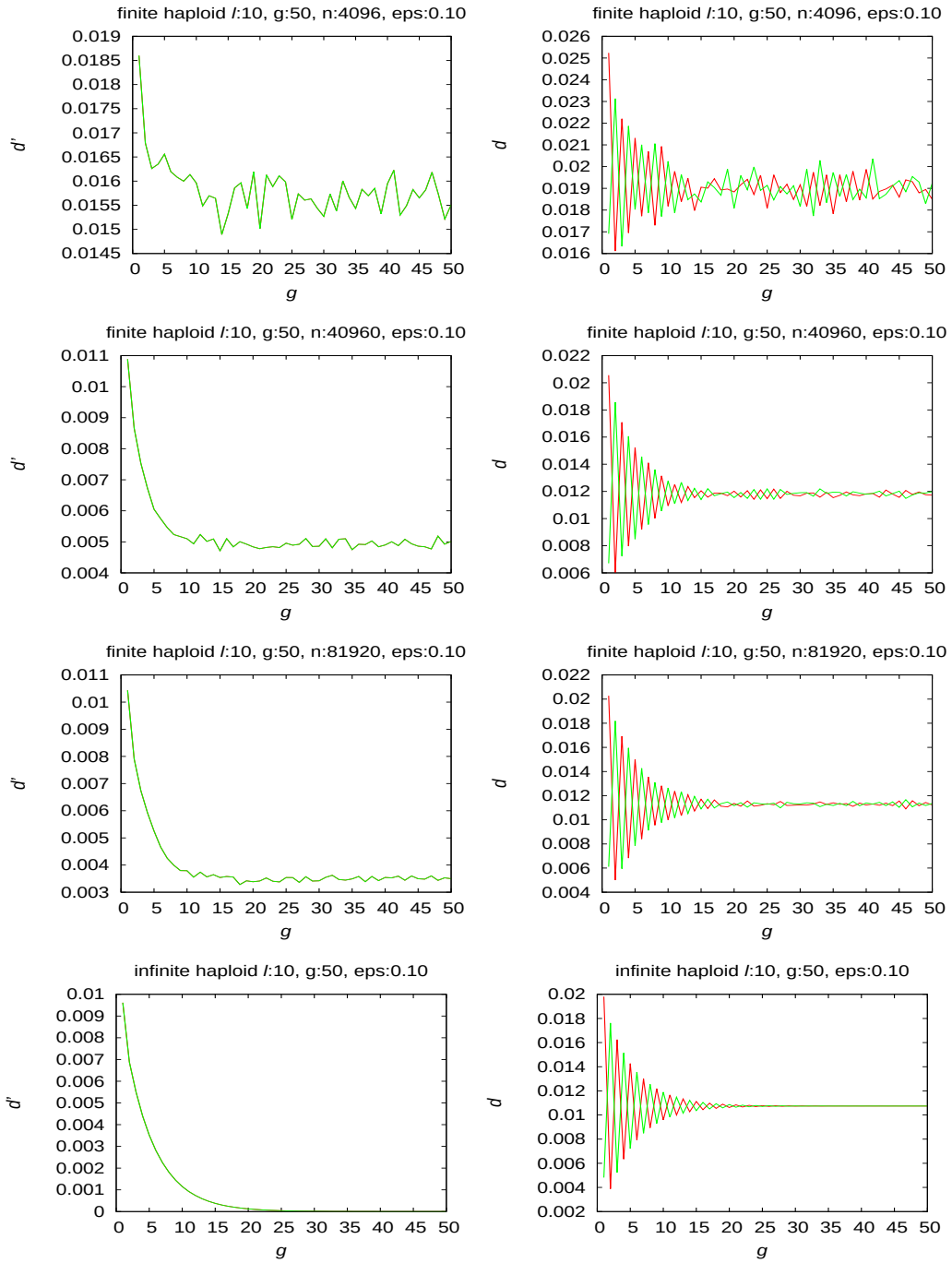


Figure 4.6: Infinite and finite haploid population behavior for μ violation, genome length $\ell = 10$ and $\epsilon = 0.1$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

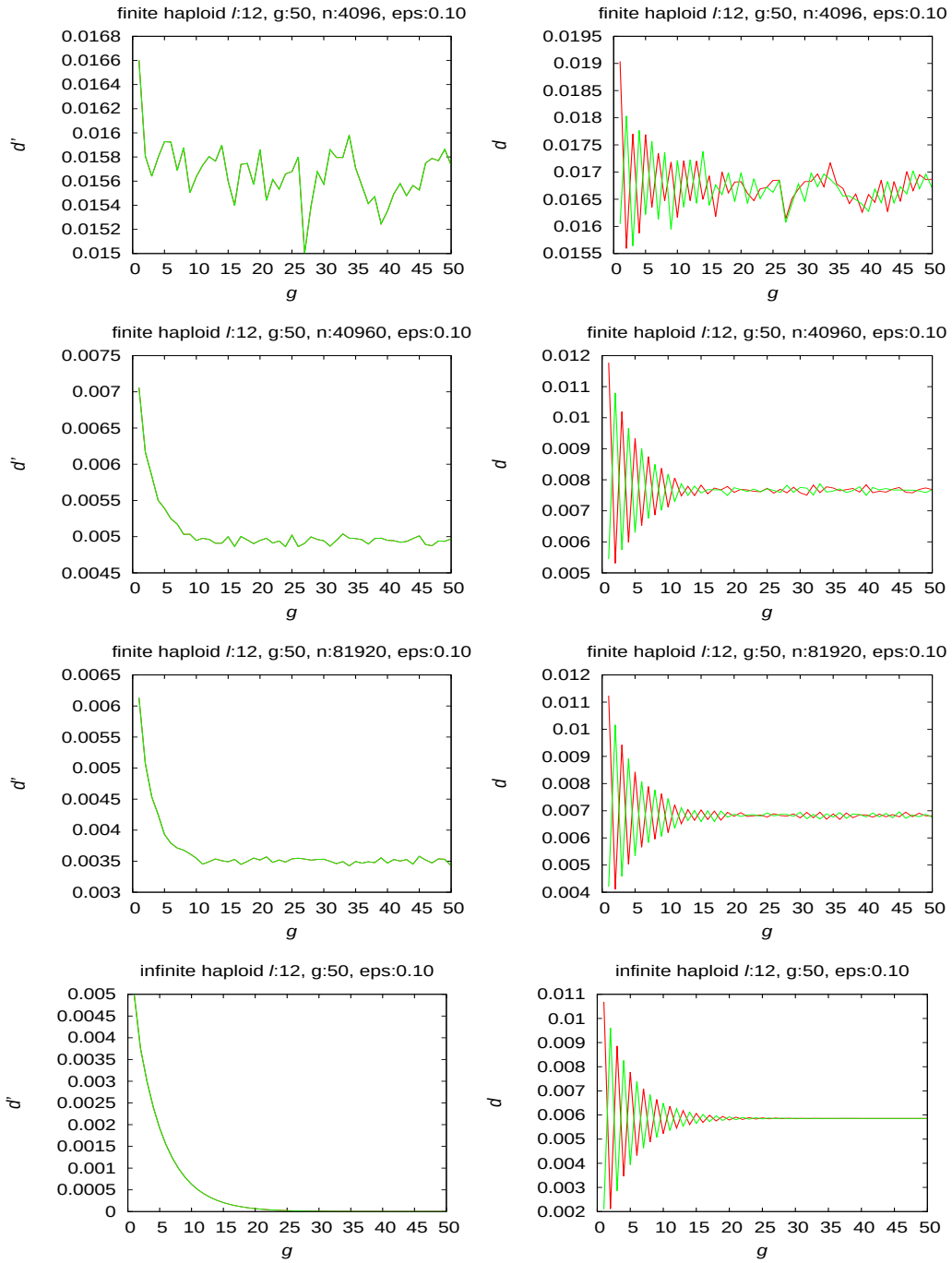


Figure 4.7: Infinite and finite haploid population behavior for μ violation, genome length $\ell = 12$ and $\epsilon = 0.1$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

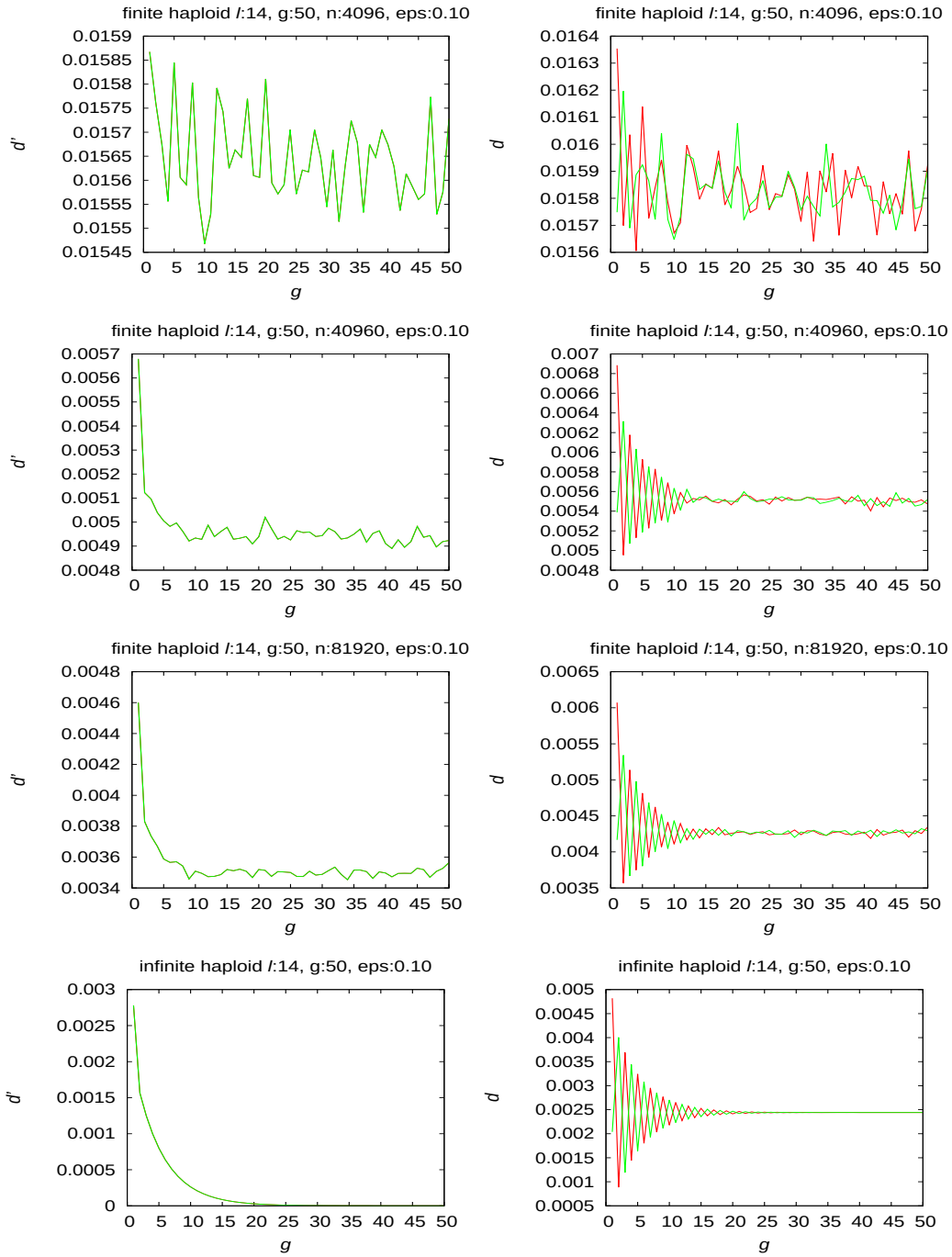


Figure 4.8: Infinite and finite haploid population behavior for μ violation, genome length $\ell = 14$ and $\epsilon = 0.1$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

Average distance data for haploid population in case of violation in $\boldsymbol{\mu}$ distribution with $\boldsymbol{\epsilon} = 0.1$ for different finite population size N is tabulated in table 4.2.

Table 4.2: Distance measured for violation in $\boldsymbol{\mu}$ with $\boldsymbol{\epsilon} = 0.1$ for haploids: ℓ is genome length, average distance between finite and infinite population is tabulated in the last three columns, and last row is expected single step distance.

ℓ	$N = 4096$	$N = 40960$	$N = 81920$
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
$1/\sqrt{N}$	0.0156	0.0049	0.0035

Table 4.2 show average distance between finite population and infinite population decreases with increasing string length, approaching the expected single step distance $1/\sqrt{N}$.

4.1.3 Haploid Population $\sim \epsilon : 0.5$

The right column in figures 4.9 through 4.12 shows distance of finite and infinite haploid populations with $\boldsymbol{\epsilon} = 0.5$ to non-violation limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Neither finite nor infinite populations show noticable oscillation given violation. The all zeros mask created in mutation distribution with $\boldsymbol{\epsilon} = 0.5$ has a large probability of being used during mutation, so oscillation decreased signficantly.

The left column of figures 4.9 through 4.12 shows distance of finite and infinite haploid populations to limit $\boldsymbol{z}^*$ (limit with violation in mutation distribution $\boldsymbol{\mu}$) when $\boldsymbol{\epsilon} = 0.5$. The distance decreases as finite population size increases, and finite population shows behavior similar to infinite population behavior as finite population size grows. Average distance data for haploid population in case of violation in $\boldsymbol{\mu}$ distribution with $\boldsymbol{\epsilon} = 0.5$ for different finite population size N are tabulated in table 4.3.

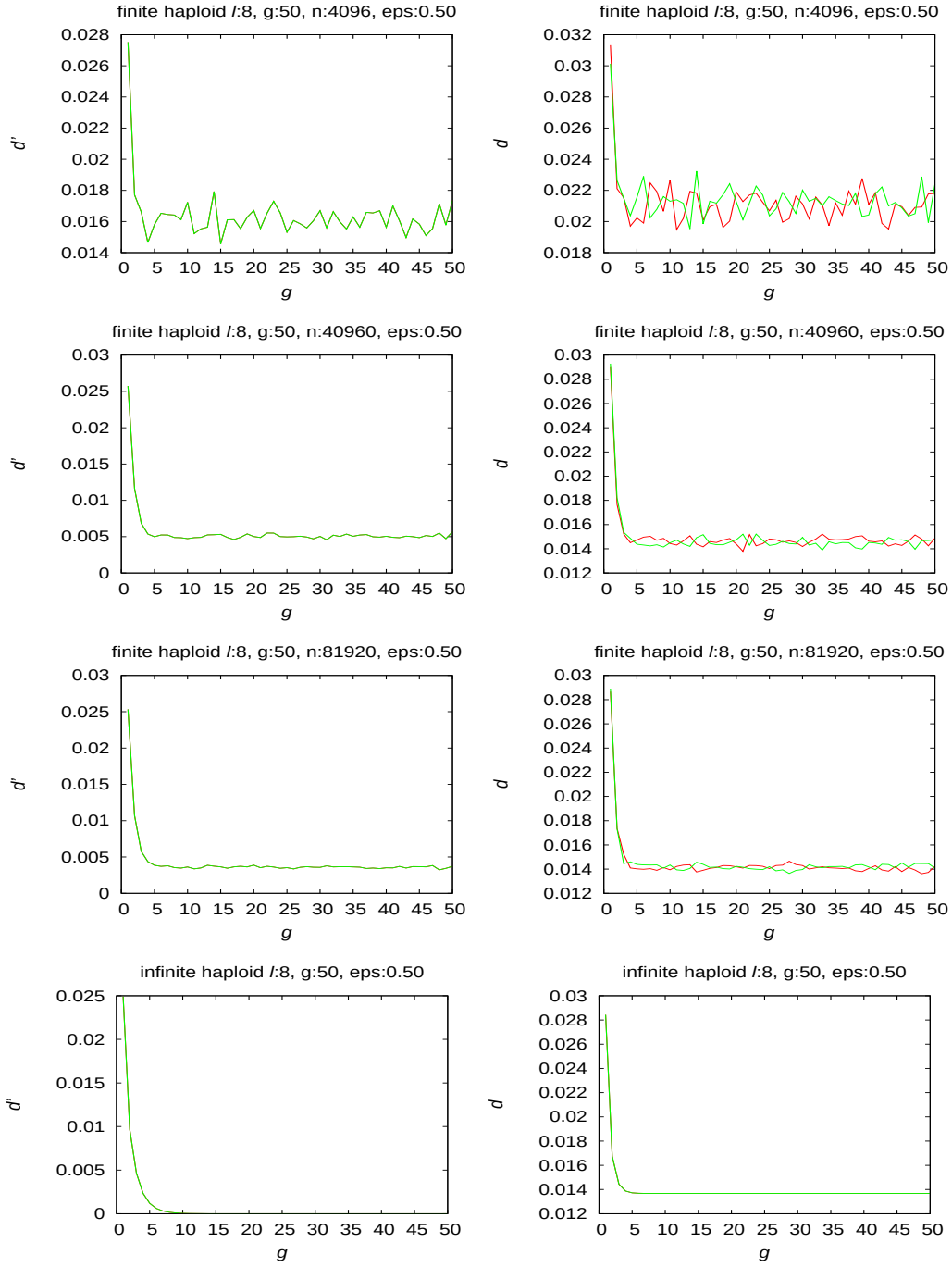


Figure 4.9: Infinite and finite haploid population behavior for μ violation and $\ell = 8$ and $\epsilon = 0.5$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

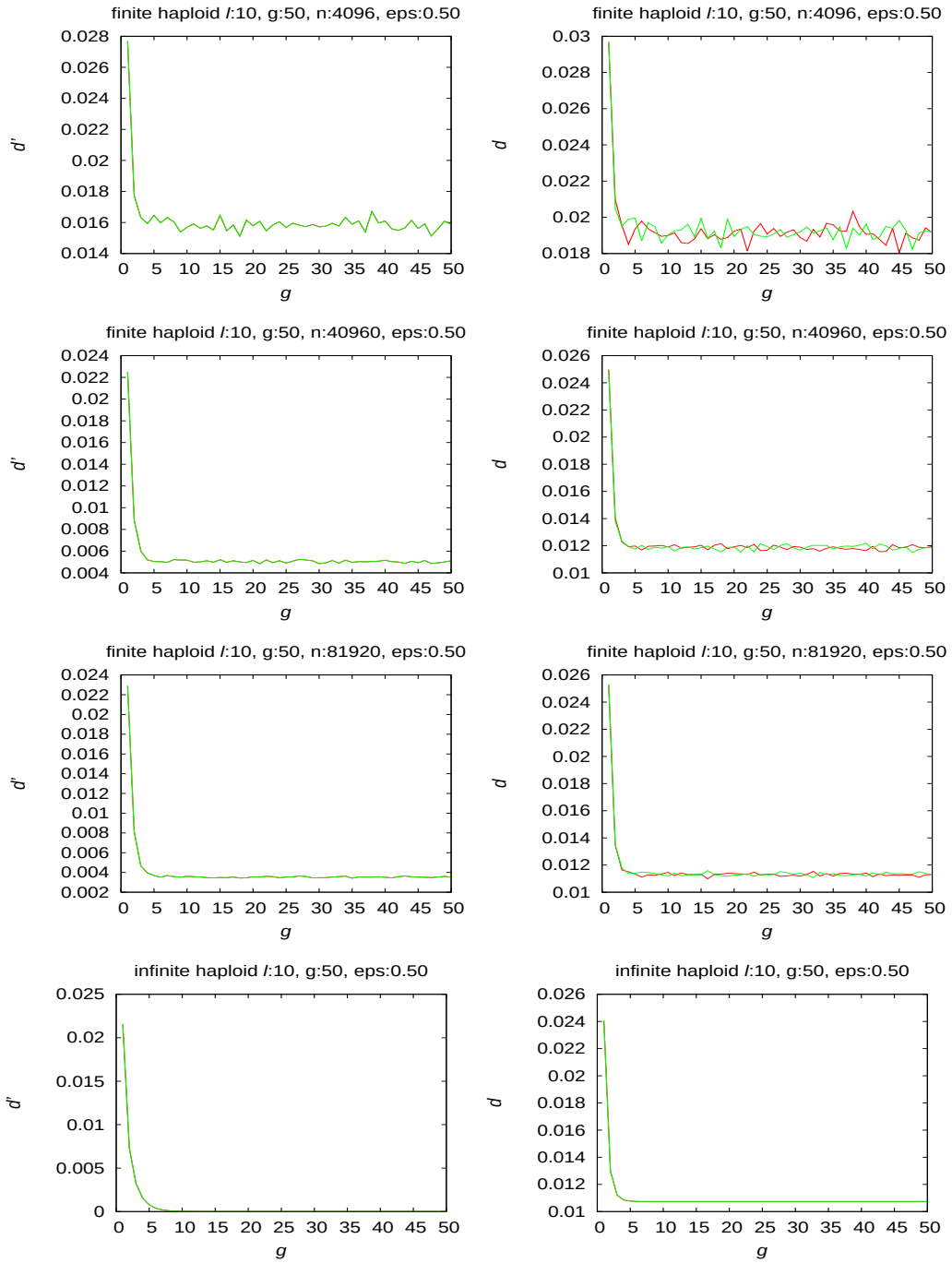


Figure 4.10: Infinite and finite haploid population behavior μ for violation, genome length $\ell = 10$ and $\epsilon = 0.5$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

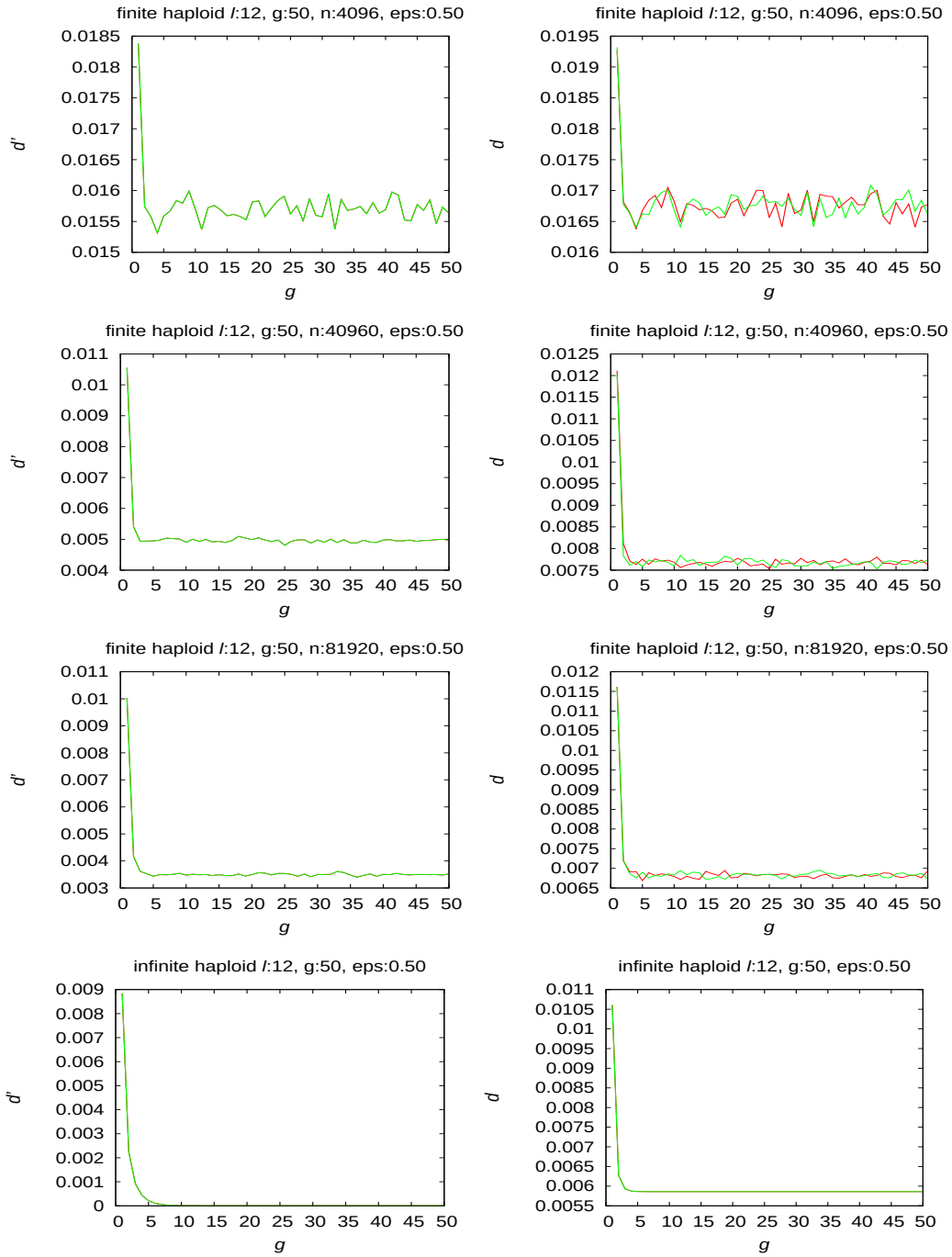


Figure 4.11: Infinite and finite haploid population behavior μ for violation, genome length $\ell = 12$ and $\epsilon = 0.5$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

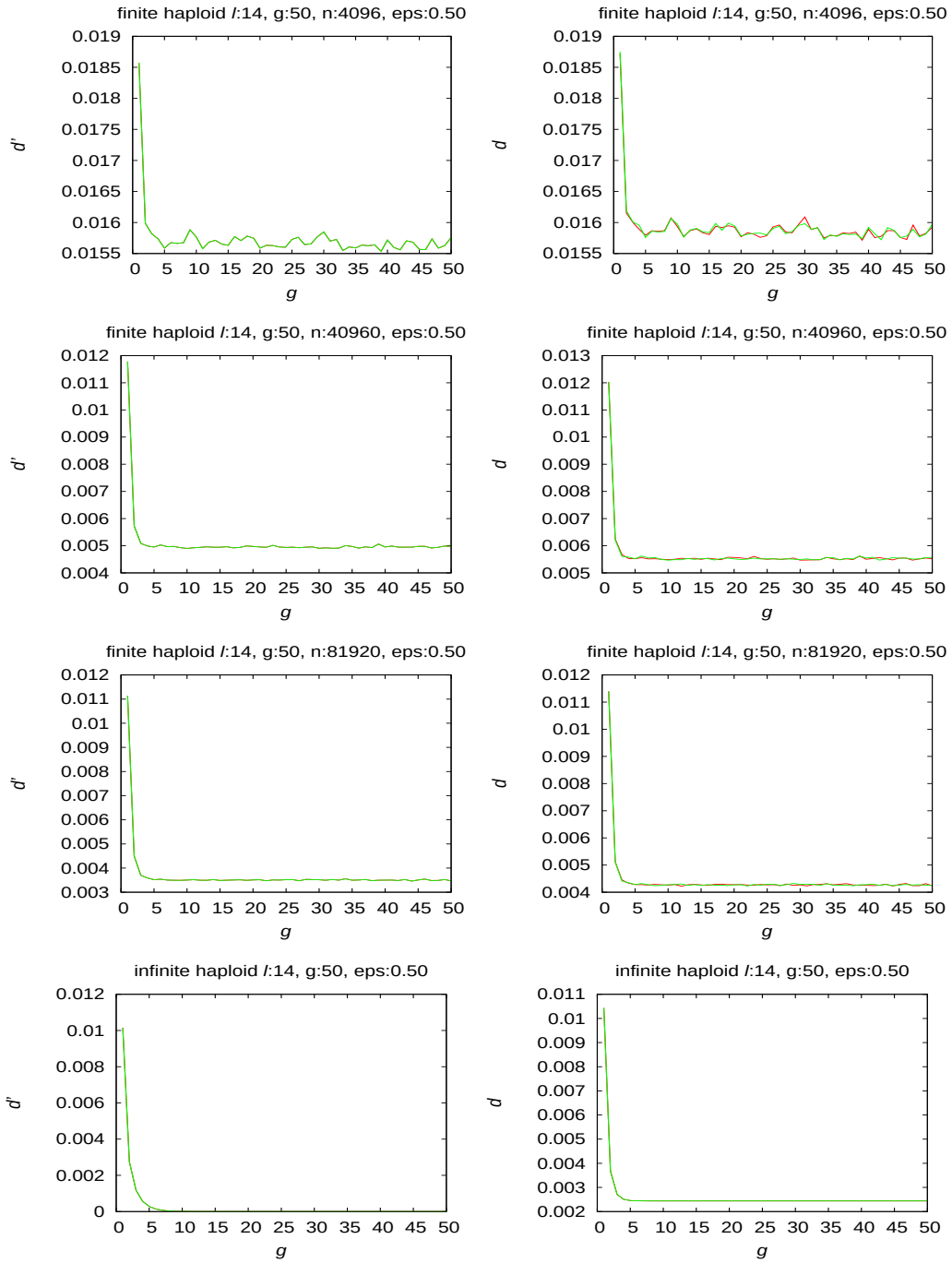


Figure 4.12: Infinite and finite haploid population behavior μ for violation, genome length $\ell = 14$ and $\epsilon = 0.5$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

Table 4.3: Distance measured for violation in $\boldsymbol{\mu}$ with $\boldsymbol{\epsilon} = 0.5$ for haploids: ℓ is genome length, average distance between finite and infinite population is tabulated in the last three columns, and last row is expected single step distance.

ℓ	$N = 4096$	$N = 40960$	$N = 81920$
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	0.0036
$1/\sqrt{N}$	0.0156	0.0049	0.0035

Table 4.3 shows that the average distance between finite population and infinite population decreases with increasing string length, approaching the expected single step distance $1/\sqrt{N}$.

4.1.4 Diploid Population $\sim \epsilon : 0.01$

The right column in figures 4.13 through 4.16 shows distance of finite and infinite diploid populations with $\boldsymbol{\epsilon} = 0.01$ to non-violation limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Those graphs indicate oscillating behavior of finite diploid population given violation. Infinite populations initially oscillate given violation but the oscillation dies out. Since the value of $\boldsymbol{\epsilon}$ is very small, damping of ripples is slow. Infinite population oscillation does not die out in 50 generations but will eventually die out. Even though oscillation in finite population is tapering down, it will not die out completely; because the Markov chain is regular, finite population oscillation will reappear infinitely often.

The left column of figures 4.13 through 4.16 shows distance of finite and infinite diploid populations to limit $\boldsymbol{z}^*$ (limit with violation in mutation distribution $\boldsymbol{\mu}$) when $\boldsymbol{\epsilon} = 0.01$. The distance decreases as finite population size increases, and finite population shows behavior similar to infinite population as population size grows. Average distance data for diploid population for $\boldsymbol{\mu}$ violation with $\boldsymbol{\epsilon} = 0.01$ are tabulated in table 4.4.

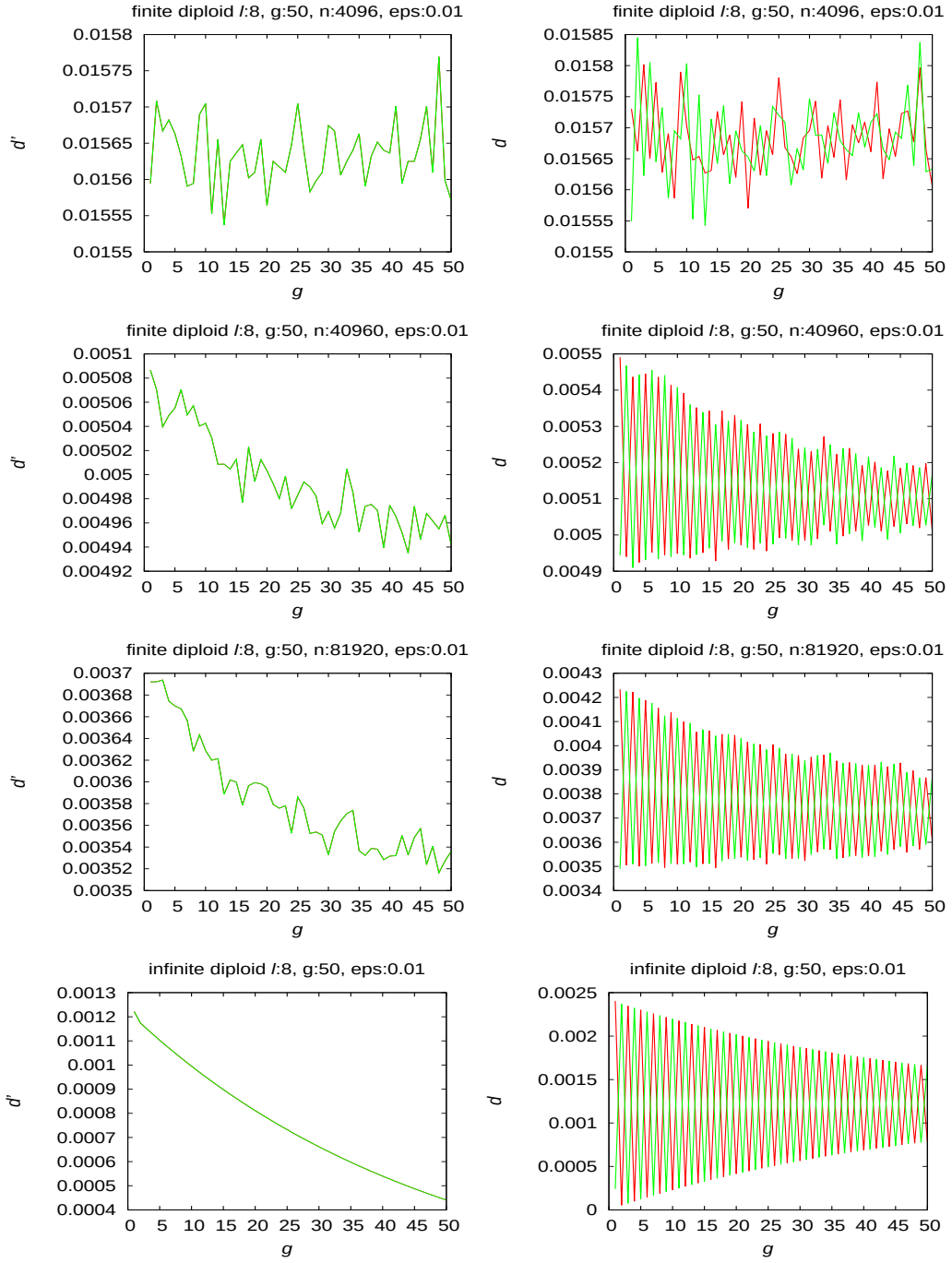


Figure 4.13: Infinite and finite diploid population behavior for μ violation, $\ell = 8$ and $\epsilon = 0.01$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

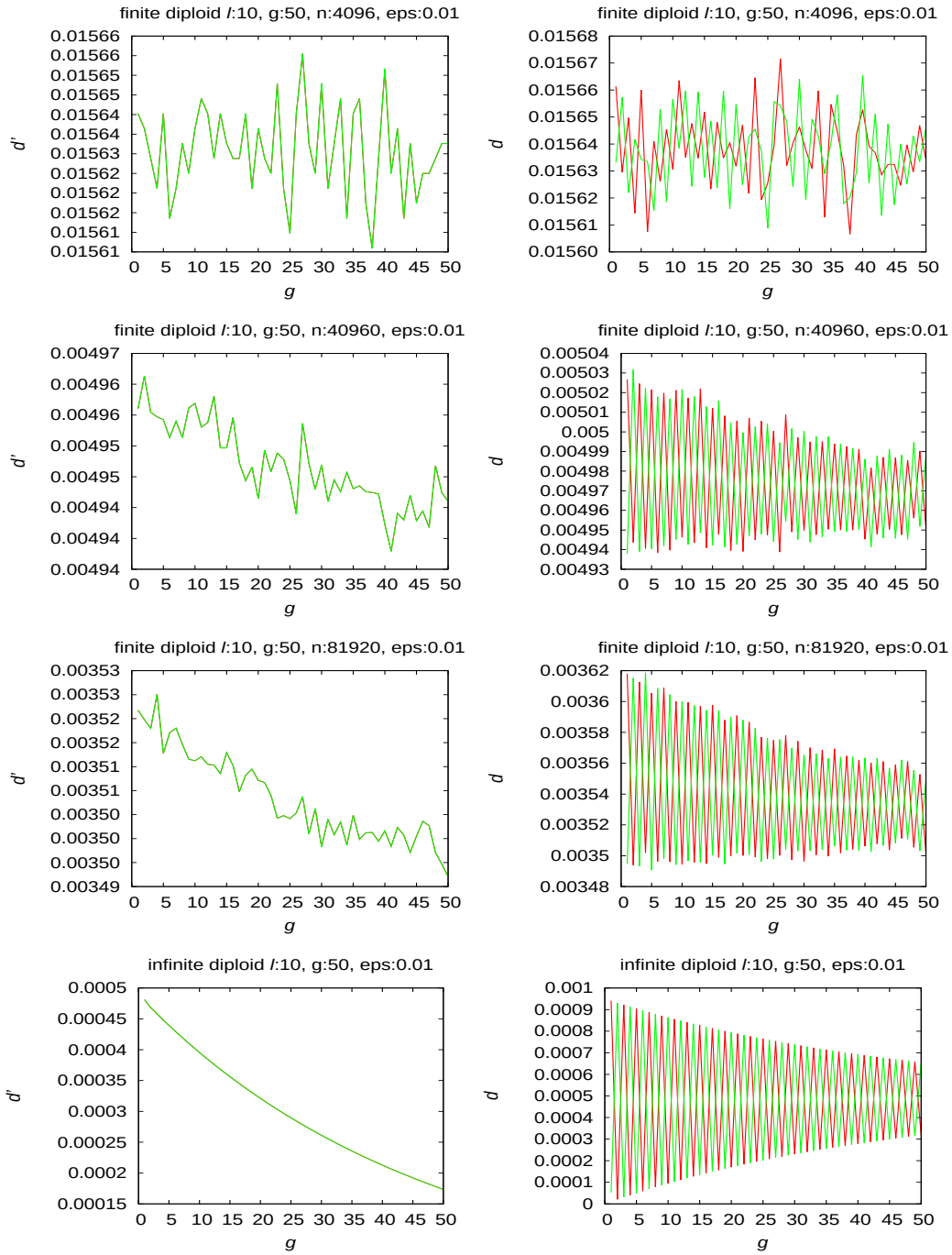


Figure 4.14: Infinite and finite diploid population behavior for μ violation, genome length $\ell = 10$ and $\epsilon = 0.01$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

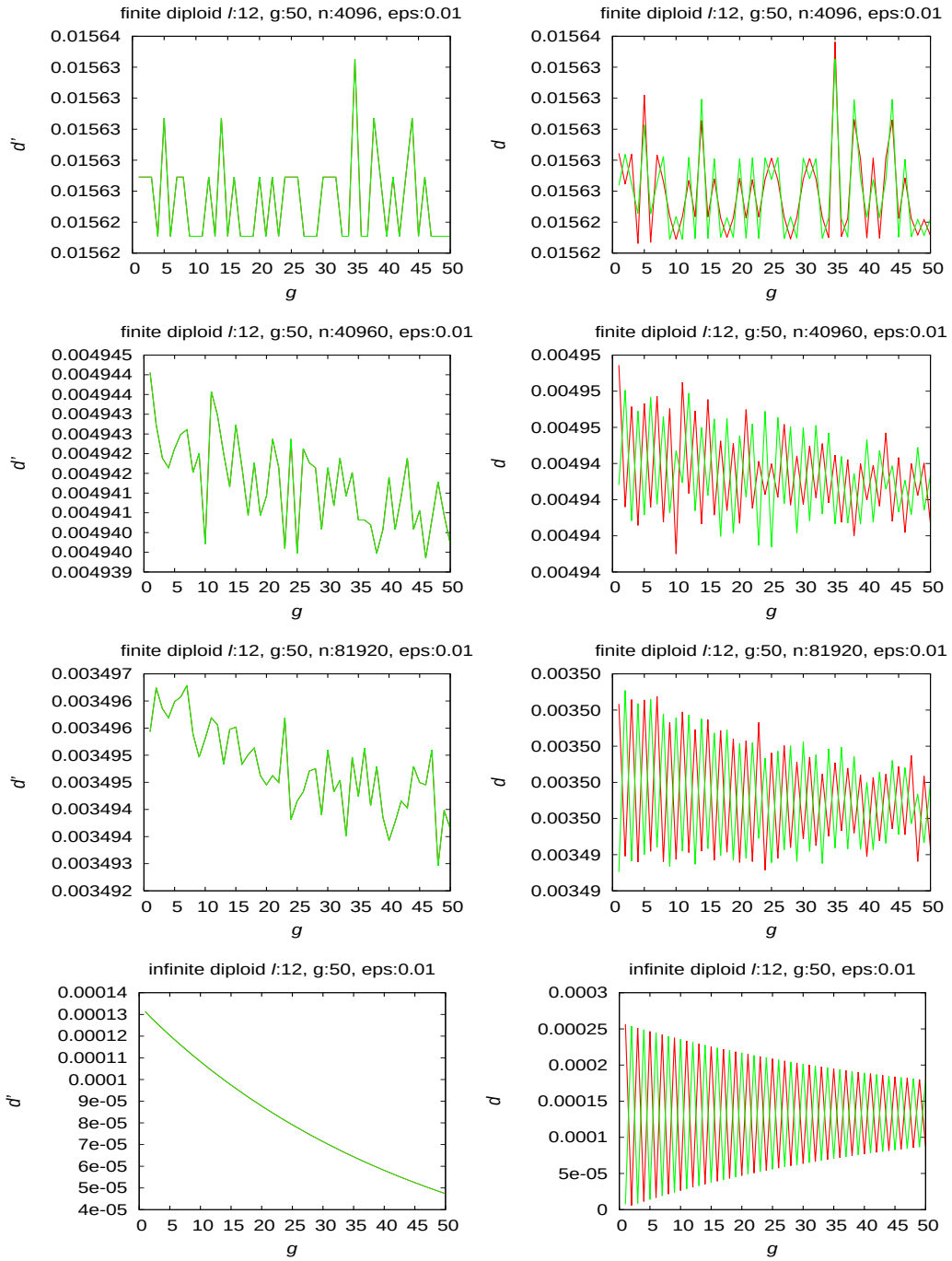


Figure 4.15: Infinite and finite diploid population behavior for μ violation, genome length $\ell = 12$ and $\epsilon = 0.01$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

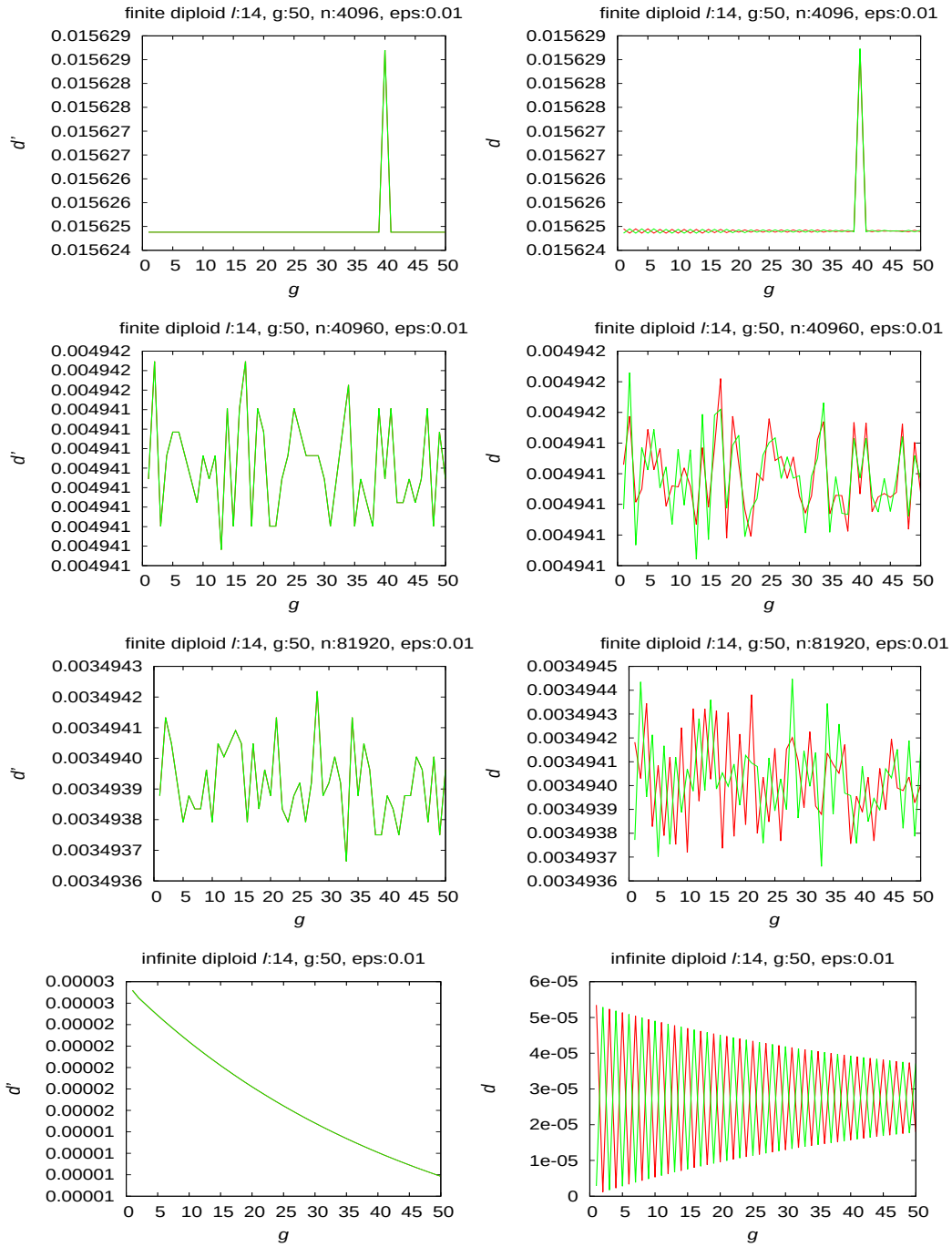


Figure 4.16: Infinite and finite diploid population behavior for μ violation, genome length $\ell = 14$ and $\epsilon = 0.01$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

Table 4.4: Distance measured for violation in μ with $\epsilon = 0.01$ for diploids: ℓ is genome length, average distance between finite and infinite population is tabulated in the last three columns, and last row is expected single step distance.

ℓ	$N = 4096$	$N = 40960$	$N = 81920$
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
$1/\sqrt{N}$	0.0156	0.0049	0.0035

Table 4.4 shows that the average distance between finite population and infinite population decreases with increasing string length, approaching the expected single step distance $1/\sqrt{N}$.

4.1.5 Diploid Population $\sim \epsilon : 0.1$

The right column in figures 4.17 through 4.20 shows distance of finite and infinite diploid populations with $\boldsymbol{\epsilon} = 0.1$ to non-violation limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Those graphs indicate the oscillating behavior of finite diploid populations given violation; oscillation amplitudes decrease with time. Like in the haploid case, (for $\boldsymbol{\epsilon} = 0.1$) oscillations in infinite populations die out quickly, but finite population oscillation does not (and will reappear infinitely often). Rate of damping of ripples with $\boldsymbol{\epsilon} = 0.1$ is larger than with $\boldsymbol{\epsilon} = 0.01$. The all zeros mask created in mutation distribution with $\boldsymbol{\epsilon} = 0.1$ has a larger probability of being used during mutation as compared with the $\boldsymbol{\epsilon} = 0.01$ case. More random wiggling of finite population is noticed than in case of $\boldsymbol{\epsilon} = 0.01$, and as value of ℓ increases, random wiggling is more prevalent.

The left column of figures 4.17 through 4.20 shows distance of finite and infinite diploid populations to limit $\boldsymbol{z}^*$ (limit with violation in mutation distribution $\boldsymbol{\mu}$) when $\boldsymbol{\epsilon} = 0.1$. The distance decreases as finite population size increases, and finite population shows behavior similar to infinite population behavior as population size grows.

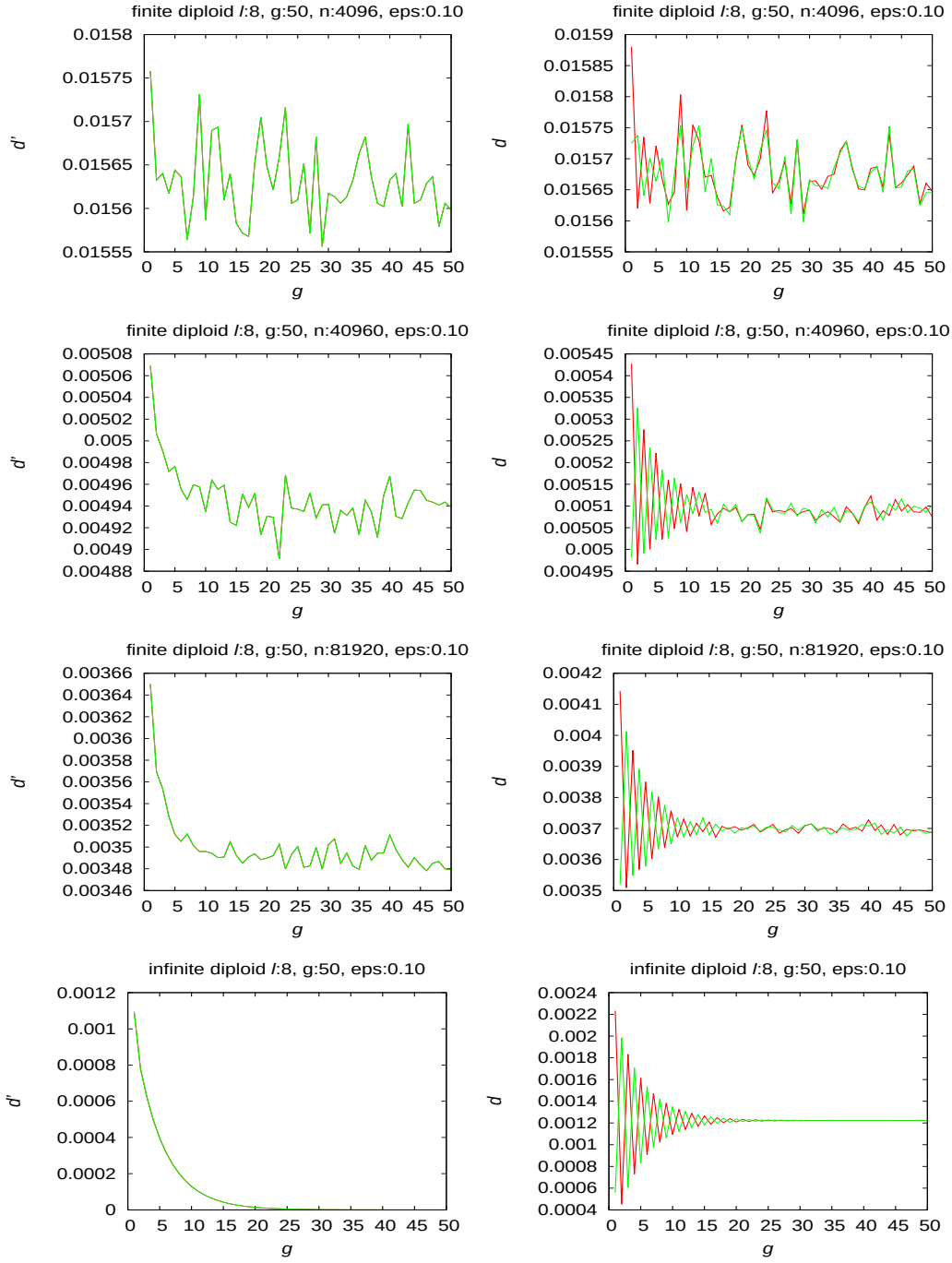


Figure 4.17: Infinite and finite diploid population behavior for μ violation, $\ell = 8$ and $\epsilon = 0.1$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

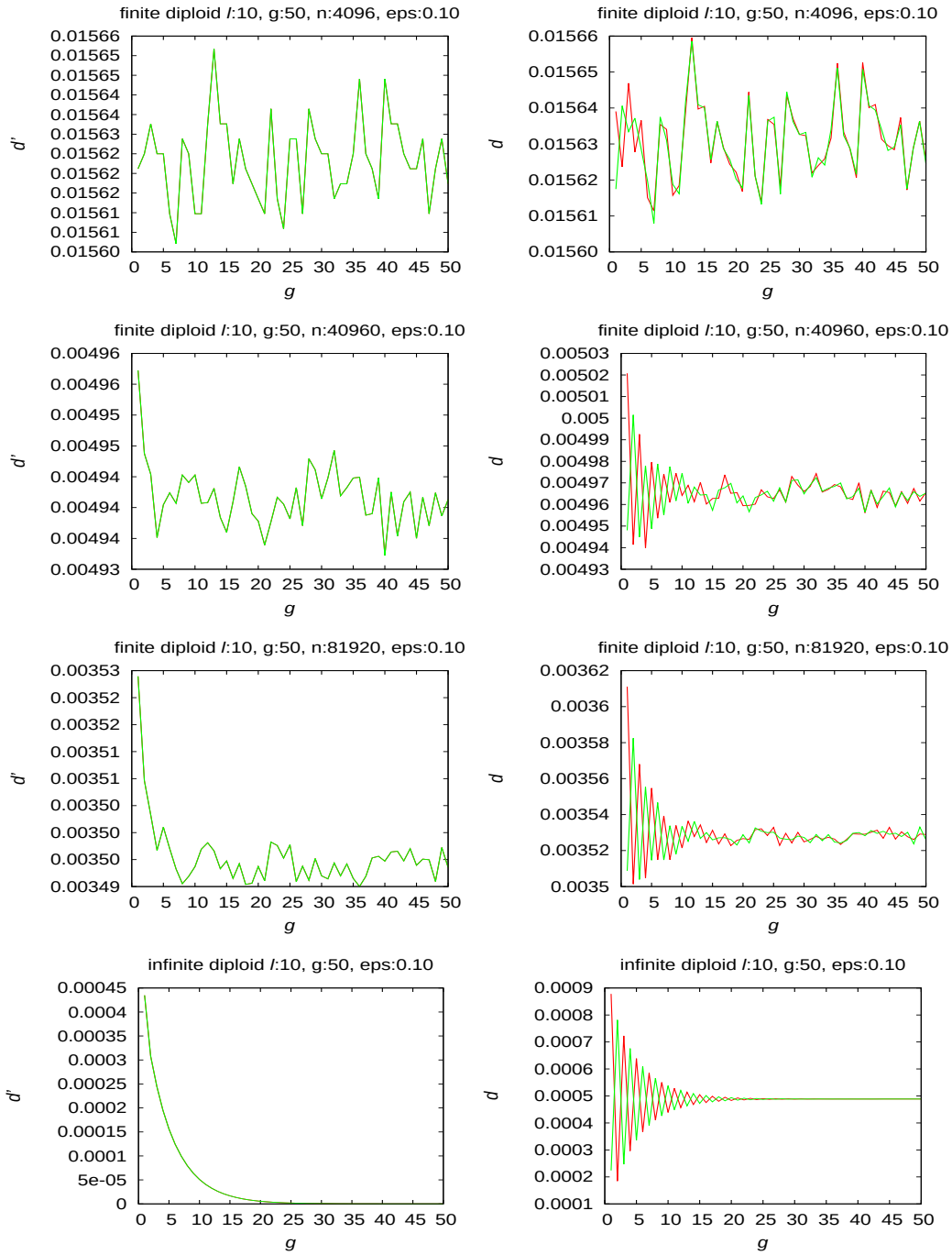


Figure 4.18: Infinite and finite diploid population behavior for μ violation, genome length $\ell = 10$ and $\epsilon = 0.1$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

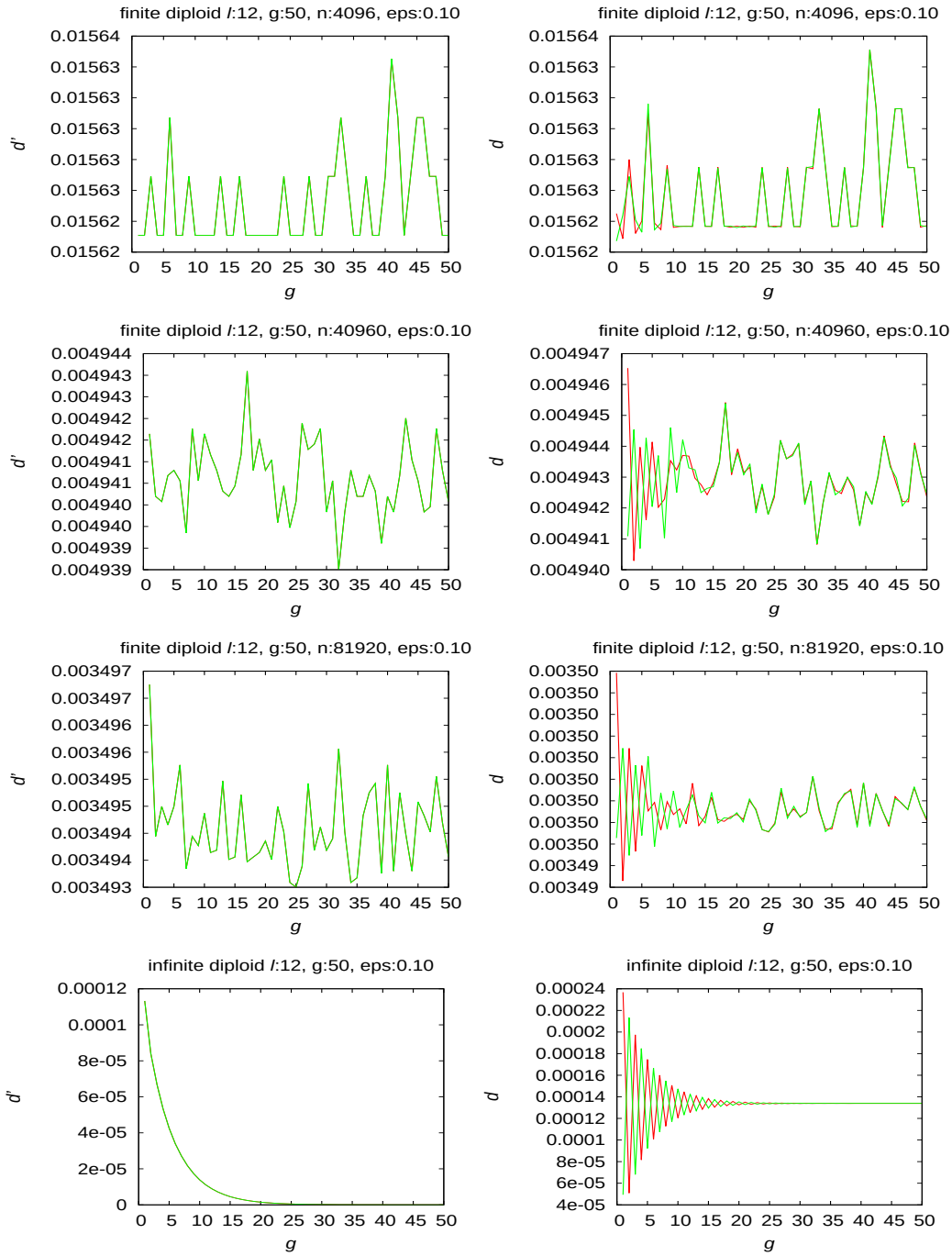


Figure 4.19: Infinite and finite diploid population behavior for μ violation, genome length $\ell = 12$ and $\epsilon = 0.1$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

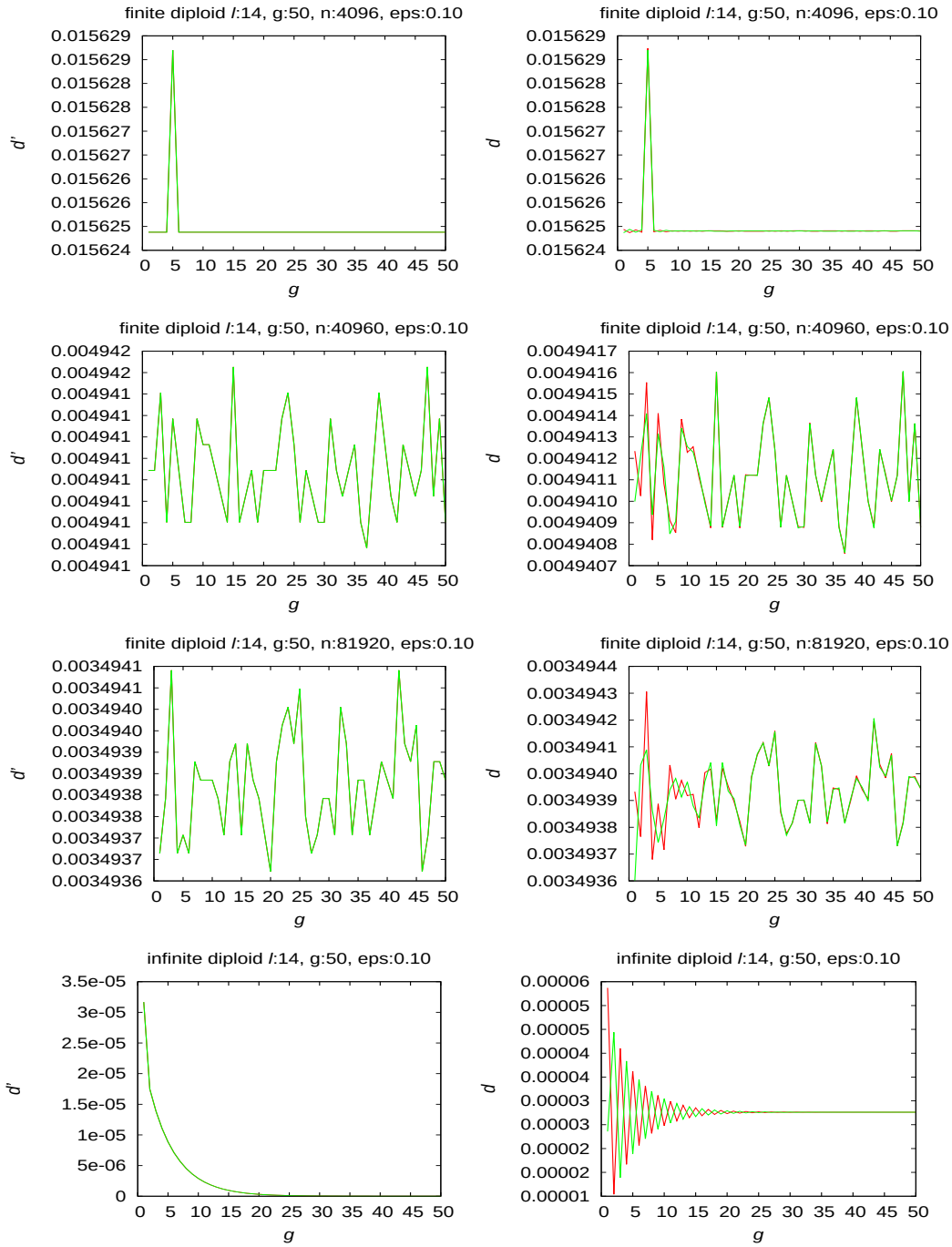


Figure 4.20: Infinite and finite diploid population behavior for μ violation, genome length $\ell = 14$ and $\epsilon = 0.1$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

Average distance data for diploid population in case of violation in $\boldsymbol{\mu}$ distribution with $\boldsymbol{\epsilon} = 0.1$ are tabulated in table 4.5.

Table 4.5: Distance measured for violation in $\boldsymbol{\mu}$ with $\boldsymbol{\epsilon} = 0.1$ for diploids: ℓ is genome length, average distance between finite and infinite population is tabulated in the last three columns, and last row is expected single step distance.

ℓ	$N = 4096$	$N = 40960$	$N = 81920$
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
$1/\sqrt{N}$	0.0156	0.0049	0.0035

Table 4.5 shows the average distance between finite population and infinite population agrees with the expected single step distance $1/\sqrt{N}$.

4.1.6 Diploid Population $\sim \epsilon : 0.5$

The right column in figures 4.21 through 4.24 shows distance of finite and infinite diploid populations with $\boldsymbol{\epsilon} = 0.5$ to non-violation limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Neither finite nor infinite populations show noticeable oscillation given violation. The all zeros mask created in mutation distribution with $\boldsymbol{\epsilon} = 0.5$ has a large probability of being used during mutation, so finite population oscillation decreased significantly.

The left column of figures 4.21 through 4.24 shows distance of finite and infinite diploid populations to limit $\boldsymbol{z}^*$ (limit with violation in mutation distribution $\boldsymbol{\mu}$) when $\boldsymbol{\epsilon} = 0.5$. The distance decreases as finite population size increases, and finite population shows behavior similar to infinite population as population size grows. Average distance data for diploid population in case of violation in $\boldsymbol{\mu}$ distribution with $\boldsymbol{\epsilon} = 0.5$ are tabulated in table 4.6.

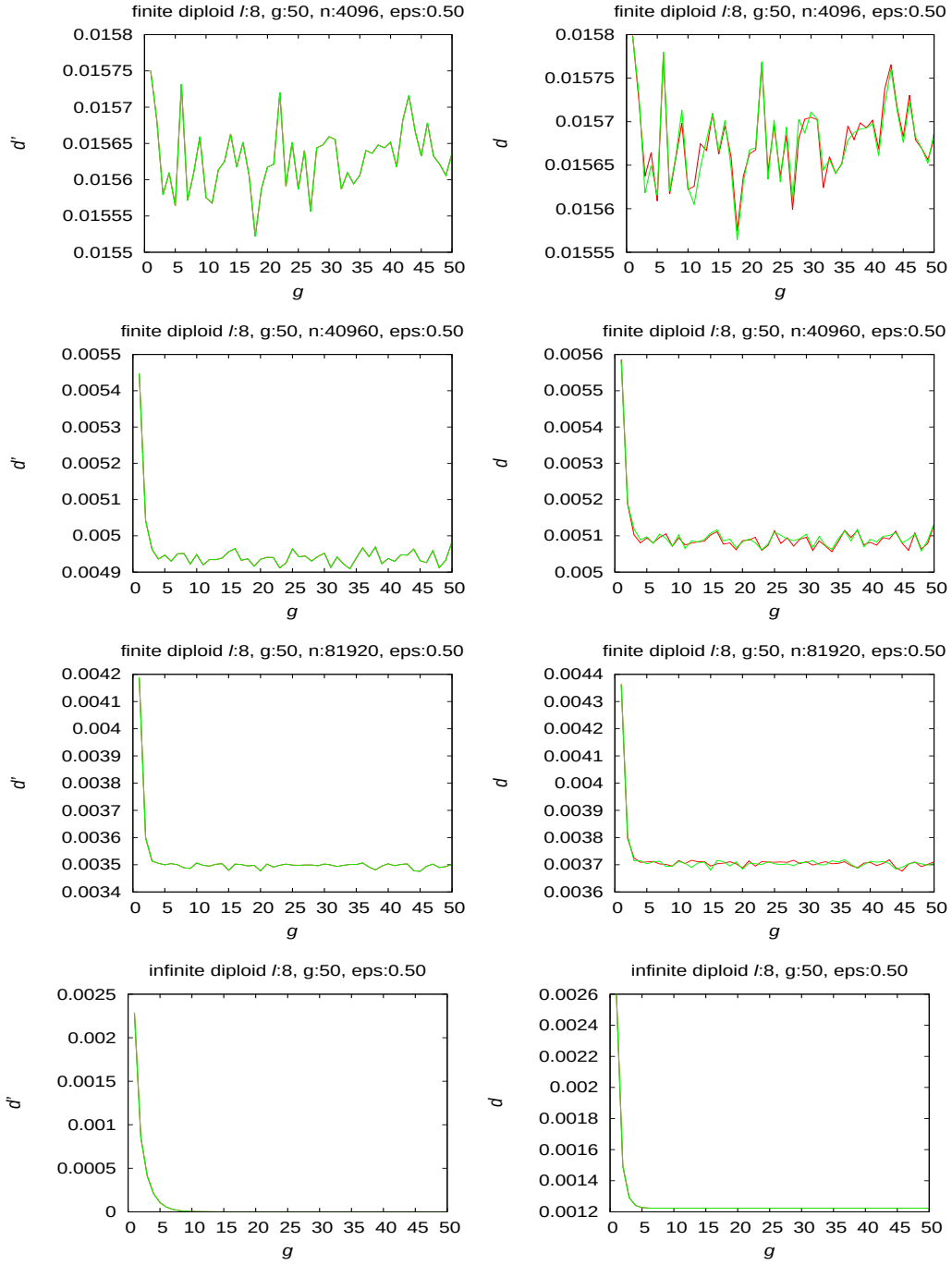


Figure 4.21: Infinite and finite diploid population behavior for μ violation, $\ell = 8$ and $\epsilon = 0.5$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

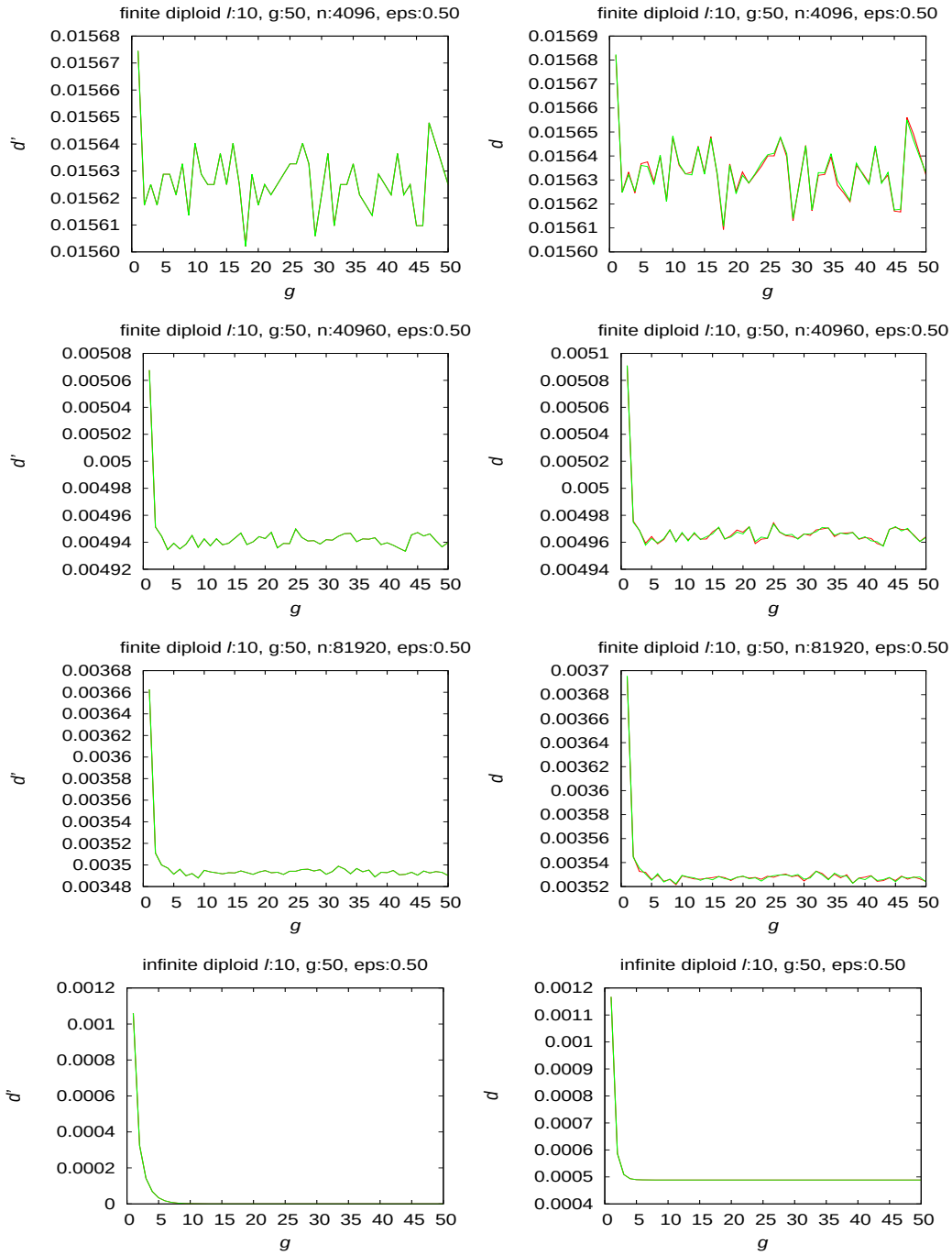


Figure 4.22: Infinite and finite diploid population behavior for μ violation, genome length $\ell = 10$ and $\epsilon = 0.5$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

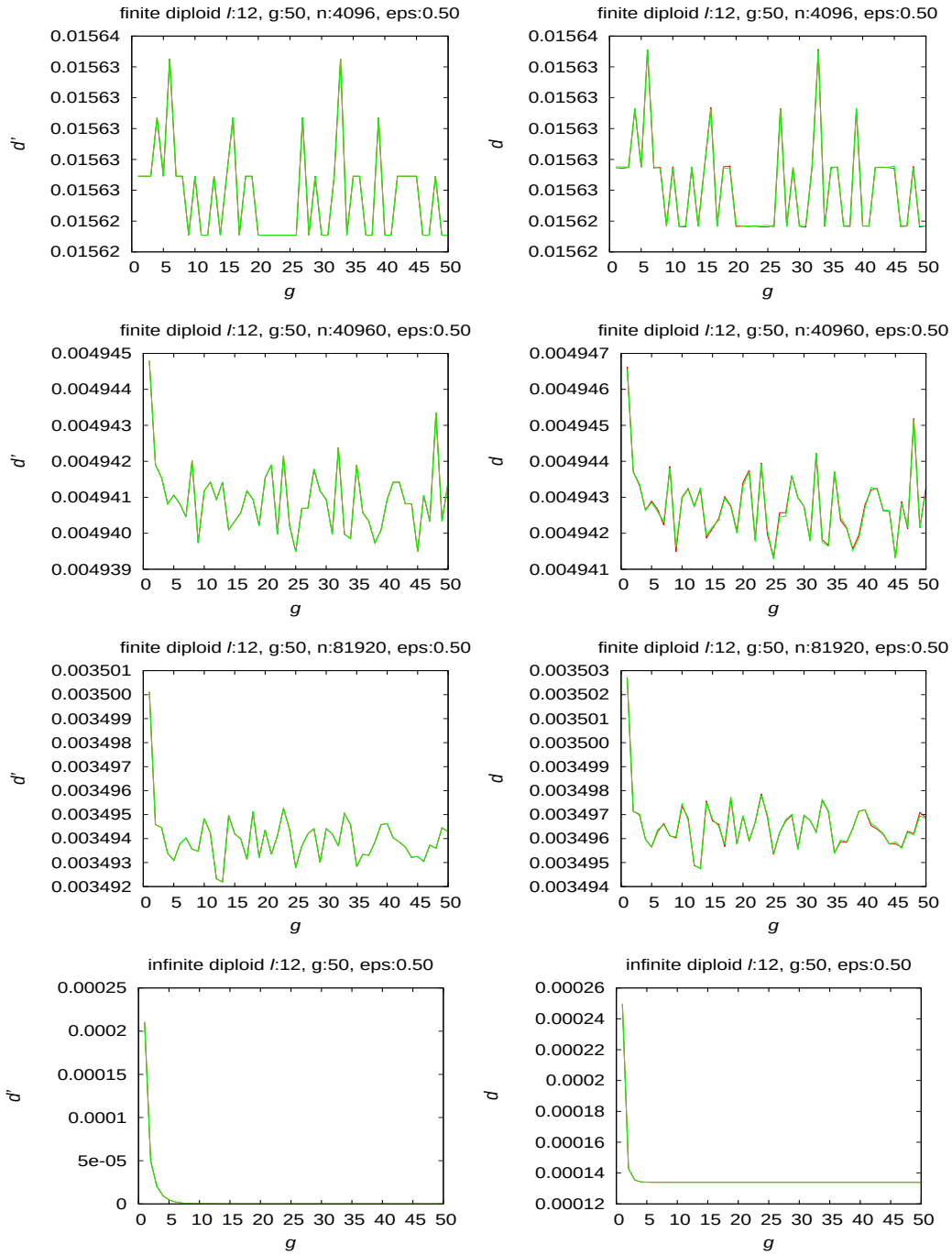


Figure 4.23: Infinite and finite diploid population behavior for μ violation, genome length $\ell = 12$ and $\epsilon = 0.5$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

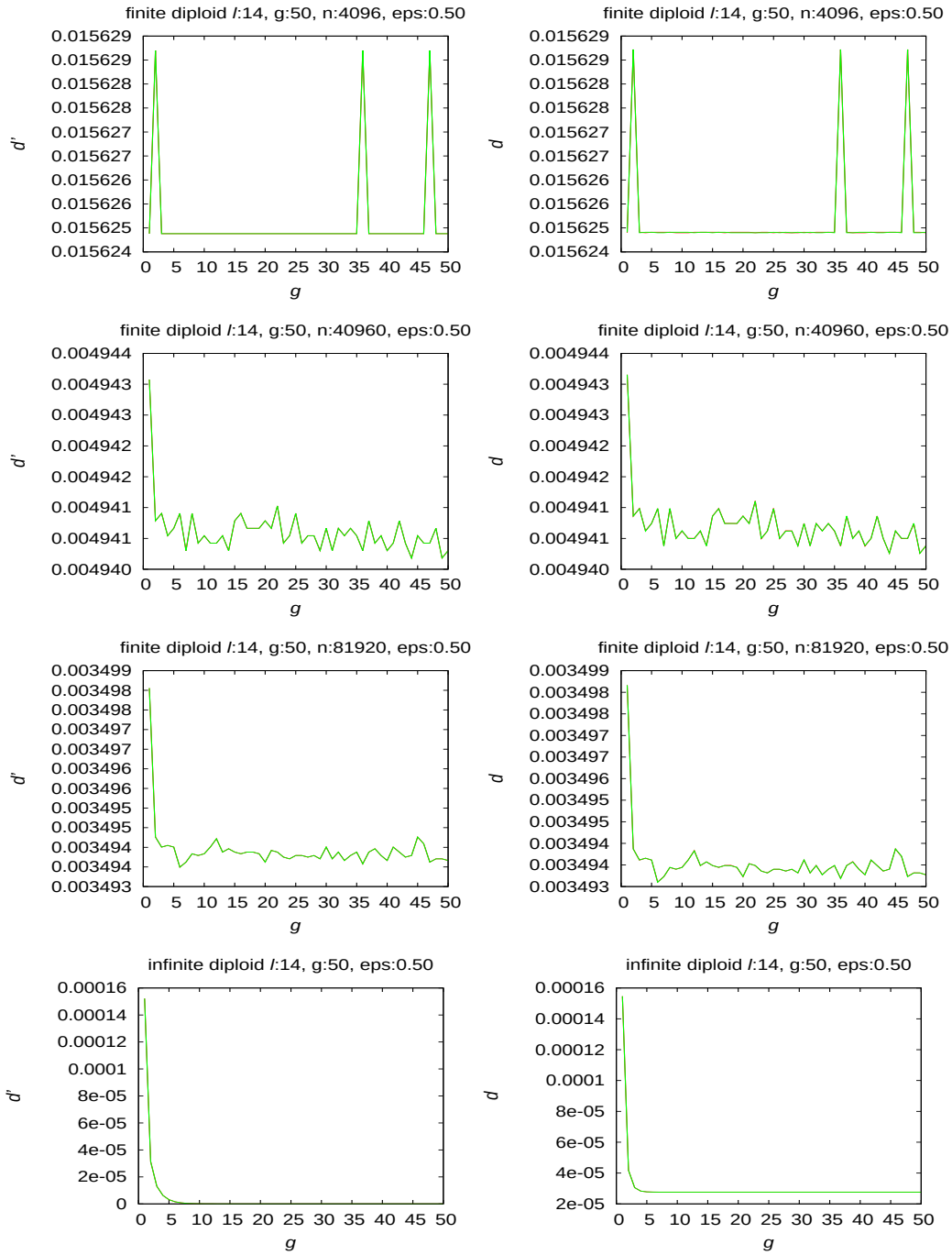


Figure 4.24: Infinite and finite diploid population behavior for μ violation, genome length $\ell = 14$ and $\epsilon = 0.5$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

Table 4.6: Distance measured for violation in $\boldsymbol{\mu}$ with $\boldsymbol{\epsilon} = 0.5$ for diploids: ℓ is genome length, average distance between finite and infinite population is tabulated in the last three columns, and last row is expected single step distance.

ℓ	$N = 4096$	$N = 40960$	$N = 81920$
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
$1/\sqrt{N}$	0.0156	0.0049	0.0035

Table 4.6 shows that the average distance between finite population and infinite population agrees with the expected single step distance $1/\sqrt{N}$.

4.2 Discussion

The previous graphs indicate that as value of ℓ increases, amplitude of oscillation decreases, and randomness in oscillation increases. Populations with larger population size show better oscillations. Since a diploid population has an effective string length twice the string length of a hapliod, diploid populations need larger population size to exhibit good oscillation. For diploid populations, increasing string length ℓ degrades convergence (as population size increases) to infinite population behavior. That is noticeable in figures 4.13 through 4.20 for violation in $\boldsymbol{\mu}$. Such behavior is less noticeable in haploid populations.

With increasing $\boldsymbol{\epsilon}$, oscillation diminishes. As observed in chapter 3, diploid populations hop to various levels (in figures 4.15, 4.16, 4.19, 4.20, 4.23 and 4.24), and such behavior is absent for large population sizes.

Figure 4.25 summarizes the distance data from tables 4.1 through 4.6. Distance data (between finite and infinite populations) are plotted for different ℓ. Plots for different violation levels $\boldsymbol{\epsilon}$ are arranged in columns. Plots for haploid and diploid populations are arranged in two rows. With increase in ℓ, distance moves closer to

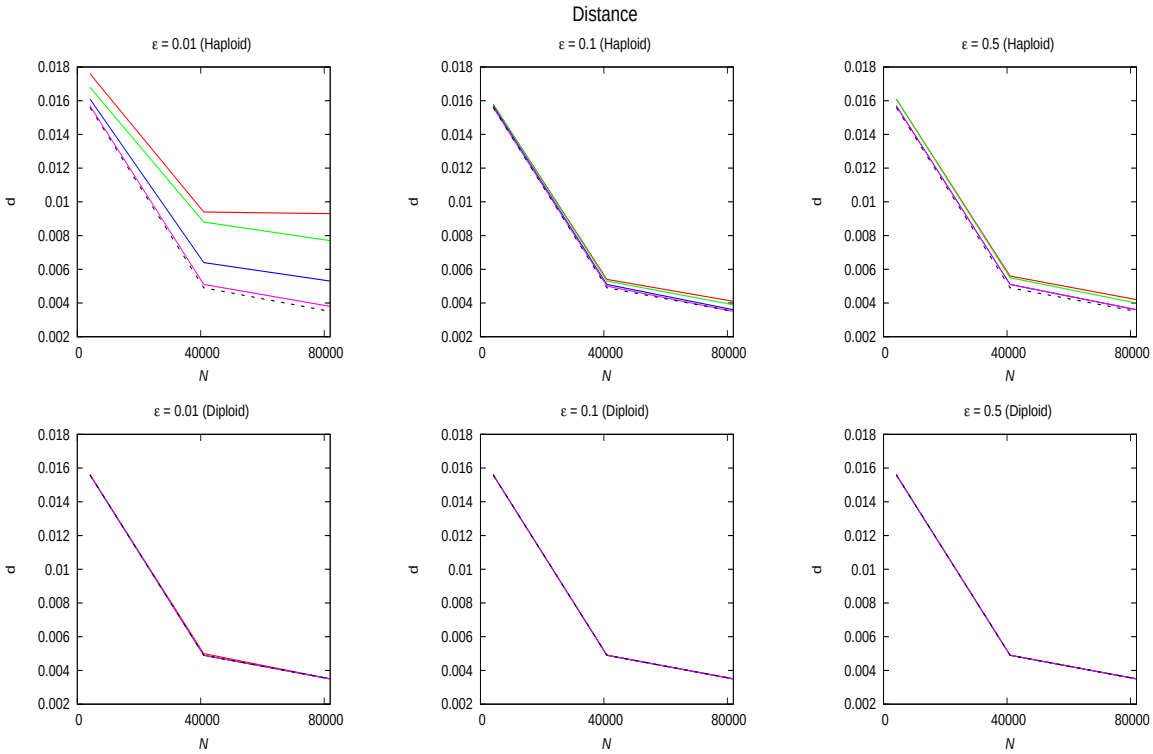


Figure 4.25: Distance between finite and infinite population in case of violation in μ: d is distance; N is finite population size; ϵ is level of violation; red line represents distance for $\ell = 8$, green line for $\ell = 10$, blue line for $\ell = 12$, pink line for $\ell = 14$ and black dotted line for expected single step distance.

the single step distance. Since diploid effective string length is twice haploid string length, distance in diploid case moves closer to the single step distance than in haploid case. It is also noticeable that in the haploid case, the distance moves closer to the single step distance as ϵ increases.

4.3 Summary

In this chapter, we violated the condition 3.3 for mutation, making the Markov chain representing finite population evolution regular, and ensuring that infinite population trajectories have no periodic orbit. Our experiments show that finite population evolution continues to approximately oscillate for small values of ϵ. For such values

of ϵ, finite population evolution might sometimes be unaware of violation in condition 3.3 because the probability of using the new mask (all 0s mask) is low, and if the new mask is not used, finite population behavior matches the behavior exhibited without the violation. As population size increases, better oscillations are observed. As string length increases, oscillation degrades and larger population sizes are required to observe good oscillation.

Chapter 5

Violation in Crossover Distribution

The results from chapter 4 show robustness of finite population oscillation demonstrating approximate oscillation can take place in finite populations when the mutation distribution $\boldsymbol{\mu}$ violates condition 3.3 . This chapter explores the robustness of finite population oscillation when condition 3.3 for the crossover distribution $\boldsymbol{\chi}$ is violated. Violation of the condition, crossover-violation, as we call it, is expressed as:

$$\text{For all } g \text{ , } g \neq 0, \qquad 1 \neq \sum_{k \in \bar{g}\mathcal{R}} \boldsymbol{\chi}_{k+g} + \boldsymbol{\chi}_k \tag{5.1}$$

The question explored in this chapter is: Can finite populations exhibit approximate oscillation when there is violation in $\boldsymbol{\chi}$ and infinite population trajectories have no periodic orbit?

Error $\boldsymbol{\epsilon}$ is introduced into the crossover distribution $\boldsymbol{\chi}$ so as to violate condition 3.3; this guarantees that infinite population trajectories have no periodic orbit. Consequently, $\boldsymbol{p}^* = \boldsymbol{q}^* = \boldsymbol{z}^*$. Going forward, we use 'limit $\boldsymbol{z}^*$' to denote evolutionary limit when crossover distribution $\boldsymbol{\chi}$ violates condition 3.3, and 'non-violation limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$' to denote limits without violations (i.e., $\boldsymbol{\epsilon} = \mathbf{0}$).

5.1 Violation

The crossover distribution $\boldsymbol{\chi}$ was modified as

$$\boldsymbol{\chi}_i = (1-\boldsymbol{\epsilon})\boldsymbol{\chi}_i \qquad \text{so that} \qquad \sum \boldsymbol{\chi}_i + \boldsymbol{\chi}_{i+g} = 1 - \boldsymbol{\epsilon}$$

Then a single j is chosen where $j \notin \bar{g}\mathcal{R}$ and set $\boldsymbol{\chi}_j = \boldsymbol{\epsilon}$.

Violation in crossover distribution $\boldsymbol{\chi}$ is different from violation in mutation distribution $\boldsymbol{\mu}$. The Markov chain formed by transition matrix Q is regular under violation in $\boldsymbol{\mu}$ but that need not be the case under violation in $\boldsymbol{\chi}$. The initial population is computed using the same procedure as described in section 3.3. To explore the effects of the degree of violation of condition 3.3 in $\boldsymbol{\chi}$, different values of $\boldsymbol{\epsilon} \in \{0.01, 0.1, 0.5\}$ are used in experiments. String length $\ell \in \{8, 10, 12, 14\}$ is considered for simulation. The distances of both infinite and finite populations to limit $\boldsymbol{z}^*$ are plotted. The distances of both infinite and finite populations to non-violation limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$ (i.e. $\boldsymbol{\epsilon} = 0$) are also plotted.

5.1.1 Haploid Population $\sim \epsilon : 0.01$

The right column in figures 5.1 through 5.4 shows distance of finite and infinite haploid populations with $\boldsymbol{\epsilon} = 0.01$ to non-violation limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Since the value of $\boldsymbol{\epsilon}$ is small, damping of ripples is slow. A new mask with probability $\boldsymbol{\epsilon} = 0.01$ has small probability of being used during crossover and when not used, behavior matches behavior without violation. Moreover, $\boldsymbol{\epsilon} = 0.01$ is small enough that infinite population oscillation persists over 50 generations even though it will die out eventually.

The left column of figures 5.1 through 5.4 shows distance of finite and infinite haploid populations with $\boldsymbol{\epsilon} = 0.01$ to limit $\boldsymbol{z}^*$. The distance decreases as population size increases, and finite population shows behavior similar to infinite population as population size grows.

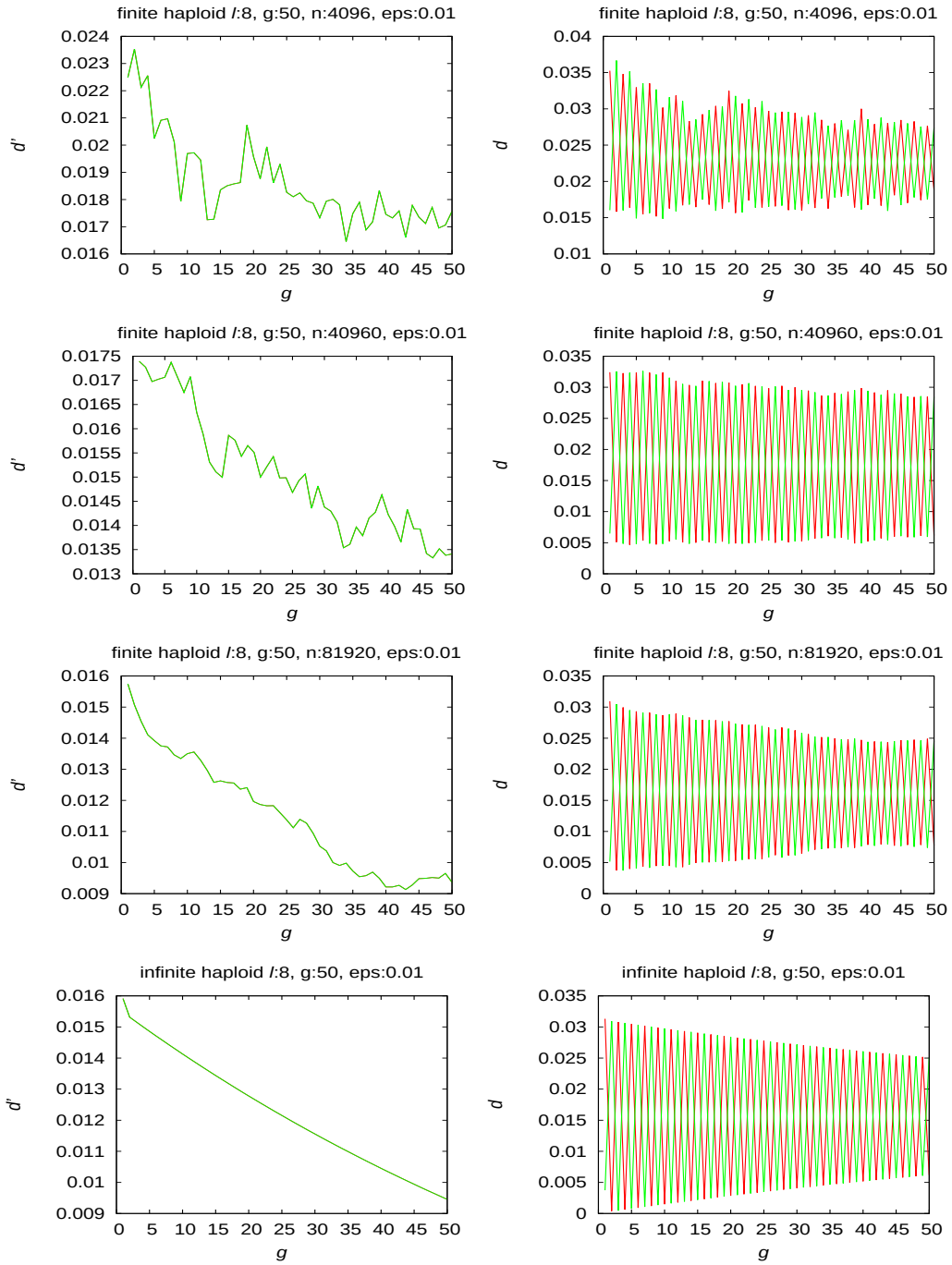


Figure 5.1: Infinite and finite haploid population behavior for χ violation, $\ell = 8$ and $\epsilon = 0.01$: In left column, d' is distance of finite population or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite population or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

Figure 5.2: Infinite and finite haploid population behavior for χ violation, genome length $\ell = 10$ and $\epsilon = 0.01$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite population or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

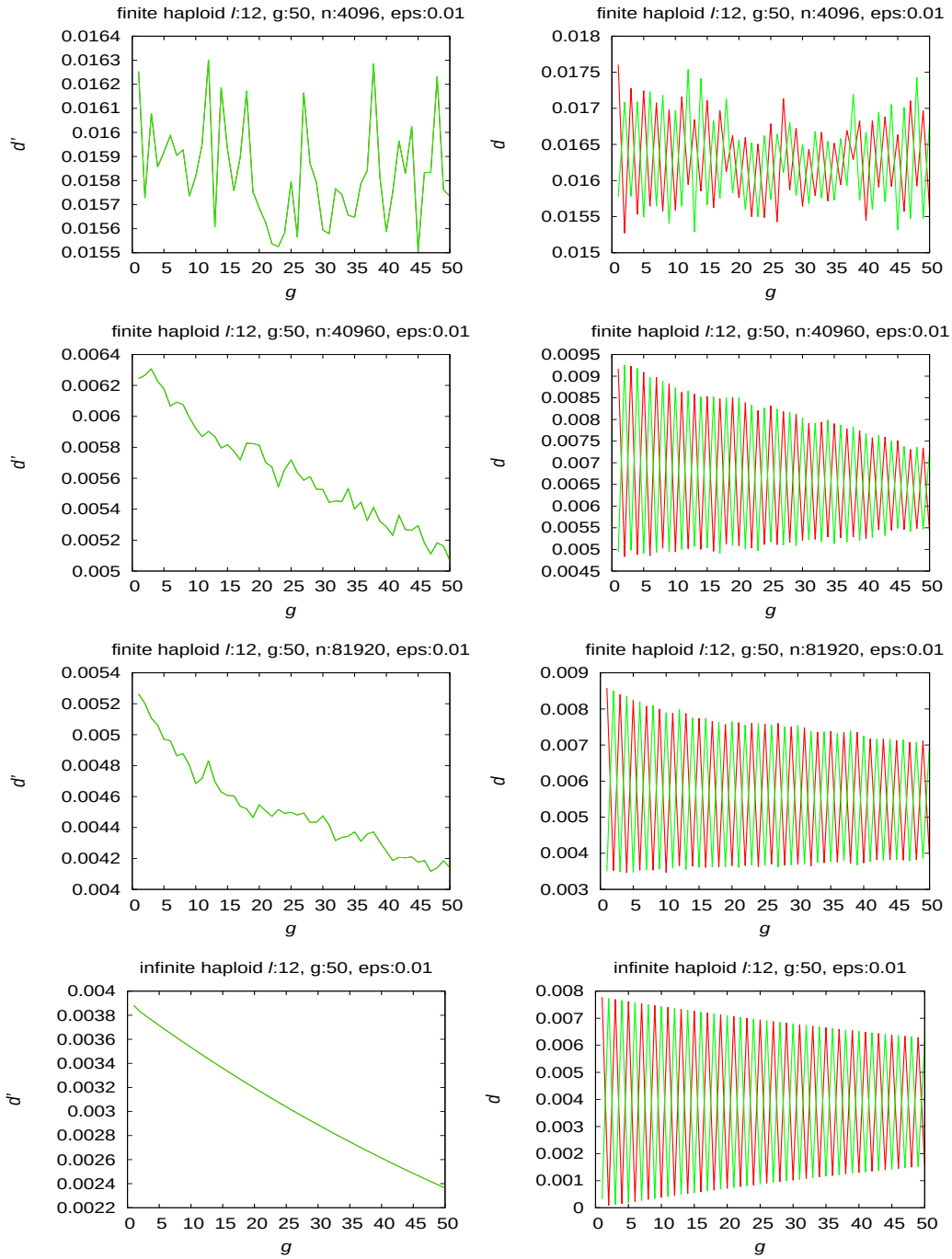


Figure 5.3: Infinite and finite haploid population behavior for χ violation, genome length $\ell = 12$ and $\epsilon = 0.01$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite population or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

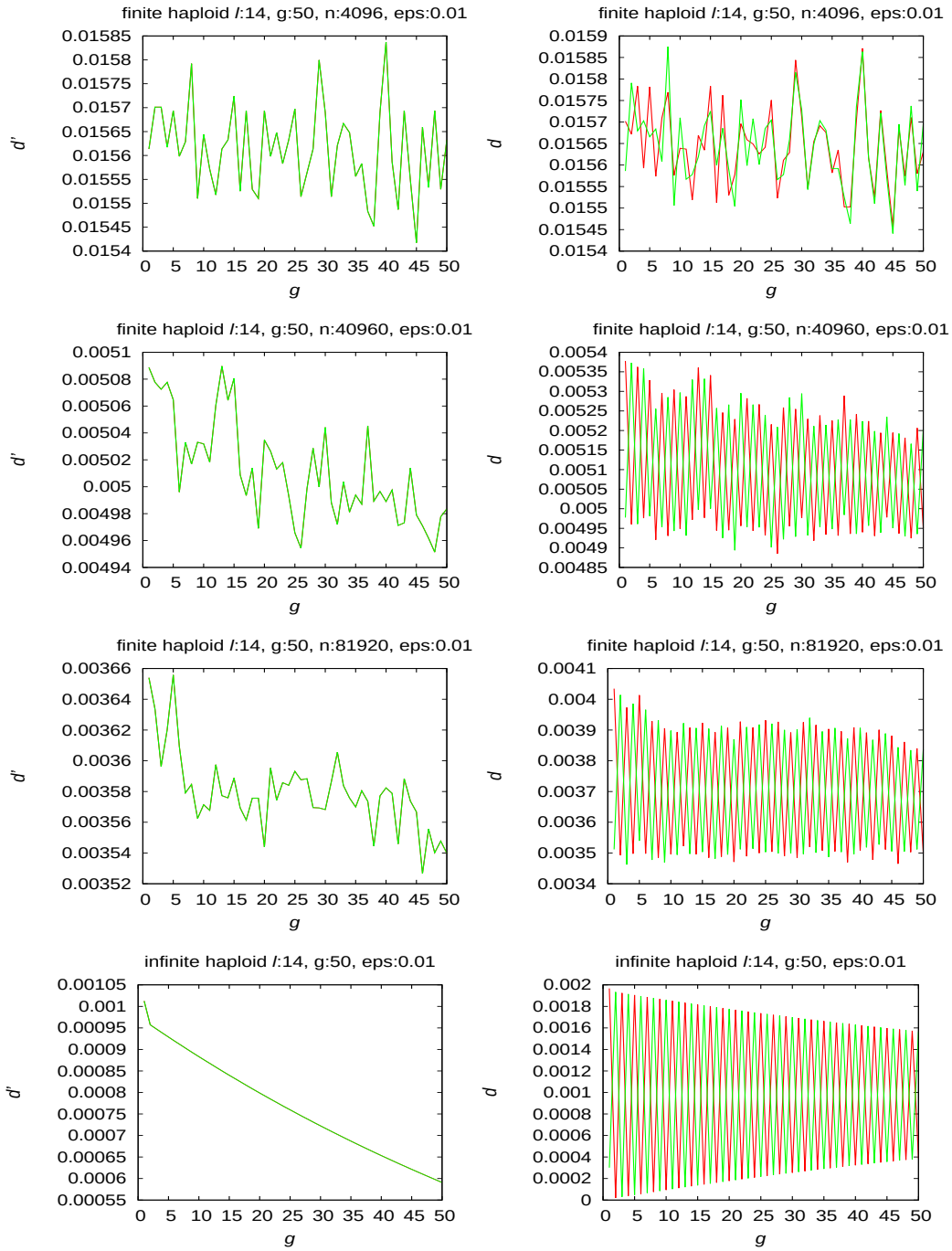


Figure 5.4: Infinite and finite haploid population behavior for χ violation, genome length $\ell = 14$ and $\epsilon = 0.01$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite population or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

Average distance data for haploid population in case of violation in $\boldsymbol{\chi}$ distribution with $\boldsymbol{\epsilon} = 0.01$ are tabulated in table 5.1.

Table 5.1: Distance measured for violation in $\boldsymbol{\chi}$ with $\boldsymbol{\epsilon} = 0.01$ for haploids: ℓ is genome length, average distance between finite and infinite population is tabulated in the last three columns, and last row is expected single step distance.

ℓ	$N = 4096$	$N = 40960$	$N = 81920$
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
$1/\sqrt{N}$	0.0156	0.0049	0.0035

Table 5.1 shows that the average distance between finite and infinite population decreases with increasing string length, approaching the expected single step distance $1/\sqrt{N}$.

5.1.2 Haploid Population $\sim \epsilon : 0.1$

The right column in figures 5.5 through 5.8 shows distance of finite and infinite haploid populations with $\boldsymbol{\epsilon} = 0.1$ to non-violation limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Those graphs indicate oscillating behavior which decreases with time. For $\boldsymbol{\epsilon} = 0.1$, infinite population oscillation dies out quickly, but oscillation in finite population does not. Rate of damping of ripples with $\boldsymbol{\epsilon} = 0.1$ is larger than with $\boldsymbol{\epsilon} = 0.01$. The new mask has probability $\boldsymbol{\epsilon} = 0.1$ of being used during crossover which is too small to significantly disrupt oscillation in those finite populations considered here.

The left column of figures 5.5 through 5.8 shows distance of finite and infinite haploid populations with $\boldsymbol{\epsilon} = 0.1$ to limit $\boldsymbol{z}^*$. The distance decreases as population size increases, and finite population behavior is similar to infinite population as population size grows. Average distance data for haploid population in case of violation in $\boldsymbol{\chi}$ distribution with $\boldsymbol{\epsilon} = 0.1$ are tabulated in table 5.2.

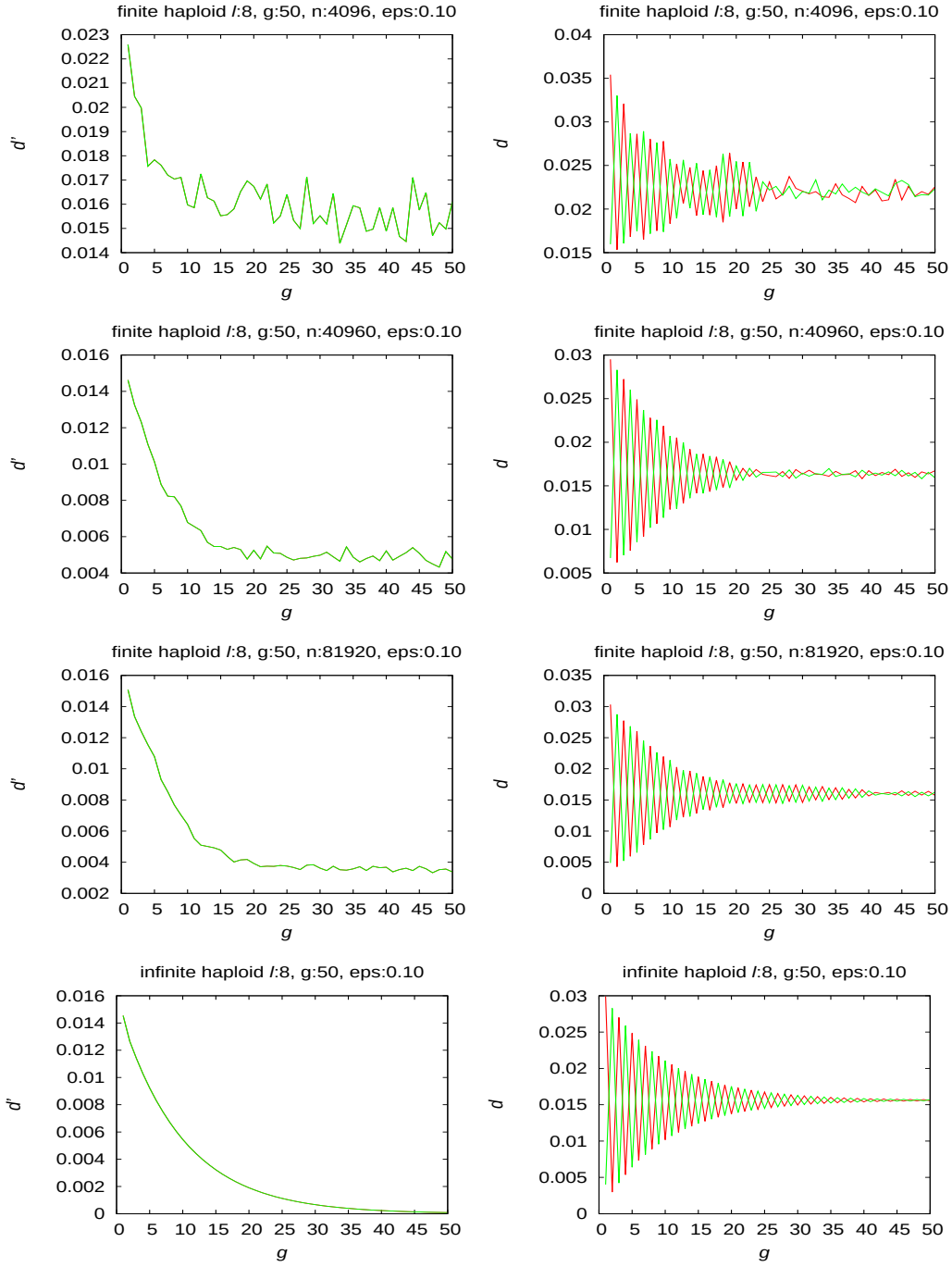


Figure 5.5: Infinite and finite haploid population behavior for χ violation, $\ell = 8$ and $\epsilon = 0.1$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

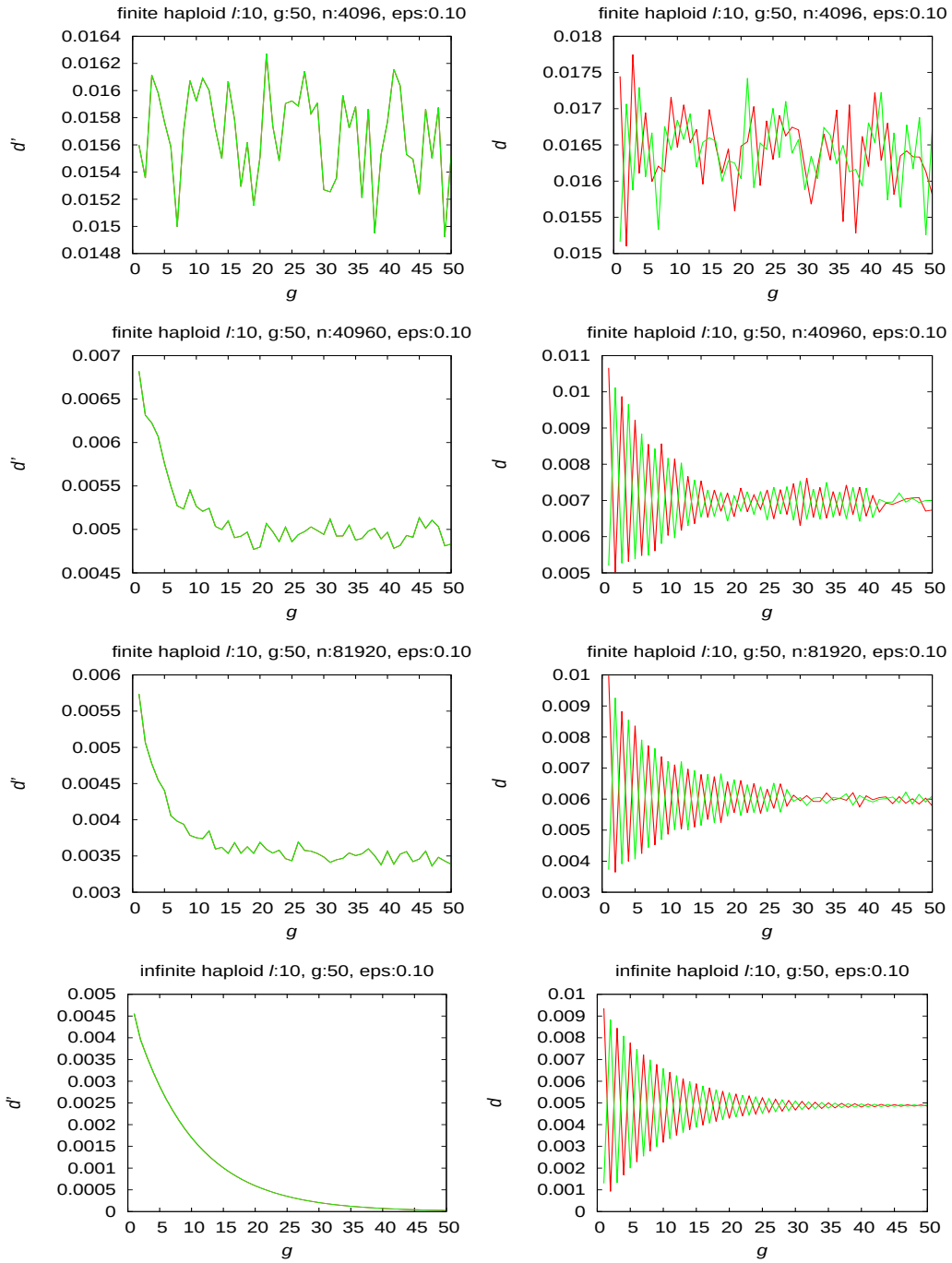


Figure 5.6: Infinite and finite haploid population behavior for χ violation, genome length $\ell = 10$ and $\epsilon = 0.1$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

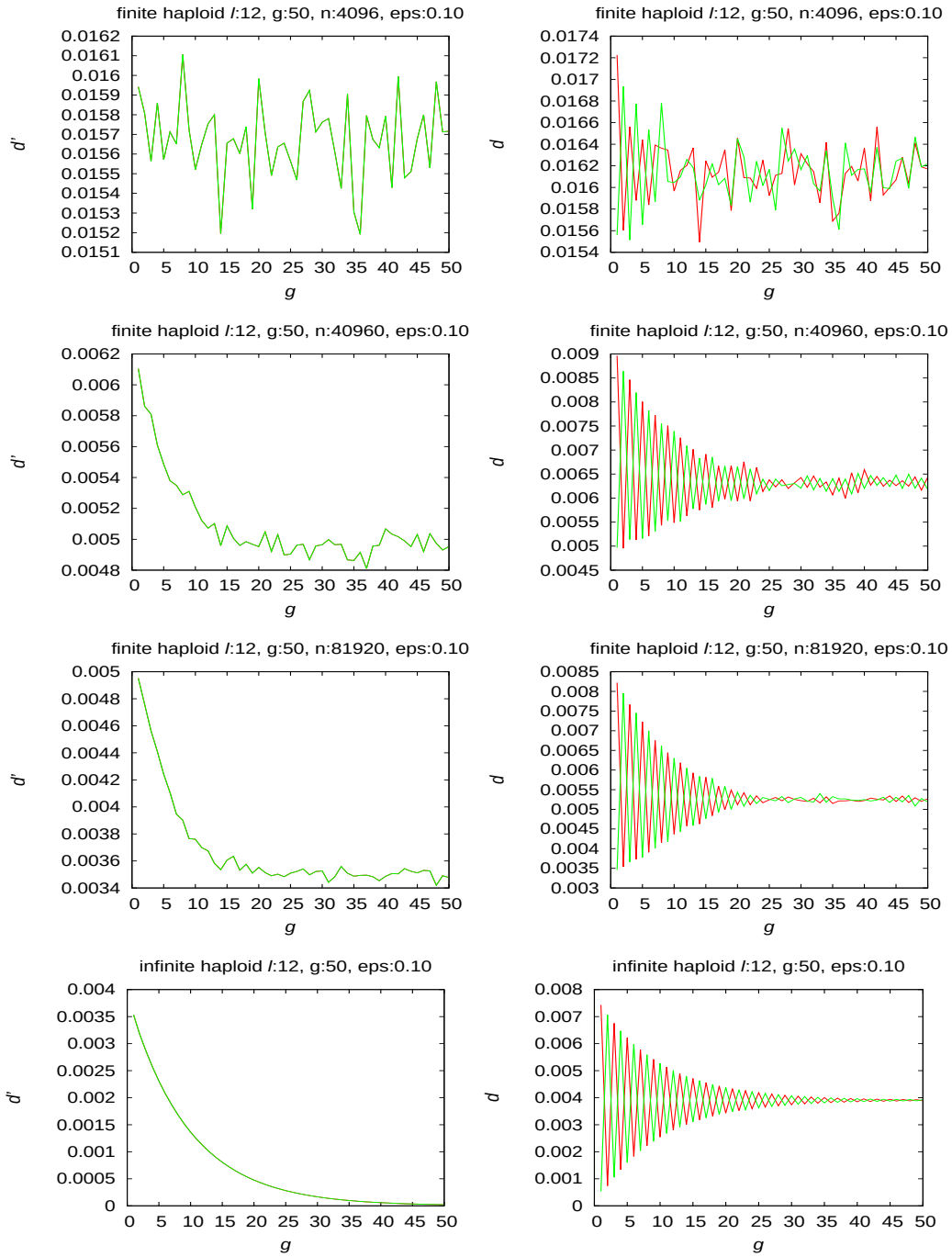


Figure 5.7: Infinite and finite haploid population behavior for χ violation, genome length $\ell = 12$ and $\epsilon = 0.1$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

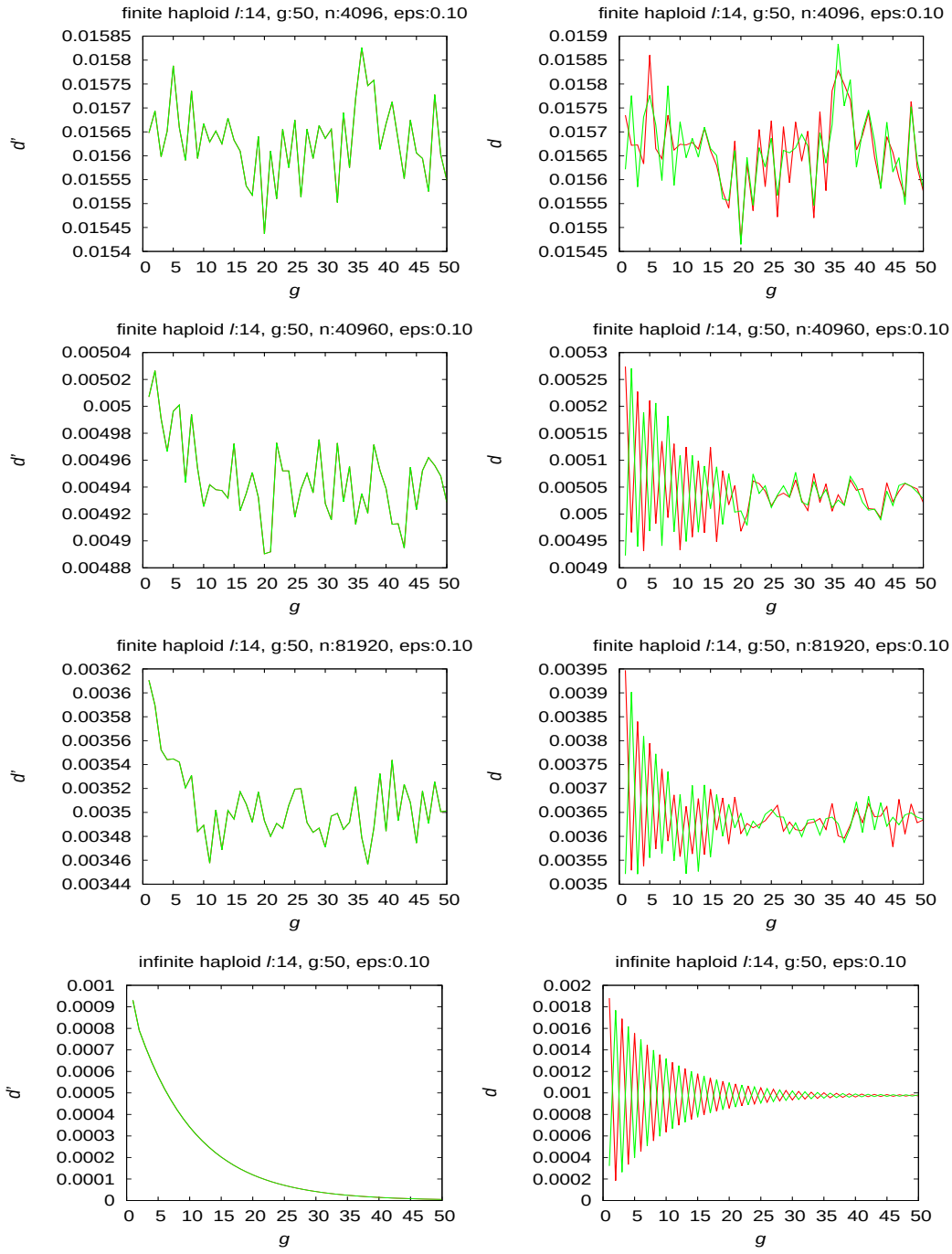


Figure 5.8: Infinite and finite haploid population behavior for χ violation, $\ell = 14$ and $\epsilon = 0.1$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

Table 5.2: Distance measured for violation in χ with $\epsilon = 0.1$ for haploids: ℓ is genome length, average distance between finite and infinite population is tabulated in the last three columns, and last row is expected single step distance.

ℓ	$N = 4096$	$N = 40960$	$N = 81920$
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
$1/\sqrt{N}$	0.0156	0.0049	0.0035

Table 5.2 shows that the average distance between finite and infinite populations decreases with increasing string length approaching the expected single step distance $1/\sqrt{N}$.

5.1.3 Haploid Population $\sim \epsilon : 0.5$

The right column in figures 5.9 through 5.12 shows distance of finite and infinite haploid populations with $\boldsymbol{\epsilon} = 0.5$ to non-violation limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Compared to mutation with violation $\boldsymbol{\epsilon} = 0.5$, oscillation is observed for more generations. Finite populations still show some, though not very clear, oscillations, and then show randomness in behavior as generations progress. The oscillation in infinite population dies out quickly. Randomness in finite population behavior increases compared to smaller values of $\boldsymbol{\epsilon}$, especially as ℓ increases.

The left column of figures 5.9 through 5.12 shows distance of finite and infinite haploid populations with $\boldsymbol{\epsilon} = 0.5$ to limit $\boldsymbol{z}^*$ (limit with violation in crossover distribution $\boldsymbol{\chi}$). The distance decreases as population size increases, and finite population shows behavior similar to infinite population behavior as finite population size grows. Average distance data for haploid population in case of violation in $\boldsymbol{\chi}$ distribution with $\boldsymbol{\epsilon} = 0.5$ for different finite population size N are tabulated in table 5.3.

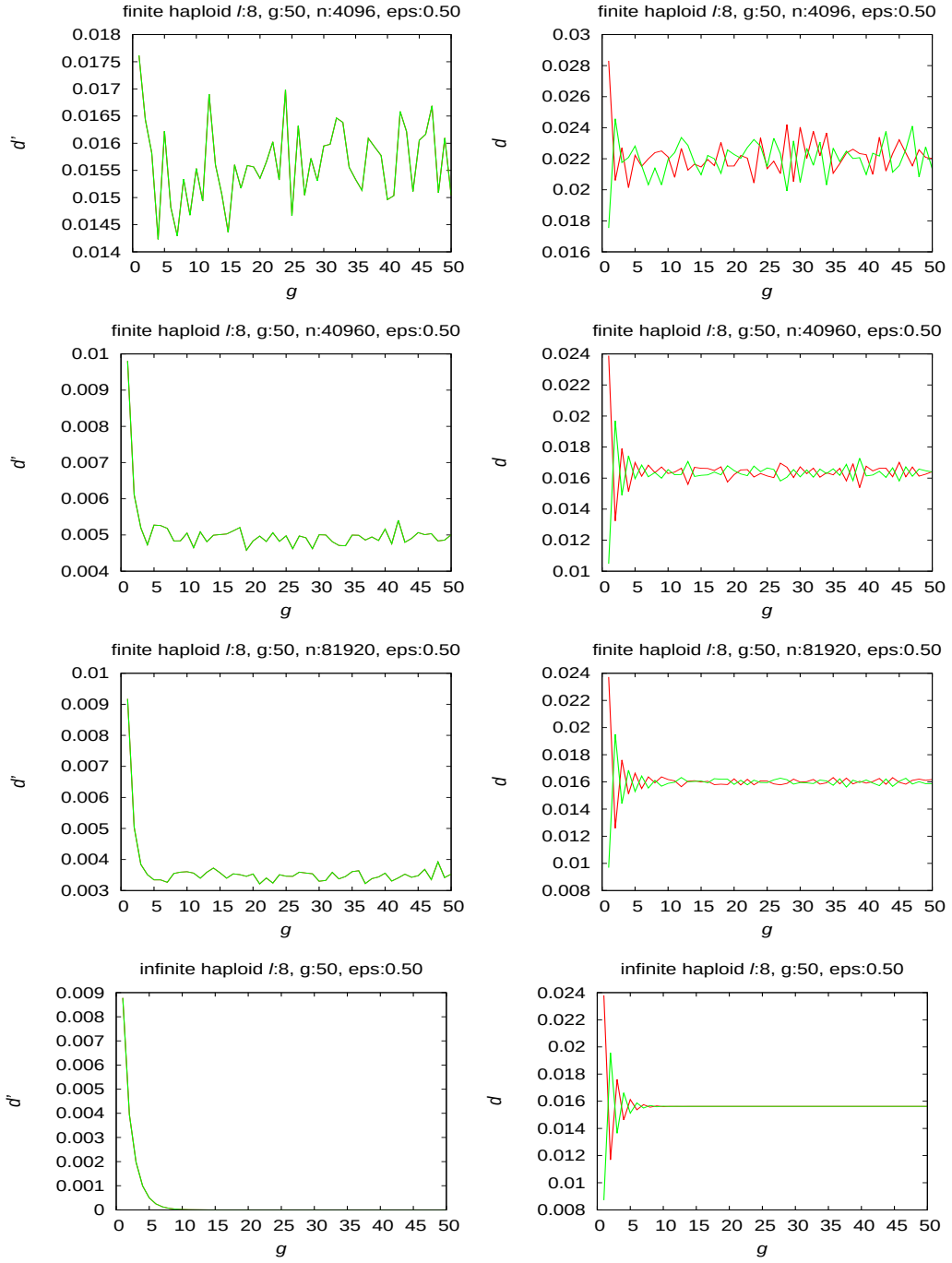


Figure 5.9: Infinite and finite haploid population behavior for χ violation, $\ell = 8$ and $\epsilon = 0.5$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

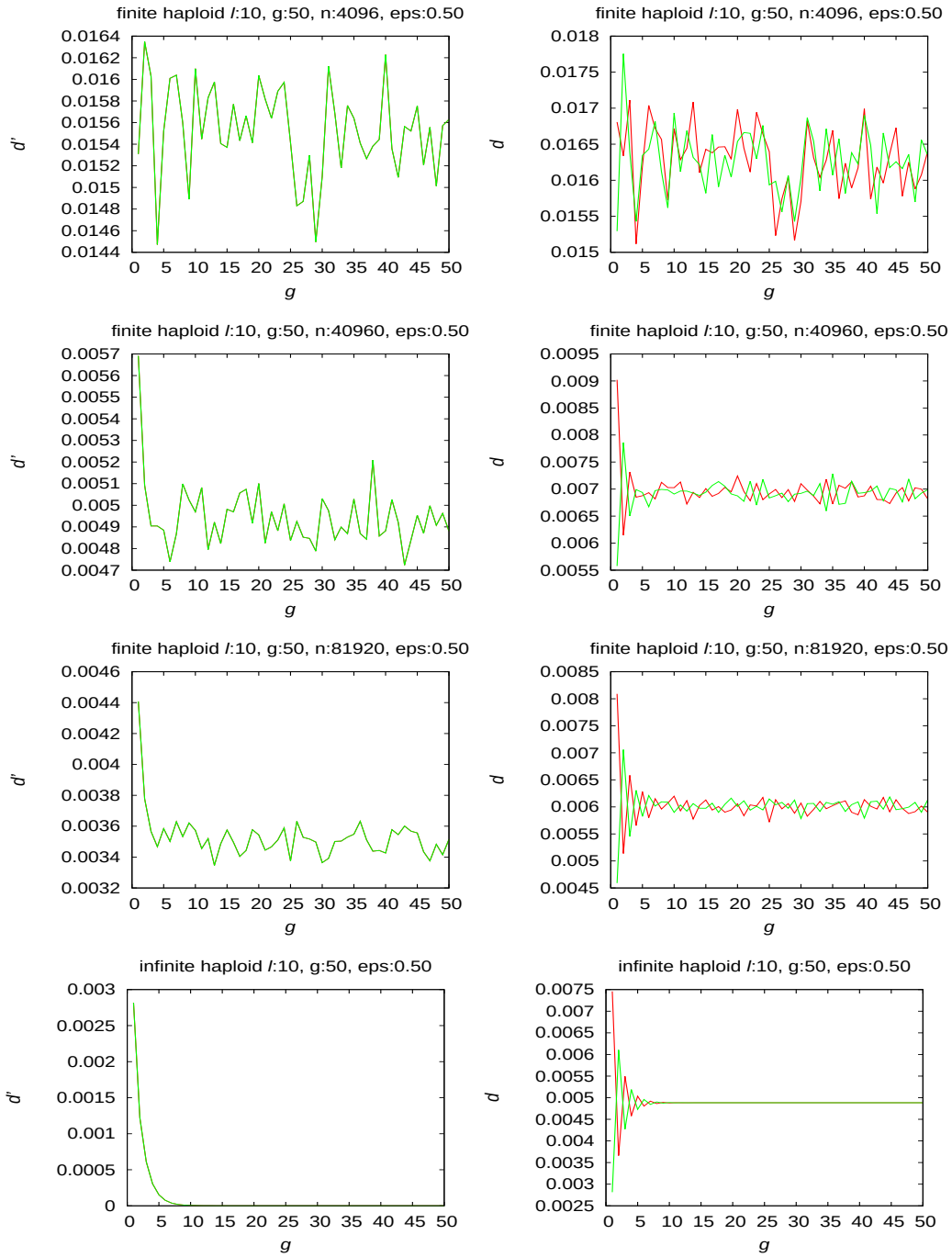


Figure 5.10: Infinite and finite haploid population behavior for χ violation, genome length $\ell = 10$ and $\epsilon = 0.5$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

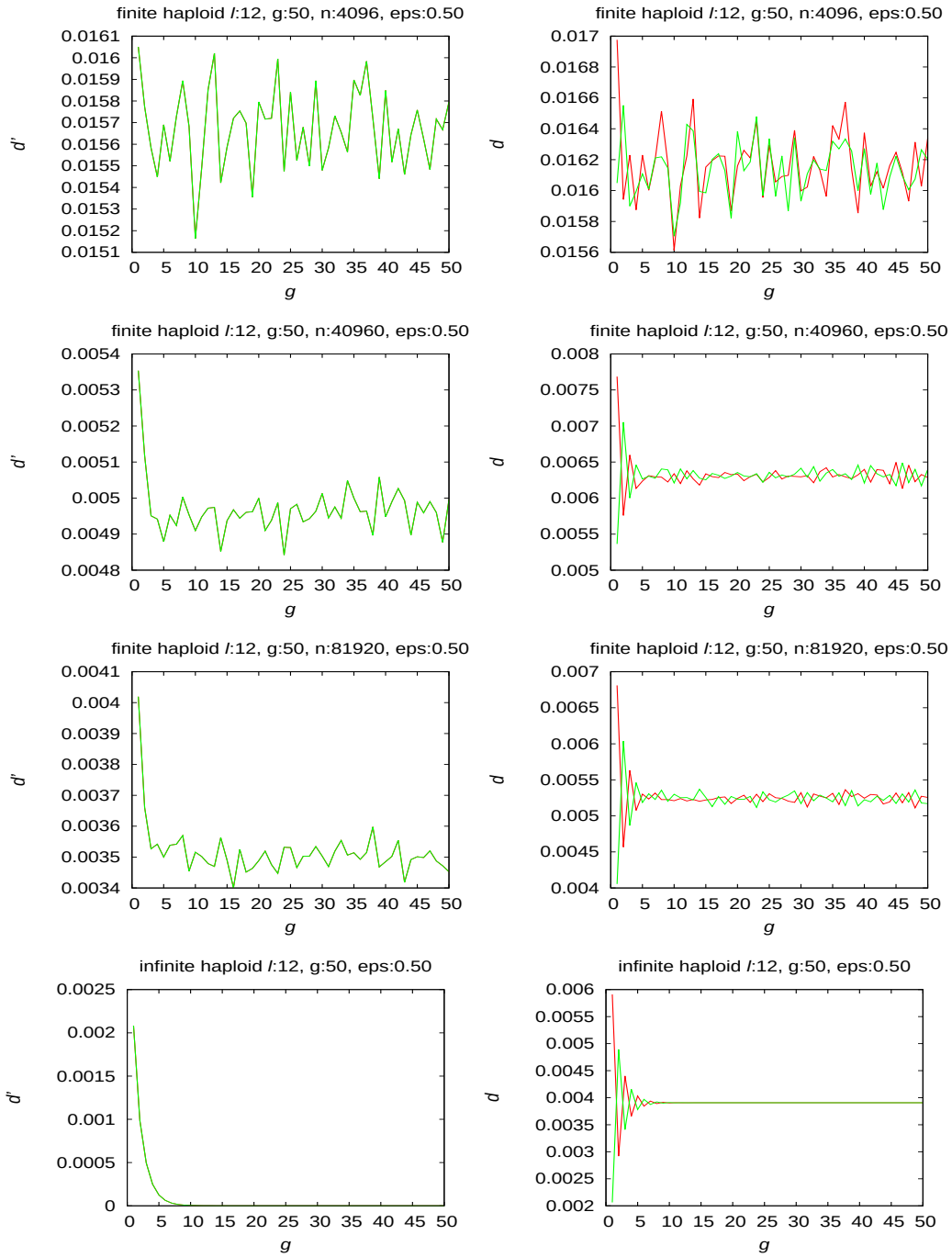


Figure 5.11: Infinite and finite haploid population behavior for χ violation, genome length $\ell = 12$ and $\epsilon = 0.5$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

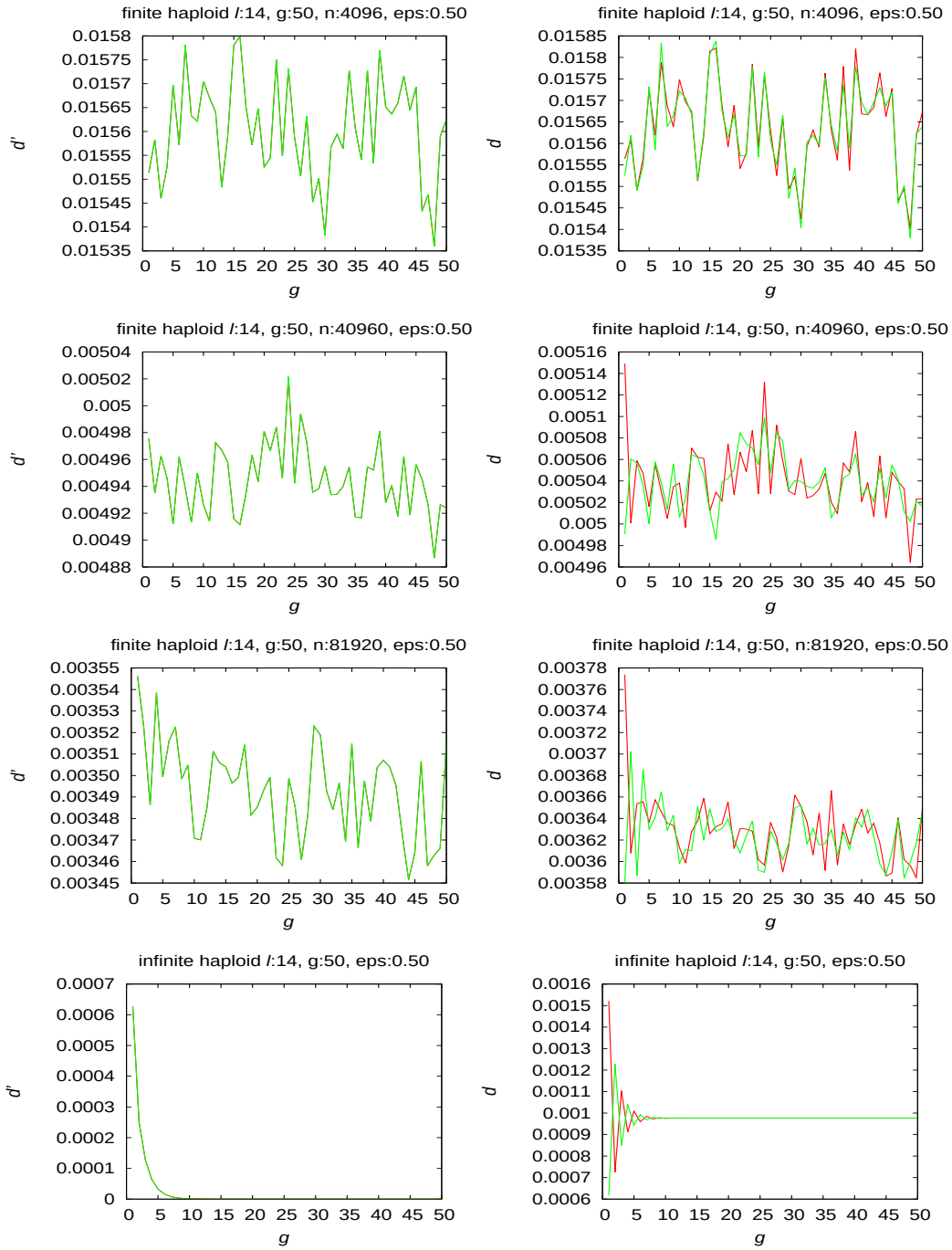


Figure 5.12: Infinite and finite haploid population behavior for χ violation, genome length $\ell = 14$ and $\epsilon = 0.5$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

Table 5.3: Distance measured for violation in $\boldsymbol{\chi}$ with $\boldsymbol{\epsilon} = 0.5$ for haploids: ℓ is genome length, average distance between finite and infinite population is tabulated in the last three columns, and last row is expected single step distance.

ℓ	$N = 4096$	$N = 40960$	$N = 81920$
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
$1/\sqrt{N}$	0.0156	0.0049	0.0035

Table 5.3 shows that the average distance between finite and infinite populations approaches the expected single step distance $1/\sqrt{N}$.

5.1.4 Diploid Population $\sim \epsilon : 0.01$

The right column in figures 5.13 through 5.16 shows distance of finite and infinite diploid populations with $\boldsymbol{\epsilon} = 0.01$ to non-violation limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Since $\boldsymbol{\epsilon}$ is small, damping of ripples is slow. Infinite population oscillation does not die out in 50 generations even though it dies out eventually. Finite population graphs show randomness, and oscillation improves with increased population size. That can be noticed more clearly in figures 5.15 and 5.16.

The left column of figures 5.13 through 5.16 shows distance of finite and infinite diploid populations with $\boldsymbol{\epsilon} = 0.01$ to limit $\boldsymbol{z}^*$ (limit with violation in crossover distribution $\boldsymbol{\chi}$). The distance decreases as population size increases. Average distance data for diploid population in case of violation in $\boldsymbol{\chi}$ distribution with $\boldsymbol{\epsilon} = 0.01$ for different finite population size N are tabulated in table 5.4.

Table 5.4 shows that the average distance between finite and infinite population approaches the expected single step distance $1/\sqrt{N}$.

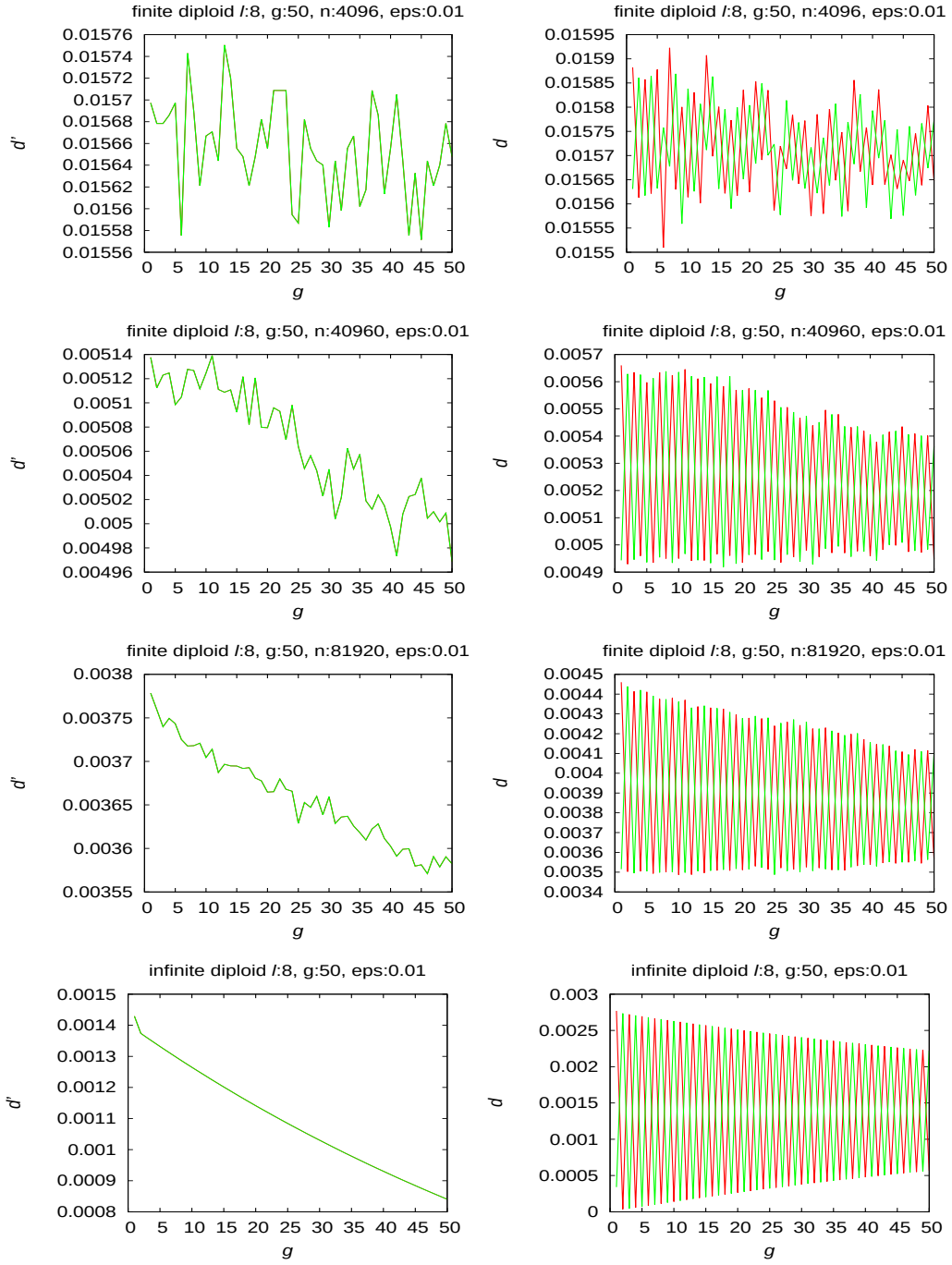


Figure 5.13: Infinite and finite diploid population behavior for χ violation, $\ell = 8$ and $\epsilon = 0.01$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

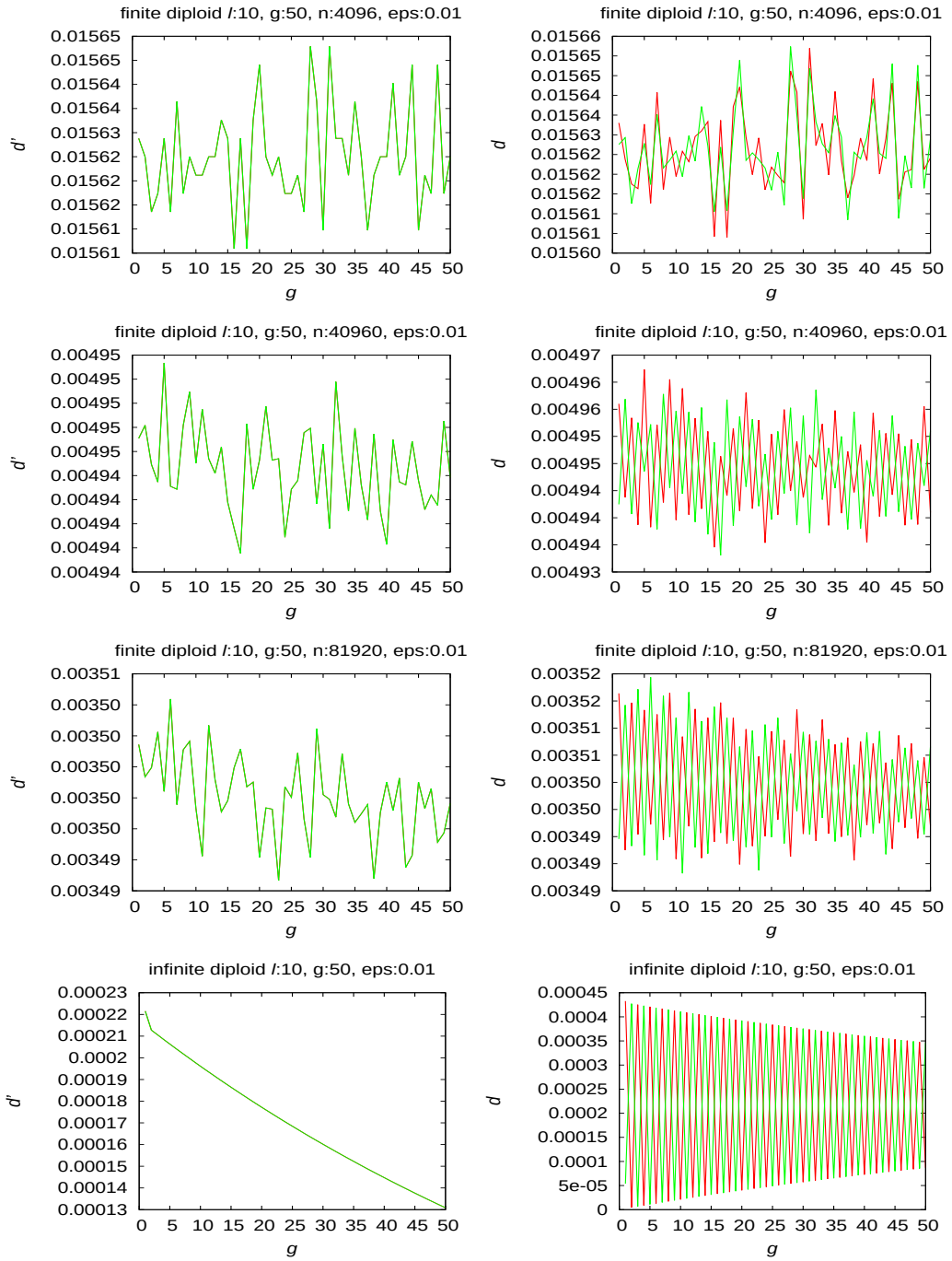


Figure 5.14: Infinite and finite diploid population behavior for χ violation, genome length $\ell = 10$ and $\epsilon = 0.01$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

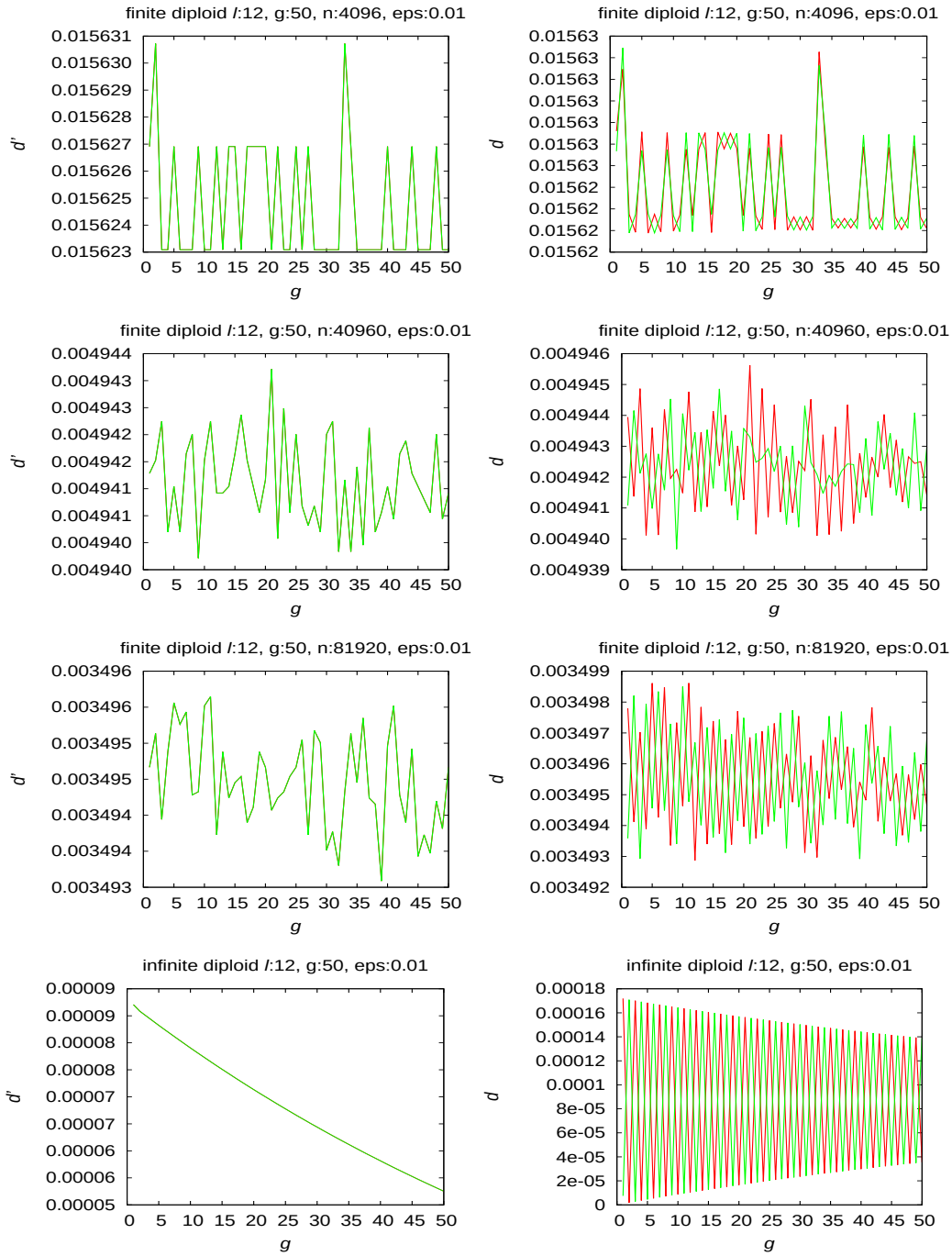


Figure 5.15: Infinite and finite diploid population behavior for χ violation, genome length $\ell = 12$ and $\epsilon = 0.01$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

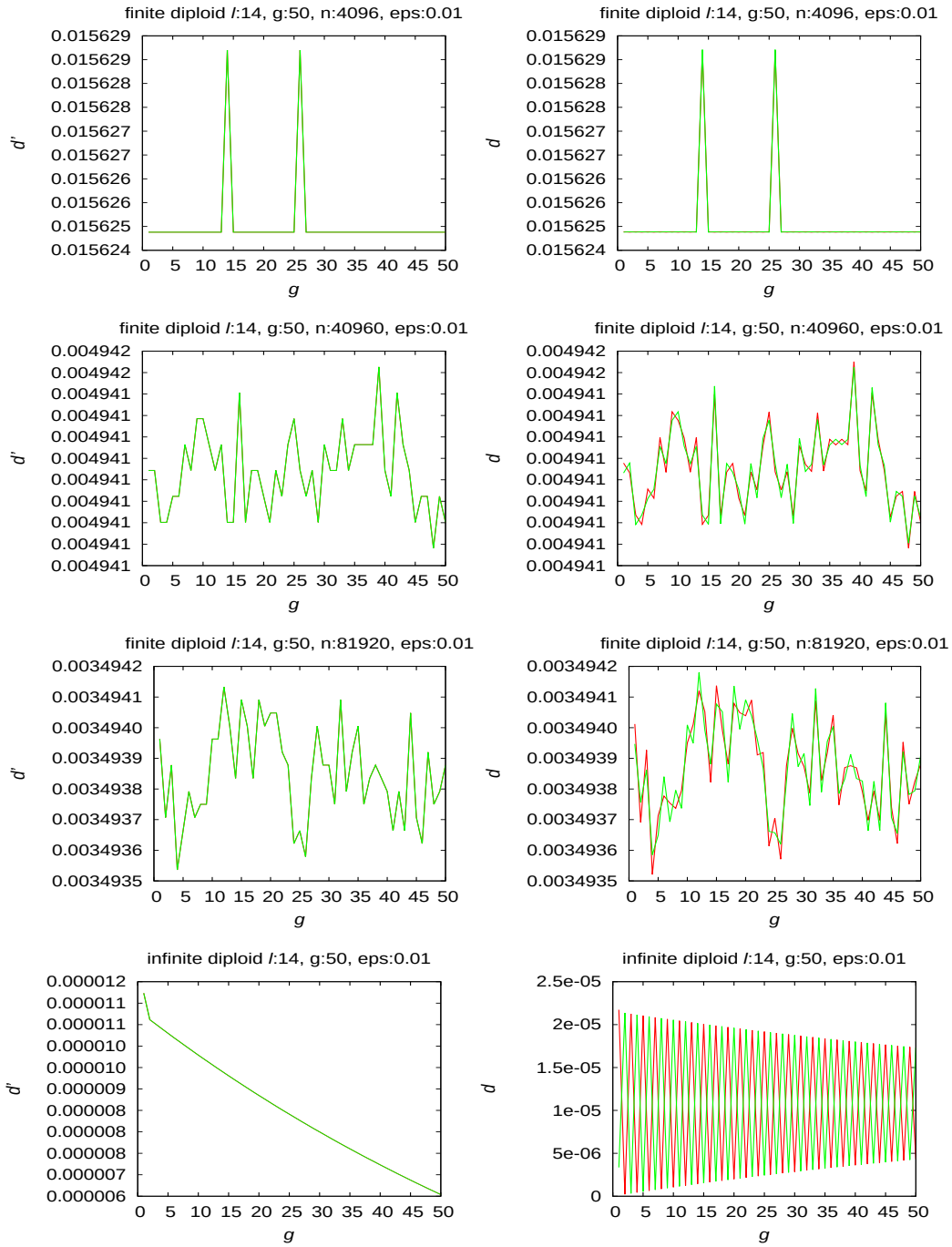


Figure 5.16: Infinite and finite diploid population behavior for χ violation, genome length $\ell = 14$ and $\epsilon = 0.01$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

Table 5.4: Distance measured for violation in $\boldsymbol{\chi}$ with $\boldsymbol{\epsilon} = 0.01$ diploids: ℓ is genome length, average distance between finite and infinite population is tabulated in the last three columns, and last row is expected single step distance.

ℓ	$N = 4096$	$N = 40960$	$N = 81920$
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
$1/\sqrt{N}$	0.0156	0.0049	0.0035

5.1.5 Diploid Population $\sim \epsilon : 0.1$

The right column in figures 5.17 through 5.20 shows distance of finite and infinite diploid populations with $\boldsymbol{\epsilon} = 0.1$ to non-violation limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Those graphs indicate oscillation amplitude decreases with increasing generations. Like in the haploid case, oscillations in infinite populations die out quickly for $\boldsymbol{\epsilon} = 0.1$. Rate of damping with $\boldsymbol{\epsilon} = 0.1$ is higher than with $\boldsymbol{\epsilon} = 0.01$. The probability $\boldsymbol{\epsilon} = 0.1$ of the new crossover mask being used is too small to significantly disrupt oscillation in those finite populations considered here. The graphs exhibit more randomness than in case of $\boldsymbol{\epsilon} = 0.01$, and as value of ℓ increases, randomness increases more for smaller population size.

The left column of figures 5.17 through 5.20 shows distance of finite and infinite diploid populations with $\boldsymbol{\epsilon} = 0.1$ to limit $\boldsymbol{z}^*$. The distance decreases as population size increases. Average distance data for diploid population in case of violation in $\boldsymbol{\chi}$ distribution with $\boldsymbol{\epsilon} = 0.1$ are tabulated in table 5.5.

Table 5.5 shows that the average distance between finite and infinite populations approaches the expected single step distance $1/\sqrt{N}$.

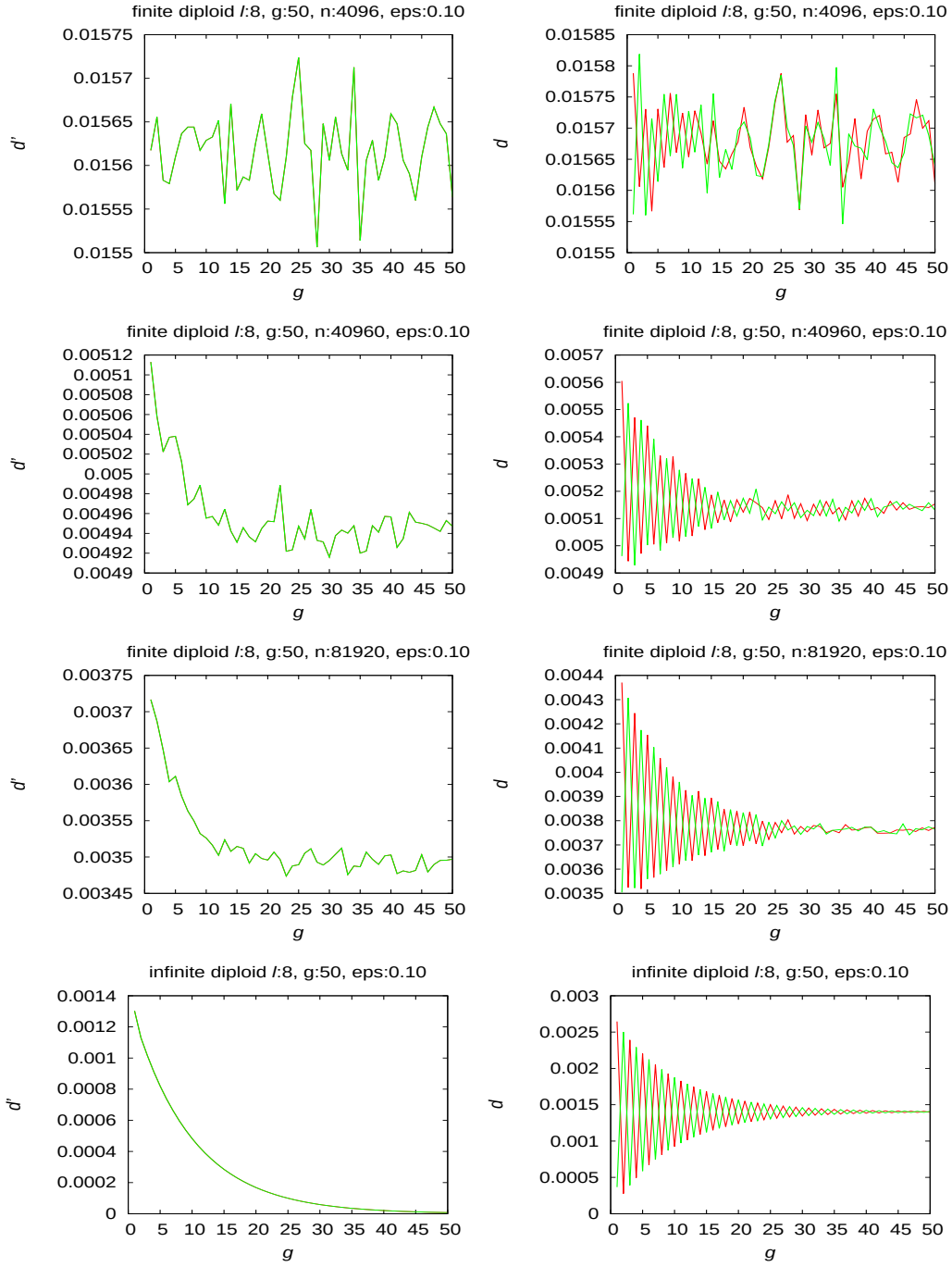


Figure 5.17: Infinite and finite diploid population behavior for χ violation, $\ell = 8$ and $\epsilon = 0.1$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

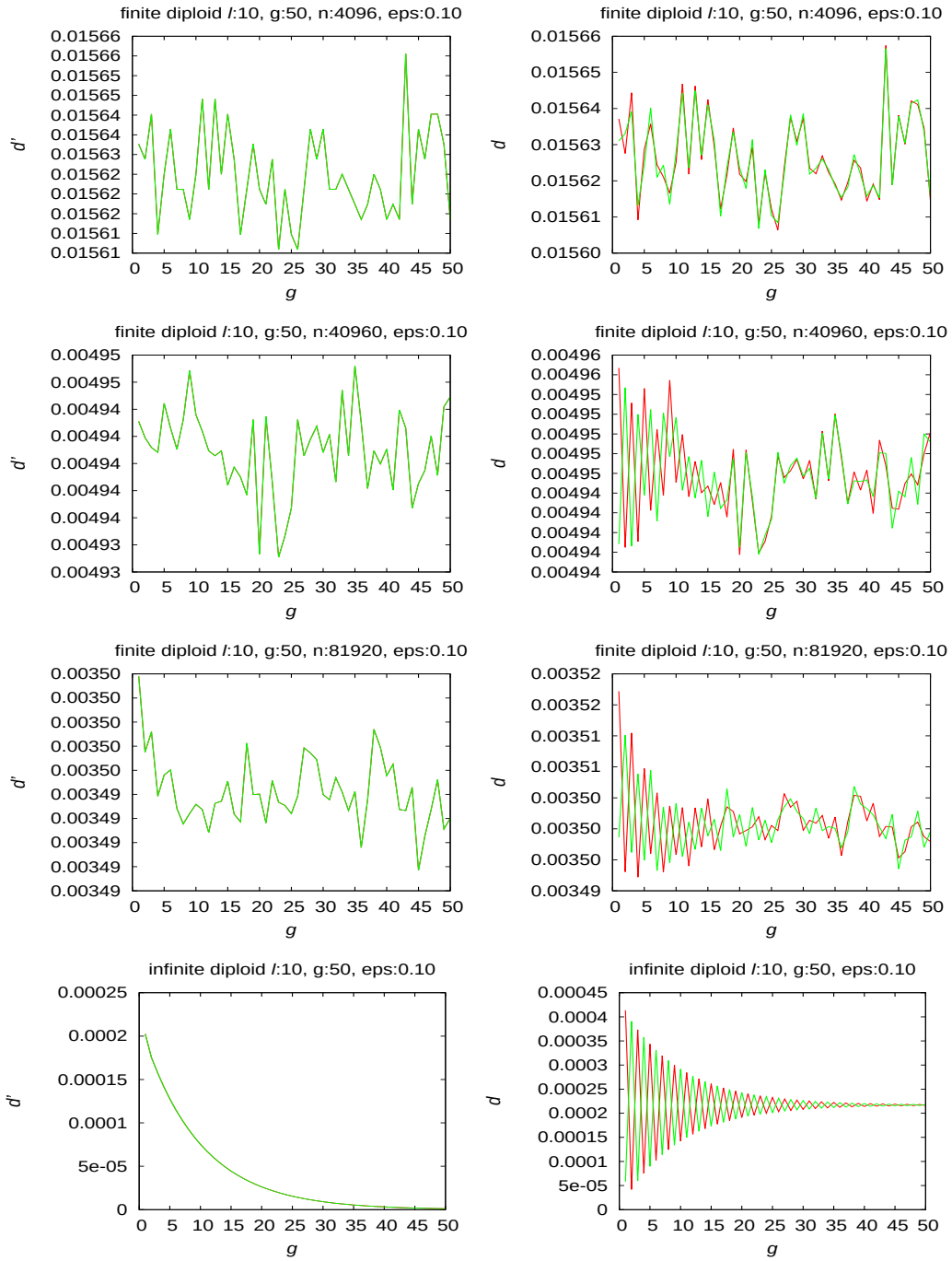


Figure 5.18: Infinite and finite diploid population behavior for χ violation, genome length $\ell = 10$ and $\epsilon = 0.1$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

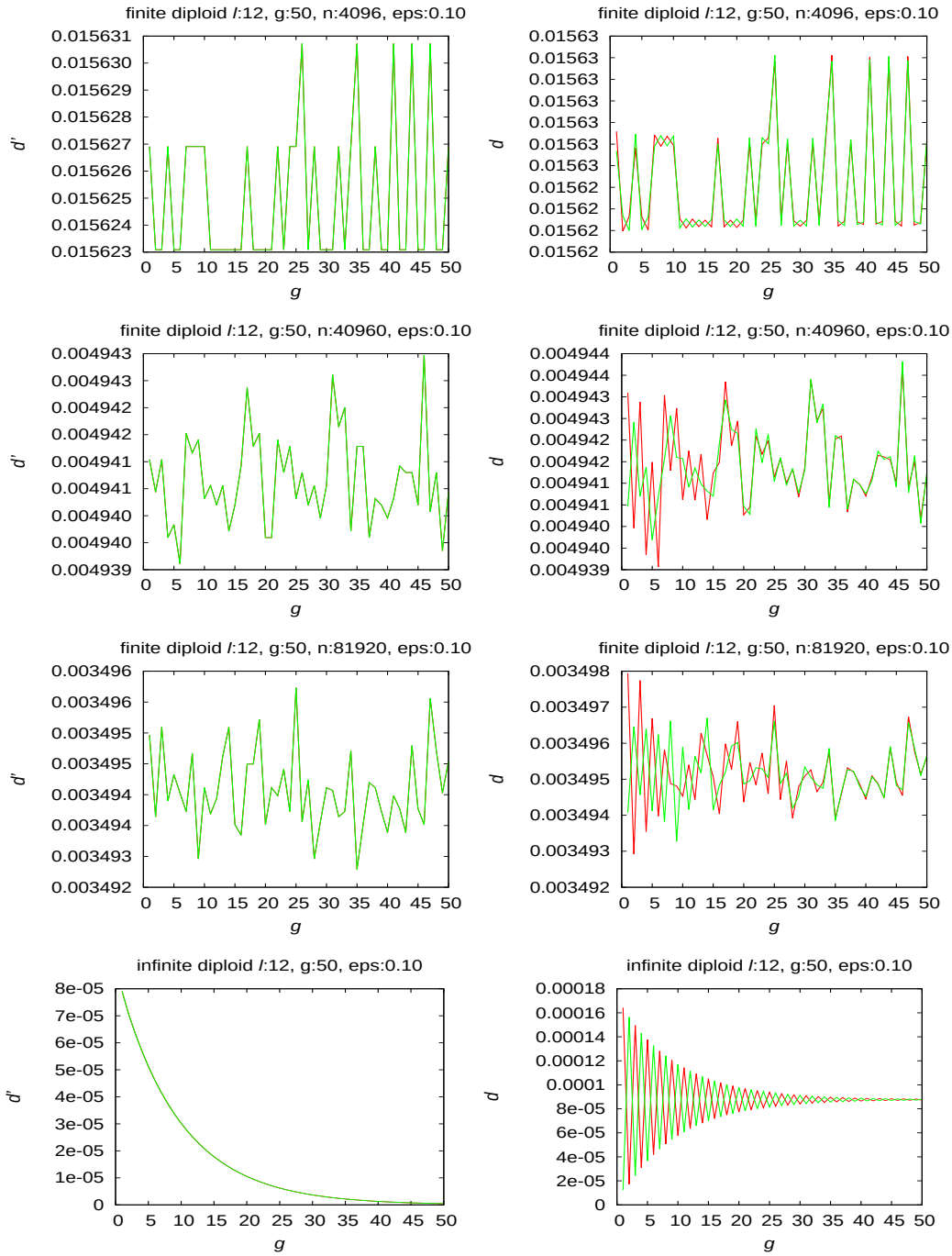


Figure 5.19: Infinite and finite diploid population behavior for χ violation, genome length $\ell = 12$ and $\epsilon = 0.1$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

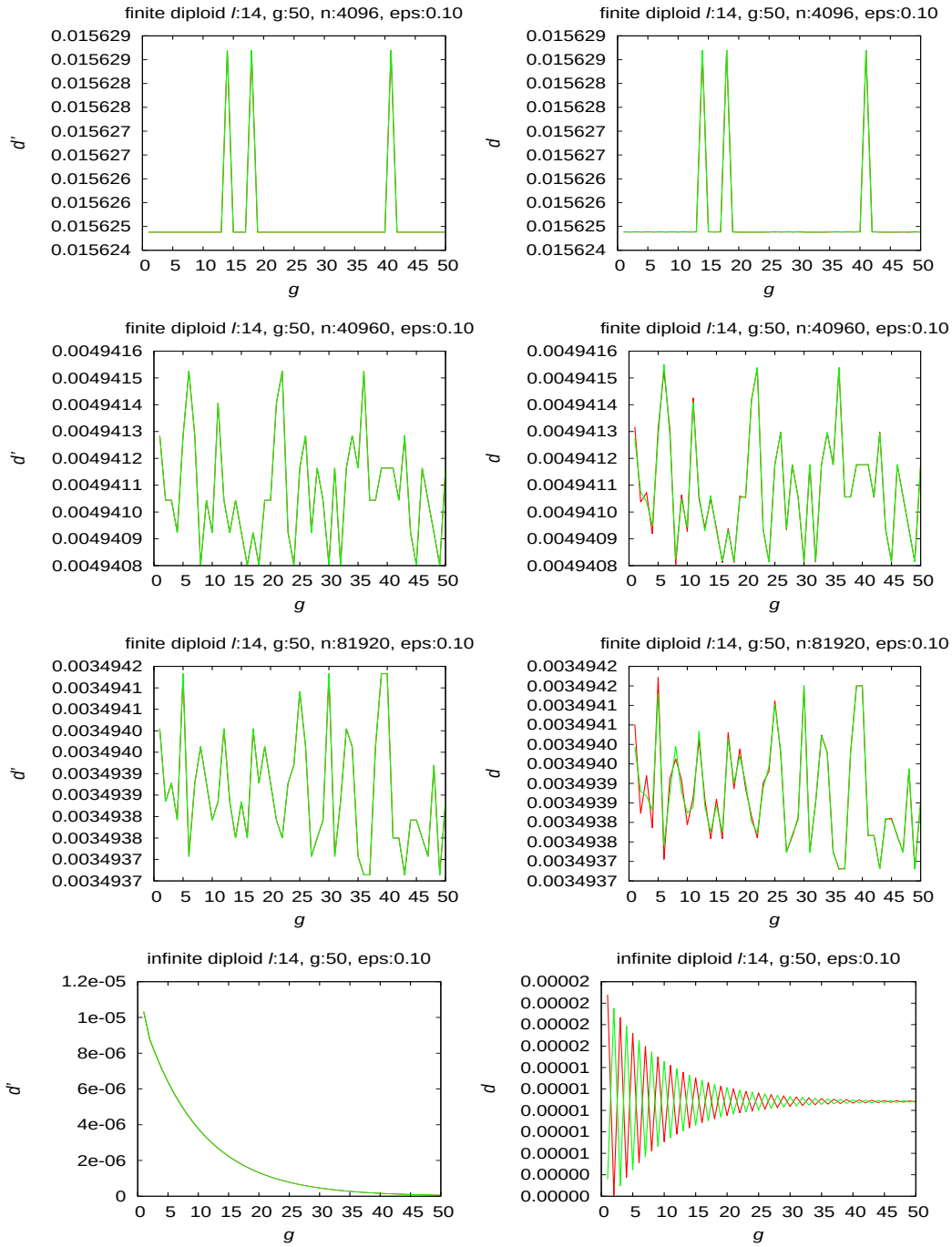


Figure 5.20: Infinite and finite diploid population behavior for χ violation, genome length $\ell = 14$ and $\epsilon = 0.1$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

Table 5.5: Distance measured for violation in χ with $\epsilon = 0.1$ for diploids: ℓ is genome length, average distance between finite and infinite population is tabulated in the last three columns, and last row is expected single step distance.

ℓ	$N = 4096$	$N = 40960$	$N = 81920$
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
$1/\sqrt{N}$	0.0156	0.0049	0.0035

5.1.6 Diploid Population $\sim \epsilon : 0.5$

The right column in figures 5.21 through 5.24 shows distance of finite and infinite diploid populations with $\boldsymbol{\epsilon} = 0.5$ to non-violation limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Infinite population oscillation quickly dies out. Finite populations show some oscillations when $\ell = 8$ for higher population size for some generations before randomness appears, as in figure 5.21, but for larger ℓ, finite populations show only randomness.

The left column of figures 5.21 through 5.24 shows distance of finite and infinite diploid populations with $\boldsymbol{\epsilon} = 0.5$ to limit $\boldsymbol{z}^*$ (limit with violation in crossover distribution $\boldsymbol{\chi}$). The distance decreases as population size increases. Average distance data for diploid population in case of violation in $\boldsymbol{\chi}$ distribution with $\boldsymbol{\epsilon} = 0.5$ are tabulated in table 5.6.

Table 5.6 shows that the average distance between finite and infinite populations approaches expected single step distance $1/\sqrt{N}$.

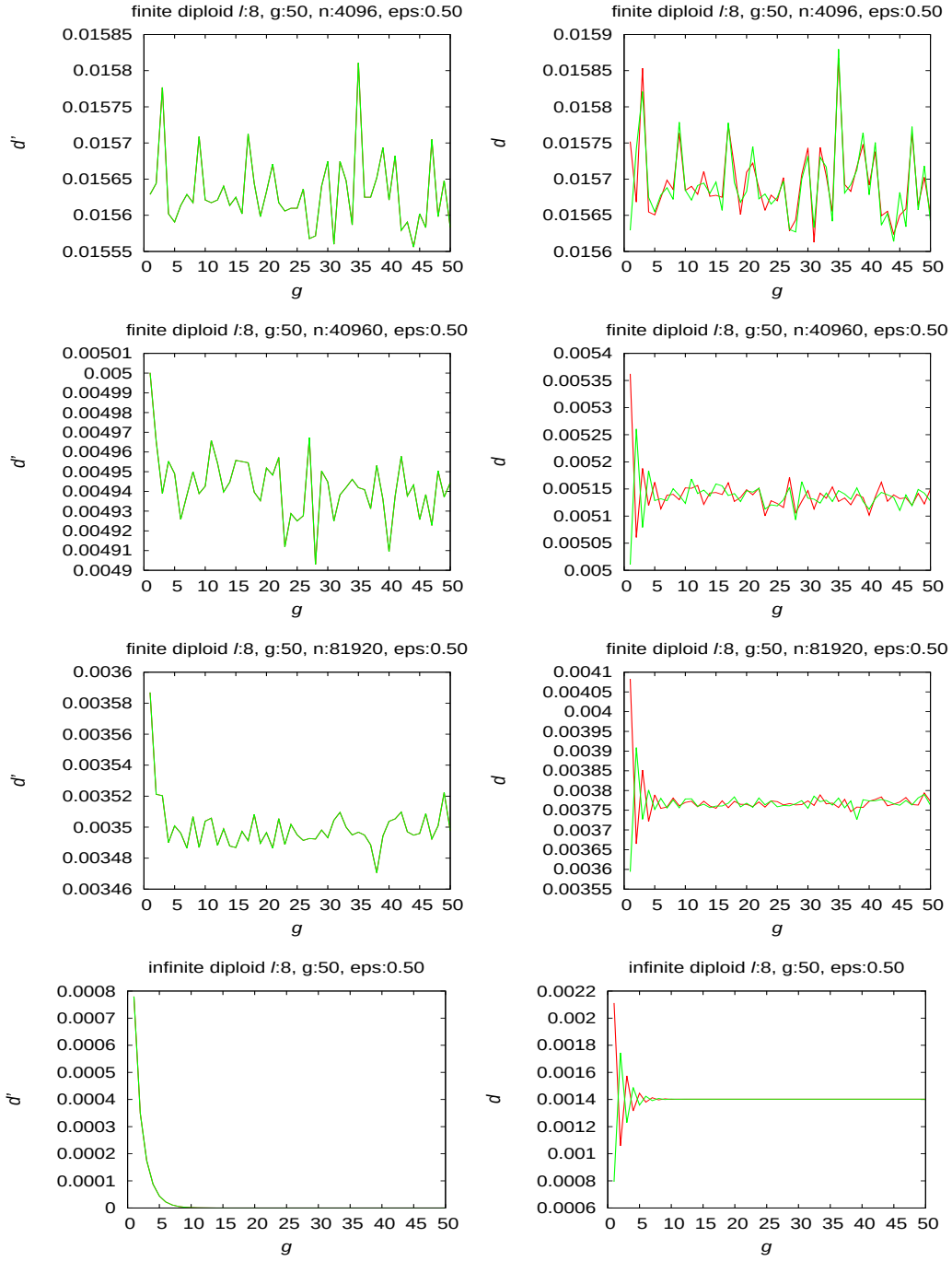


Figure 5.21: Infinite and finite diploid population behavior for χ violation, $\ell = 8$ and $\epsilon = 0.5$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

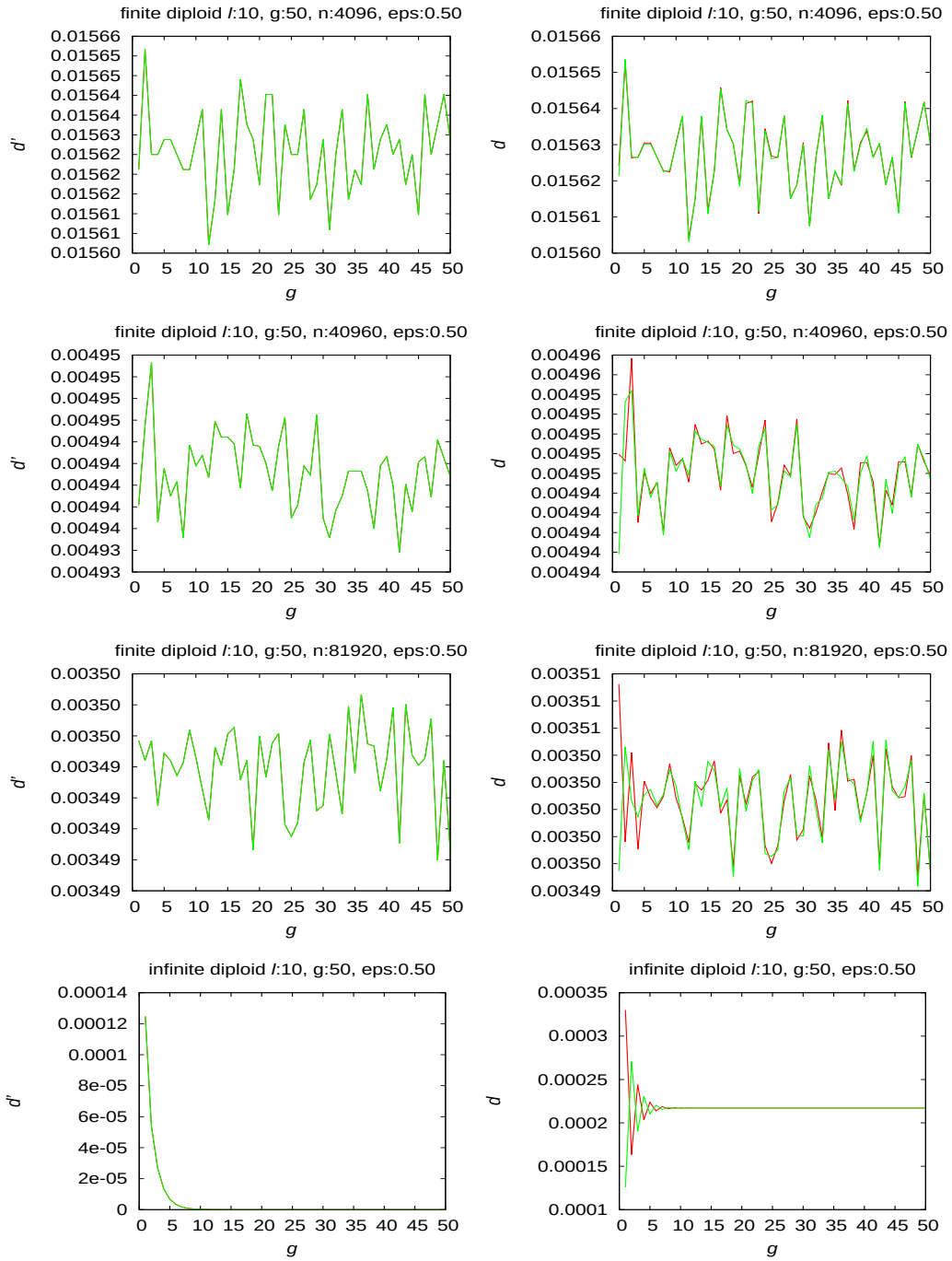


Figure 5.22: Infinite and finite diploid population behavior for χ violation, genome length $\ell = 10$ and $\epsilon = 0.5$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

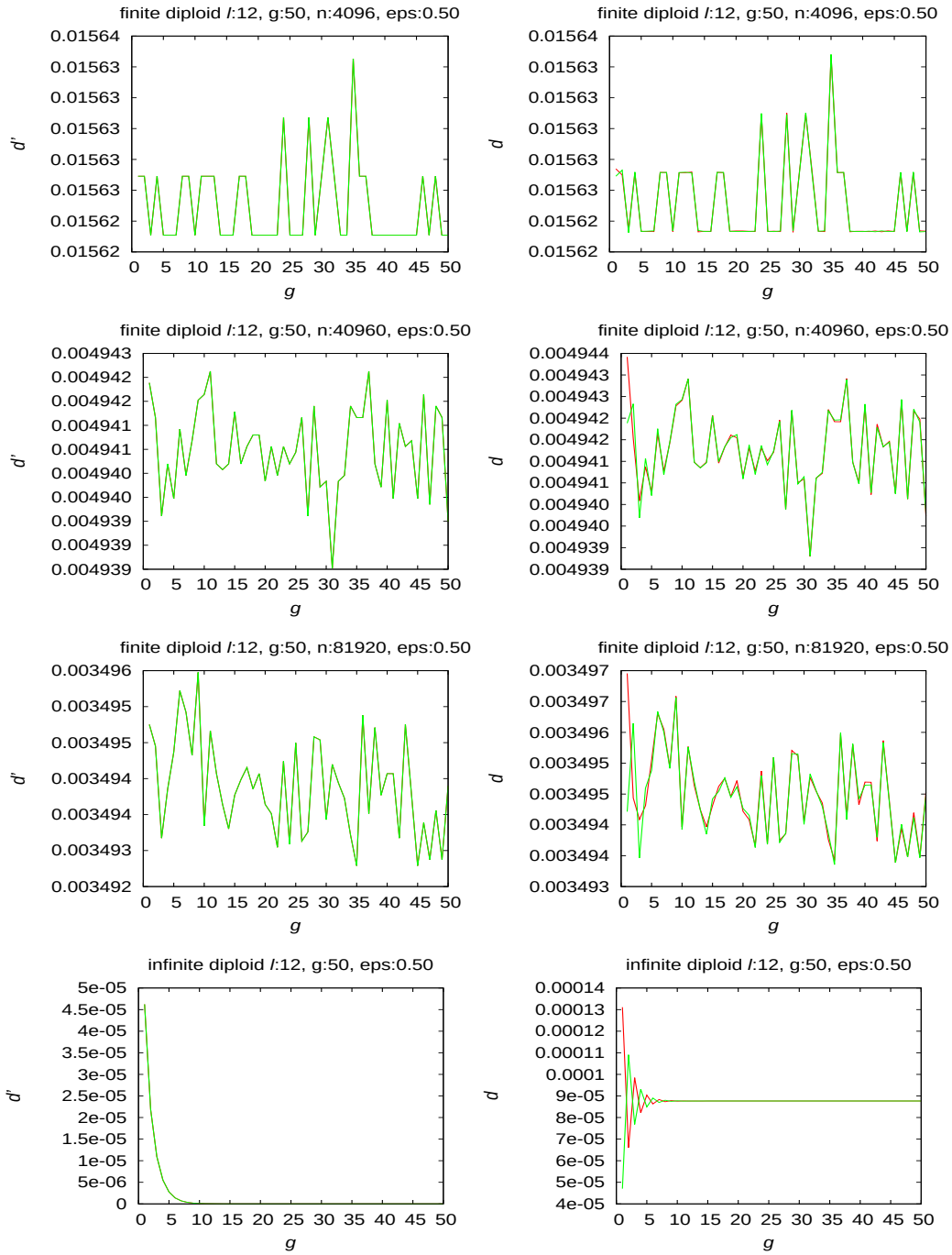


Figure 5.23: Infinite and finite diploid population behavior for χ violation, genome length $\ell = 12$ and $\epsilon = 0.5$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

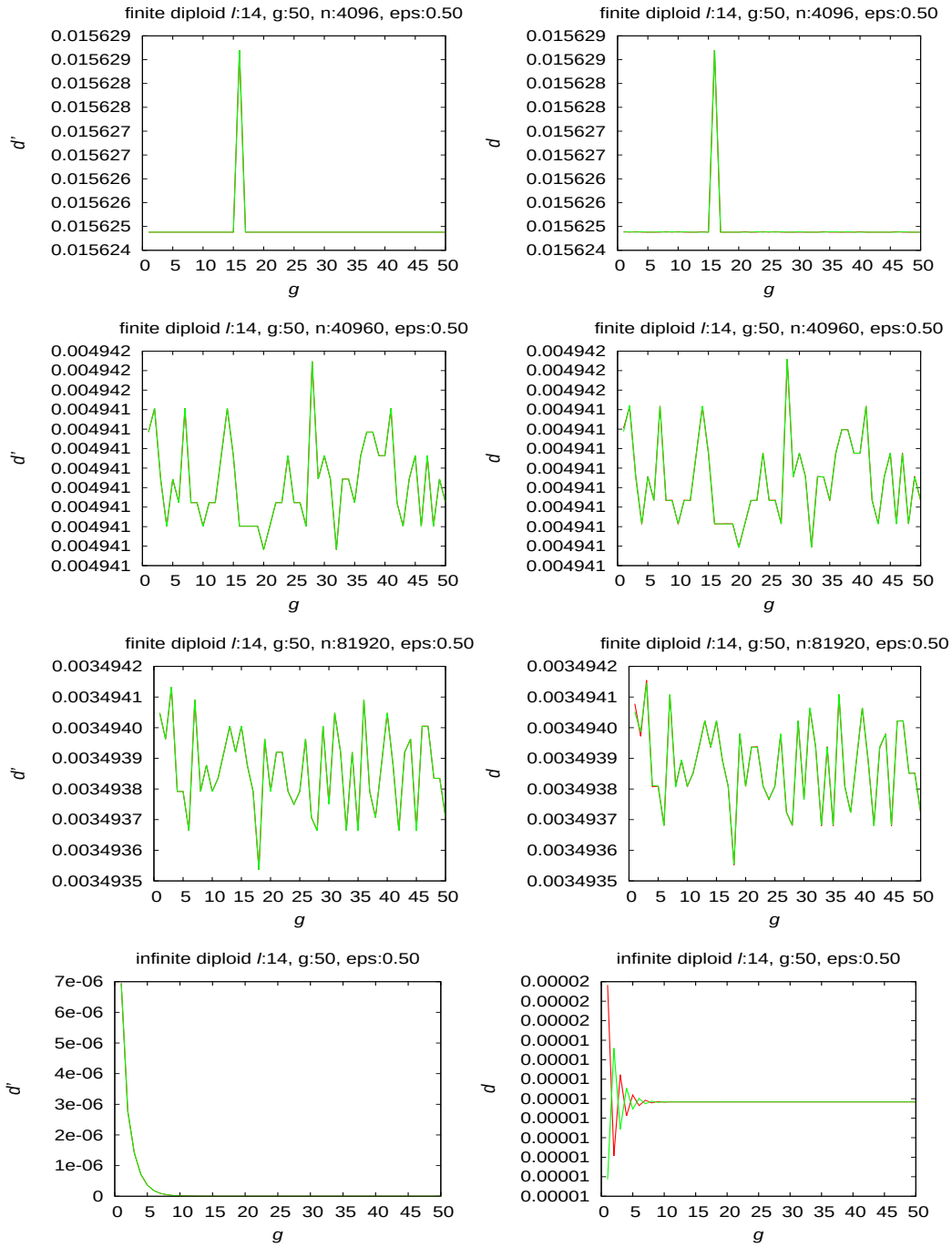


Figure 5.24: Infinite and finite diploid population behavior for χ violation, genome length $\ell = 14$ and $\epsilon = 0.5$: In left column, d' is distance of finite or infinite population to limit $\boldsymbol{z}^*$ for g generations. In right column, d is distance of finite or infinite population to limits $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. Green line is distance to $\boldsymbol{p}^*$ and red line is distance to $\boldsymbol{q}^*$.

Table 5.6: Distance measured for violation in χ with $\epsilon = 0.5$ for diploids: ℓ is genome length, average distance between finite and infinite population is tabulated in the last three columns, and last row is expected single step distance.

ℓ	$N = 4096$	$N = 40960$	$N = 81920$
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
$1/\sqrt{N}$	0.0156	0.0049	0.0035

5.2 Discussion

In the presence of violation in $\boldsymbol{\mu}$, the amplitude of oscillation decreases as string length ℓ increases. Larger population sizes show better oscillation. Since diploid populations have effective string length twice the size of haploid populations, diploid populations need larger population size than haploid population to exhibit good oscillation. As in the case of violation in $\boldsymbol{\mu}$, increasing string length ℓ degrades convergence (as finite population size increases) to infinite population behavior for diploid populations. That behavior is noticeable in figures 5.13 through 5.24 for violation in $\boldsymbol{\chi}$. That behavior is less noticeable in haploid populations.

With increase in the value of $\boldsymbol{\epsilon}$, oscillation in population diminishes and dampening of oscillation increases. Randomness increases with increasing $\boldsymbol{\epsilon}$. Comparing oscillation with violation in $\boldsymbol{\mu}$ and $\boldsymbol{\chi}$, rate of dampening of oscillation with violation in $\boldsymbol{\chi}$ seems to be slower than with violation in $\boldsymbol{\mu}$. Diploid populations jumping to other levels were observed for string lengths 12 and 14 and population size 4096 (figures 5.15, 5.16, 5.19, 5.20, 5.23 and 5.24), but unlike the case of violation in $\boldsymbol{\mu}$, the behavior is noticeable when the population size is larger (figure 5.19).

Figure 5.25 summarizes the distance data from tables 5.1 through 5.6. Distance between infinite and finite populations for population sizes 4096, 40960, 81920 are

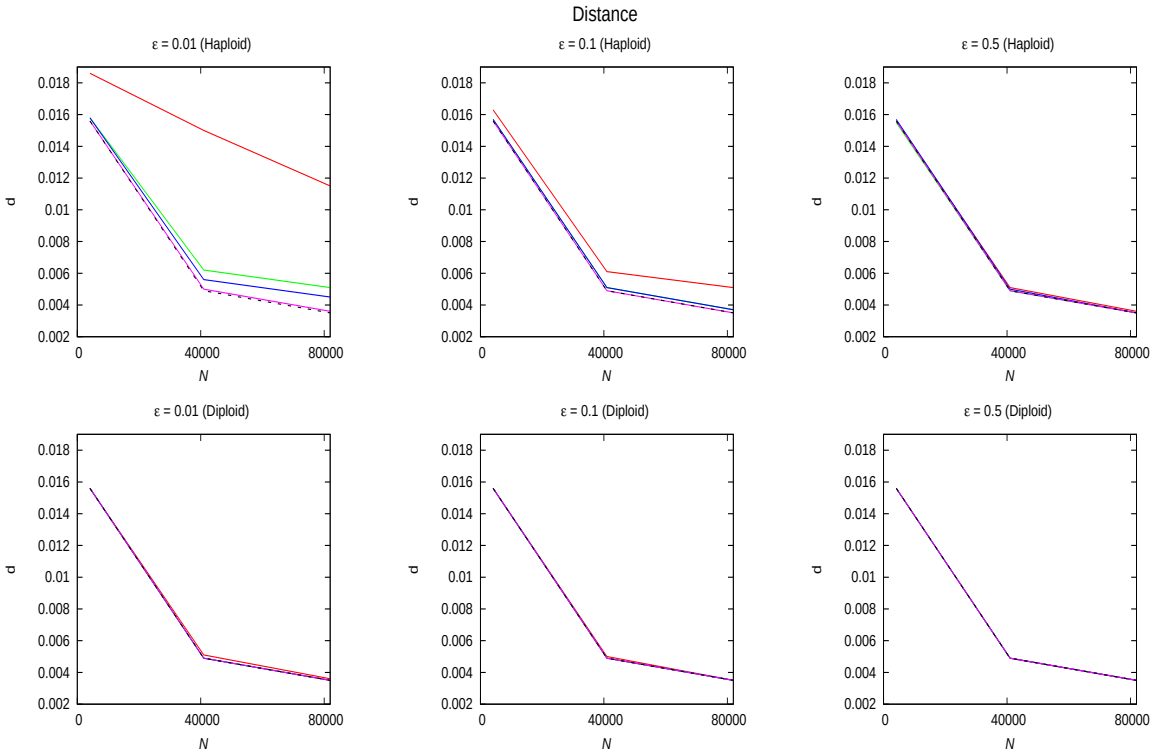


Figure 5.25: Distance between finite and infinite population in case of violation in χ: d is distance; N is finite population size; ϵ is level of violation; red line represents distance for $\ell = 8$, green line for $\ell = 10$, blue line for $\ell = 12$, pink line for $\ell = 14$ and black dotted line for expected single step distance.

plotted for different ℓ. Plots for different violation levels ϵ are arranged in columns. Plots for haploid and diploid populations are arranged in two rows. With increase in ℓ, distance between finite and infinite population moves closer to the single step distance. So, since diploid effective population string length is twice that of haploid population, distance in diploid case moves closer to the single step distance for the same value of ℓ than in haploid case. Like in the case of μ violation, it is more noticeable in haploid population case that as ϵ increases, the distance moves closer to the single step distance.

5.3 Summary

In this chapter, we violated condition 3.3 for the crossover distribution, so that infinite population trajectories have no periodic orbit. We explored infinite and finite population oscillation behavior with the violation through experiments. We did not prove that the Markov chain is not regular in this case, but we suspect it is not. Like in case of $\boldsymbol{\mu}$ violation, infinite population oscillation dies out when condition 3.3 for convergence to periodic orbits is violated, but finite populations approximately oscillate for small values of $\boldsymbol{\epsilon}$ because the probability of using the new mask is low, and when not used, finite population evolution behavior follows behavior of infinite population without violation in the condition for convergence to periodic orbits. However, rate of dampening of oscillation with violation in $\boldsymbol{\chi}$ is observed to be slower than with violation in $\boldsymbol{\mu}$. Also more randomness in oscillations are observed in this case than in violation in mutation, especially for diploid population.

Chapter 6

Conclusion And Future Work

6.1 Conclusion

This research shows how Vose's haploid model for Genetic Algorithms extends to the diploid case, facilitating the computation of infinite population evolutionary trajectories by significantly reducing the time and space used. Efficiency is achieved through reducing diploid evolution to the evolution of haploid 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 to infinite population evolutionary trajectories. Results confirm that distance can be inversely proportional to the square root of population size.

Simulations showed that when the necessary and sufficient condition for oscillation in infinite populations is met, finite populations also exhibit approximate oscillation. Amplitude of oscillation increases with increase in population size, and larger population exhibit better oscillation. Moreover, amplitude of oscillation decreases with increase in genome length.

When the condition for inifinite population oscillation is violated for the mutation distribution, the Markov chain representing finite population evolution is regular, and hence, perfect oscillation can not occur. However, simulation results show finite populations continue to approximately oscillate if the violation is small, and when the violation is larger, oscillation dies out and randomness in behavior increases.

When the condition is violated for the crossover distribution, we did not prove that the Markov chain formed is regular or not, but results show finite populations continue to approximately oscillate when the violation is small, and randomness in behavior increases when the violation is larger. As genome length increases oscillation in population degrades. Moreover, larger population shows better oscillation as in the case of oscillation with violation.

6.2 Future Work

In figures 4.19, 4.20, 5.19 and 5.20, infinite population oscillation dies out symmetrically to give graph of single straight line. But infinite population is converging to limit $\boldsymbol{z}^*$. This suggests $\boldsymbol{z}^*$ may be somewhere equidistant from $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$. We devised a test to check whether $\boldsymbol{z}^*$ lies between hyperplanes H_1 and H_2, both perpendicular to the line joining $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$, H_1 containing $\boldsymbol{p}^*$ and H_2 containing $\boldsymbol{q}^*$. Let $\boldsymbol{n}$ be unit vector parallel to the line joining $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$ as shown in figure 6.1

$$\boldsymbol{n} = \frac{\boldsymbol{p}^* - \boldsymbol{q}^*}{\|\boldsymbol{p}^* - \boldsymbol{q}^*\|}$$

Then a point x is *between* H_1 and H_2 if

$$\boldsymbol{n}^T(x - \boldsymbol{p}^*) < 0 \qquad \text{and} \qquad \boldsymbol{n}^T(x - \boldsymbol{q}^*) > 0.$$

Note that $\boldsymbol{n}^T(x - \boldsymbol{p}^*)$ is dot-product of $\boldsymbol{n}$ and $(x - \boldsymbol{p}^*)$; its value is

$$\|x - \boldsymbol{p}^*\| \cos\phi$$

where ϕ is angle between vectors $\boldsymbol{n}$ and $(x-\boldsymbol{p}^*)$. Likewise, $\boldsymbol{n}^T(x-\boldsymbol{q}^*)$ is dot-product of $\boldsymbol{n}$ and $(x-\boldsymbol{q}^*)$; its value is

$$\|x-\boldsymbol{q}^*\|\cos\theta$$

where θ is angle between vectors $\boldsymbol{n}$ and $(x-\boldsymbol{q}^*)$.

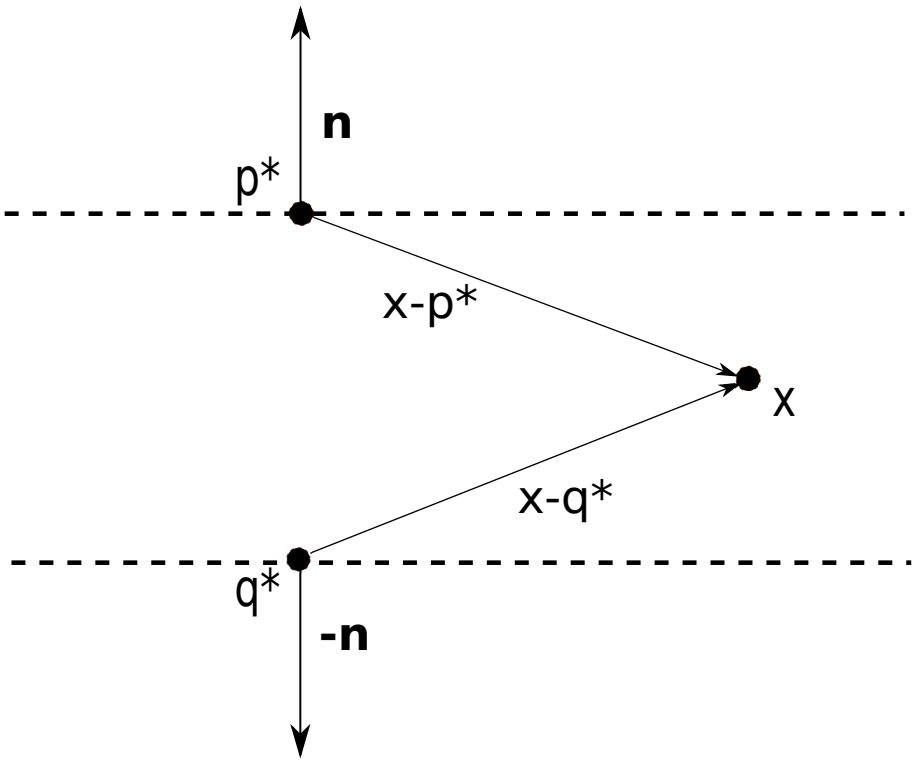


Figure 6.1: Geometry of GA: $\boldsymbol{p}^*$, $\boldsymbol{q}^*$ and $\boldsymbol{z}^*$

Our tests show $\boldsymbol{z}^*$ is between H_1 and H_2 and also equidistant from $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$ in both the haploid and diploid case. We also ran tests for population points. In haploid case, both infinite and haploid populations were between H_1 and H_2. In diploid case, infinite populations were between H_1 and H_2 but finite populations were not. These geometric properties of GA were uncovered by our simulations. Whether these observations persist to simulations we have not checked, or whether they only are true for those we considered is at this point unknown. Perhaps there are more

geometric properties and details that can be discovered through further simulations of evolutionary system. That is a topic for future investigation.

Bibliography

Akin, E. (1982). Cycling in simple genetic systems. *J. Math. Biology*, 13:305–324. 13

Beauchamp, K. (1975). *Walsh functions and their applications*. Academic Press. 24

Bethke, A. D. (1980). *Genetic Algorithms As Function Optimizers*. PhD thesis, The University of Michigan. 4, 7

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. 25

Crow, J. and Kimura, M. (1970). *An introduction to population genetics theory*. New York, Evanston and London: Harper & Row. 32

Geiringer, H. (1944). On the probability theory of linkage in mendelian heredity. *Ann. Math. Stat.*, 15(1):25–27. 19, 22

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. 7

Goldberg, D. E. (1989b). Genetic algorithms and walsh functions-partii: Deception and its analysis. *Complex systems*, 3(153–171). 7

Häggström, O. (2002). *Finite Markov Chains and Algorithmic Applications*. Cambridge University Press. 15

Haldane, J. B. S. (1932). *The Causes Of Evolution*. Longmans, New York. 4

Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science*, 28(706):49–50. 18, 19

Hastings, A. (1981). Stable cycling in discrete-time genetic models. *Proc. Nat. Acad. Sci.*, 78:7224–7225. 13

Holland, J. H. (1992). *Adaptation in natural and artificial systems*. Cambridge : MIT Press. 2, 4

Iosifescu, M. (1980). *Finite Markov Processes and Their Applications*. Dover Publications Inc, Mineola, New York. 15, 55

Koehler, G. J. (1994). A proof of the vose-liepins conjecture. *Annals of Mathematics and Artificial Intelligence*, 10(4):409–422. 7

Koehler, G. J., Bhattacharyya, S., and Vose, M. D. (1997). General cardinality genetic algorithms. *Evol. Comput.*, 5(4):439–459. 7

Mendel, G. (1865). Versuche über plflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn*, IV:3–47. 19

Minc, H. (1988). *Nonnegative Matrices*. A Wiley-Interscience Publication. 15, 55

Mitchell, M. (1999). *An Introduction to Genetic Algorithms*. The MIT Press. 2

Nix, A. E. and Vose, M. D. (1992). Modeling genetic algorithms with markov chains. *Annals of Mathematics and Artificial Intelligence*, 5(1):79–88. 6

Shanks, J. L. (1969). Computation of the fast walsh-fourier transform. *IEEE Trans. Comput.*, 18(5):457–459. 7, 24, 28

Vose, M. and Liepins, G. E. (1991). Punctuated equilibria in genetic search. *Complex systems*, 5(1):31–44. 5, 7

Vose, M. D. (1999). *The simple genetic algorithm: foundations and theory*, volume 12. MIT press. 1, 7, 8, 9, 10, 11, 12, 13, 16, 21, 29, 31, 34, 35, 51

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. 7, 21, 22, 23, 25

Wikipedia, C. (2016a). Chebyshev's inequality. 11

Wikipedia, C. (2016b). Jensen's inequality. 12

Wright, A. H. and Agapie, A. (2001). Cyclic and chaotic behavior in genetic algorithms. *Proceedings of the 3rd Annual Conference on Genetic and Evolutionary Computation*, (7):718–724. 14, 16

Wright, A. H. and Bidwell, G. L. (1997). A search for counterexamples to two conjectures on the simple genetic algorithm. *Foundations of genetic algorithms*, 4:73–84. 13

Vita

Mahendra Duwal Shrestha was born in Kathmandu, Nepal, to the parents Krishna Prasad and Laxmi Duwal Shrestha. He is the first of three children: Ambika and Anuka. He attended Panga Secondary School up to tenth grade. Graduating tenth grade in distinction division, he attended NIC Higher Secondary School, where he was also awarded scholarship, taking Physics and Mathematics major. After graduation, he attended Pulchowk Engineering Campus, Institute of Engineering, Tribhuvan University where he was merit listed student for four years. He was introduced to applied mathematics, programming in C and C++, and electronic circuits and communication. He got bachelor's degree of Engineering in Electronics and Communication from Pulchowk Engineering Campus, IOE, Tribhuvan University in March 2011. He worked for two and half years as software engineer at Deerwalk Services Pvt. Ltd, developing health care business applications for clients in US. During the short stint at Deerwalk Services, he learnt and applied knowledge in programming and scripting languages as C#, Java, Groovy and javascript, web framworks as ASP.NET, ASP.NET MVC, Grails, Ajax and jquery, and databases as MYSQL and MSSQL. He accepted teaching assistantship at The University of Tennessee, Knoxville in Electrical Engineering and Computer Science program. He is continuing his education in Masters of Science in Computer Science working as research assistant at Sergey's lab in Ecology and Evolutionary Biology.