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

Michael D. Vose, Major Professor

We have read this thesis
and recommend its acceptance:

Michael D. Vose

Hairong Qi

Judy D. Day

Accepted for the Council:

Dixie Thompson

Vice Provost and Dean of the Graduate School

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

Michael D. Vose, Major Professor

We have read this thesis
and recommend its acceptance:

Michael D. Vose

Hairong Qi

Judy D. Day

Accepted for the Council:

Dixie Thompson

Vice Provost and Dean of the Graduate School

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

Analysis and Simulation Of A Simple Evolutionary System

A Thesis Presented for

The Master of Science

Degree

The University of Tennessee, Knoxville

Mahendra Duwal Shrestha

August 2016

dedication...

Acknowledgements

I would like to thank...

Some quotation...

Abstract

Abstract text goes here...

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

Genetic algorithm (GA) is inspired by population genetics, and is population based, and proceeds over a number of generations to obtain optimal solution. GA is powerful, broadly applicable techniques to solve problems not yielding to other known methods. The basic mechanism of a GA, although there are many variants, consists of:

1. Evaluation of individual fitness and formation of a gene pool
2. Recombination and mutation.

The members of next generation population is formed by individuals produced from these operations. Criteria for fitness evaluation exerts evolutionary force for populations to produce more fit individuals. The process is repeated until system stops to improve or threshold is met.

The members of population are typically fixed length binary strings (also called genome length in later chapters). These members contribute to gene pool according to their relative fitness which is calculated using some objective function. They are mutated and recombined by crossover. Mutation corresponds to flipping the bits of an individual with some small probability, the mutation rate. During crossover, two parents are selected from the pool, a random but same position is chosen within each

parent string and segments are exchanged. Small probability, crossover rate, is used perform crossover and otherwise parents are cloned. New offsprings produced after mutation and crossover form next generation.

In the book "An Introduction to Genetic Algorithms" by Mitchell (see **?**), she mentioned 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 were given to them. Mitchell mentioned genetic algorithms were developed by Holland and his colleagues in the 1960s and the 1970s. Mitchell mentioned in her book that Holland introduced a population-based algorithm with crossover and mutation. With schema theorem (see Holland (1992)), Holland gave some perspective on expected next generation by showing expectation of schema survival in next generation of population. A schema is a template that identifies a subset of strings with similarities at certain string positions, and a template is made up of 1s, 0s, and $*$s where $*$ is the 'don't care' symbol that matches either 0 or 1. However, the schema theorem by Holland did not compute expectation of strings in next generation which is considered of higher importance in GA. Bethke (see Bethke (1980)) gave equations for computing expectation of number of a string in next generation. Goldberg (see **?**) used equations for the expected next generation to model the evolutionary trajectory of a two bit GA under crossover and proportional selection. Vose and Liepins (see Vose and Liepins (1991)) simplified and extended these equations integrating mutation into the recombination of arbitrarily long binary strings. Vose and Liepins modeled simple GA by computing expected population trajectories through time based on infinite population. Like Goldberg, their equations are deterministic. Vose has proved infinite population models can be used to explain some aspect of finite population behavior. Given a finite population with proportional representation vector $\boldsymbol{p}^n$ at generation n with component $\boldsymbol{p}_i^n$ as proportion of string i in finite population, infinite population model can be used to compute the expected proportion $\boldsymbol{p}_i^{n+1}$ of string i as result of *selection* and *mixing* in next generation finite population $\boldsymbol{p}^{n+1}$. If $\boldsymbol{r}_{i,j}(k)$ is probability

that parents i and j recombine to produce child k, and $\boldsymbol{s}_i^t$ and $\boldsymbol{s}_j^t$ are probability of selection of i and j as parents, Vose and Liepins computed expected proporiton of k in next generation as

$$\mathcal{E}(\boldsymbol{p}_k^{t+1}) = \sum_{i,j} \boldsymbol{s}_i^t \boldsymbol{s}_j^t \boldsymbol{r}_{i,j}(k); \quad \mathcal{E} \text{ denotes expectation}$$

If M is recombination matrix with elements $\boldsymbol{m}_{i,j} = \boldsymbol{r}_{i,j}(0)$ and permutation σ_j defined as

$$\sigma_j \langle S_0, .., S_{2^\ell - 1} \rangle^T = \langle S_{0+j}, .., S_{(2^\ell - 1)+j} \rangle^T$$

where T is transpose and ℓ is bit length of binary string, Vose and Liepins represented expected proportion of k in next generation in using recombination or mixing matrix M as

$$\mathcal{E}(\boldsymbol{p}_k^{t+1}) = (\sigma_k \boldsymbol{s})^T M (\sigma_k \boldsymbol{s})$$

Vose and Nix (see Nix and Vose (1992)) further explored issues regarding relationship between real GA and infinite population model. Vose and Nix obtained an exact model for real GAs in the form of a Markov chain. For a non-zero positive mutation rate, mutation will produce any possible string in finite population with non-zero probability and hence a real GA will form ergodic Markov chain, visiting every state infinitely often in a long run. Another result Vose and Nix obtained was that trajectory followed by finite populations was related to the evolutionary path predicted by the infinite population model. Vose and Nix proved that for large populations, the evolutionary path of a real GA follows very closely, with large probability, and for a long period of time that path predicted by the infinite population model. So if we form a geometrical cylinder around the path of infinite population model, a real GA will stay inside the pipe for short term and then escape out of it after some period of time. Larger population stay inside the pipe for a longer period of time and smaller population stay inside for shorter period of time.

In his book *Simple Genetic Algorithm: Foundations and Theory* (see Vose (1999)), Vose compiled and extended all of his previous work regarding infinite population model and application of walsh transforms to mixing matrix. Vose provided mathematical implementation of simple GA for binary strings; he provided mathematical implementation of selection, crossover, and mutation. Vose gave formula for calculating mixing matrix under his implementation of crossover and mutation. Vose discussed how application fo Walsh transform to mixing matrix simplified the matrix, giving computational efficiency in calculating matrix.

There had been previous applications of Walsh transform in field of GA. Bethke, in his dissertation "Genetic Algorithms as Function Optimizers" (see Bethke (1980)), first introduced idea of using Walsh transforms to analyse process of GA in case of binary-coded strings. The idea of using Walsh transforms were given greater incitement in papers by Goldberg (see Goldberg (1989a), Goldberg (1989b)). But these usage of Walsh transforms does not involve the direct application of the Walsh transform to crossover and mutation, or to any of their associated mathematical objects. Vose and Liepins use Walsh transform directly to mutation and recombination, and show that the twist (denoted by M_*) of the mixing matrix (M) is triangularized by the Walsh transform and used M_* in study of fixed points where $(M_*)_{i,j} = M_{i+j,i}$. In a related paper, Koehler (see Koehler (1994)) gives a congruence transformation defined by lower triangular matrix that diagonalizes the mixing matrix for 1-point crossover and mutation given by a rate and mathecally proved conjecture provided by Vose and Liepins on eigenvalues of matrix M_*. Koehler, Bhattacharyya and Vose (see Koehler et al. (1997)) applied Fourier transform to mixing in generalizing results concerning simple genetic algorithm which were previously established for binary case (in binary case, Fourier transform is Walsh transform) extending analysis to strings over an alphbet of cardinality c. Vose and Wright (see Vose and Wright (1998)) applied Walsh transform to mixing matrix and simplified the matrix from dense to sparse in Walsh basis giving advantage of computational efficiency from $O(n^3)$ to $O(n^{\log_2 3})$. The cost of conversion of standard

coordinates to Walsh basis need not be sustained since fast Walsh transform (see Shanks (1969)) can do that in $O(n \log n)$ time.

1.2 Random Heuristic Search

Vose (see Vose (1999)) introduced abstract model, a generalized heuristic search method referred to as *Random Heuristic Search (RHS)*, defined upon the central concept of state and transition between states. An instance of *RHS* is an initial collection of elements P_0 chosen from some search space Ω , together with a stochastic transition rule τ , which from P_i will produce another collection P_{i+1}; iterating τ produces a sequence of generations.

The beginning collection P_0 is referred to as the *initial population*. Let n be the cardinality of Ω and $\mathbf{1}$ denotes column vector of all 1s. The *simplex* is the set of population descriptors:

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

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 is a constant r, called the population size. 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 new vector whose ith component is the probability that 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

*$\langle .. \rangle$ represents a column vector; $\mathbf{1}^T$ is $\langle 1, .., 1 \rangle^T$

Vose (1999))

$$Q_{\boldsymbol{p},\boldsymbol{q}} = r! \prod \frac{(\mathcal{G}(\boldsymbol{p})_j)^{r\boldsymbol{q}_j}}{(r\boldsymbol{q}_j)!} \tag{1.1}$$

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

where summation is restricted to indices for which $\boldsymbol{q}_j > 0$.

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. Such a sequence forms a Markov chain with transition matrix

$$Q_{\boldsymbol{p},\boldsymbol{q}} = r! \prod \frac{(\mathcal{G}(\boldsymbol{p})_j)^{r\boldsymbol{q}_j}}{(r\boldsymbol{q}_j)!}$$

So the conceptualization of RHS can be replaced by Markov chain model abstraction which makes no reference to sampling Ω. That is from current population $\boldsymbol{p}$, produce $\boldsymbol{q} = \tau(\boldsymbol{p})$ with probability $Q_{\boldsymbol{p},\boldsymbol{q}}$. With transition matrix defined for Markov chain model, Vose (see Vose (1999)) says the expected next generation $\mathcal{E}(\tau(\boldsymbol{p}))$ is $\mathcal{G}(\boldsymbol{p})$ and the expression in transition matrix

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

gives the qualitative information regarding probable next generation which 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 the expected next population. Hence the factor

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

in the expression of transition matrix indicates the probability that $\boldsymbol{q}$ is the next generation decays exponentially, with constant r , as the discrepancy between $\boldsymbol{q}$ and

the expected next population 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 factor

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

indicates the probability that $\boldsymbol{q}$ is the next generation decays exponentially with increasing dispersion and $\theta = \sum\log(e^{x_j}x_j!/x_j^{x_j})$.

Vose (see Vose (1999)) calculated variance of next generation population with respect to expected population as

$$\mathcal{E}(\|\tau(\boldsymbol{p}) - \mathcal{G}(\boldsymbol{p})\|^2) = (1 - \|\mathcal{G}(\boldsymbol{p})\|^2)/r \tag{1.3}$$

and mentioned $\tau(\boldsymbol{p})$ converges in probability to $\mathcal{G}(\boldsymbol{p})$ as the population size increases. Therefore, τ corresponds to $\mathcal{G}$ in the infinite population case and distance between finite and infinite population in next generation decreases as $1/r$.

1.3 Overview

In chapter two, we describe a simple Markov model for diploid case under influence of mutation and crossover. The model is non-overlapping, generational, infinite population model assuming random mating and no selective pressure. Through abstract development, we show that the diploid model can be specialized by using mask based mutation and crossover operators to Vose's infinite population model which is a haploid model. Computational simplifications due to reduction of diploid model to haploid model and application of Walsh transform are exploited in experimental simulation of model, and through the experiment we demonstrate

convergence of finite diploid population to infinite population behavior implied by equation 1.3.

In chapter three, we study evolutionary limits predicted by Vose using infinite popu- lation model under no selective pressure. We use computation of predicted limits of infinite population and discuss necessary and sufficient conditions stated by Vose for population to converge in to periodic orbits. We investigate predicting the convergence of finite population short-term behavior to infinite population evolutionary limits under no selective pressure. Then it studies case of violation in the necessary and sufficient conditions for population to converge periodic orbits. We then study behavior of finite and infinite population when there is violation in necessary condition mentioned by Vose.

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 contributes to the elegance and simplicity of the abstract development and manifests diploid evolution equations can be represented by haploid equations.

A basic syntactic model for haploid and diploid genomes is considered in the beginning and commented on its expressive power. Then the mechanics of how the $(n+1)$th generation is obtained from the nth 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.

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 $\ell = \ell_1 + \cdots + \ell_h$). Partitioning may be extended to chromosomes so as to interpret each as a collection of genes. If continued to the granularity of pairs of bits, partitioning allows, for example, representing the four possibilities Adenine, Guanine, Cytosine, and Thymine.

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

Following Gieringer (see Geiringer (1944)), let the transmission function $t_\alpha(g)$ be the probability that gamete g is produced from parental genome α. It follows from the above that the equation determining the next generation q^{n+1} is

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

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

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,

$$q^{n+1}_{\langle\gamma_0,\gamma_1\rangle} \;=\; q^{n+1}_{\langle\gamma_1,\gamma_0\rangle}$$

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)) (issues he considered and results he obtained relating to invariant distributions for a particular sort of transmission function are not here mentioned because they are irrelevant to the purpose of this section).

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 $p^{n+1}_{\gamma_0}$ and the second by $p^{n+1}_{\gamma_1}$ so that

$$q^{n+1}_{\langle\gamma_0,\gamma_1\rangle} \;=\; p^{n+1}_{\gamma_0}\, p^{n+1}_{\gamma_1} \tag{2.2}$$

[†]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

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

Let $[\mathit{expression}]$ denote 1 if *expression* is true, and 0 otherwise.[§] 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 g th component of the distribution p^n. Moreover, (2.5) and (2.2) show (for $n>0$) invertibility of the map

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

[§] $[\cdots]$ is sometimes referred to as an *Iverson bracket*.

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

The set of haploids (i.e., length ℓ binary strings) is a commutative ring $\mathcal{R}$ under component-wise addition and multiplication modulo 2. This algebraic structure is crucial to Vose's specialization and subsequent analysis of (2.8). Denote the additive identity by $\mathbf{0}$ and the multiplicative identity by $\mathbf{1}$, and 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 $g \in \mathcal{R}$.

2.3.1 Mutation

Mutation simulates effects of error that happen with low probability during duplication of chromosome. Mutation provides mechanism to inject new strings into the next generation population which gives *RHS* ability to search beyond the confines of initial population.

Symbol $\boldsymbol{\mu}$ is used to represent mutation distribution describing the probability $\boldsymbol{\mu}_i$ with which $i \in \Omega$ is selected to be a mutation mask. $\boldsymbol{\mu} : \Omega \to \Omega$ is nondeterministic mutation function where the result $\boldsymbol{\mu}(x)$ of applying mutation function on x is $x + i$ with probability $\boldsymbol{\mu}_i$ of distribution $\boldsymbol{\mu}$ where i is *mutation mask*. Mutating x using mutation mask i alters the bits of x in those positions the mutation mask i is 1. $\boldsymbol{\mu} \in [0, 0.5)$ is regarded as a *mutation rate* which implicitly specifies distribution $\boldsymbol{\mu}$ according to rule (see Vose and Wright (1998))

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

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 $\boldsymbol{\mu}_{g+g'}$.

2.3.2 Crossover

Crossover refers to crossing over (also termed recombination) between two chromosomes (strings in our case). Crossover like mutation also provides mechanism for injection of new strings into new generation population. Masked based crossover is used in this document. Geiringer (see Geiringer (1944)) used crossover mask with probability (distribution) associated with the mask to generate offsprings from parent chromosomes in absence of mutation and selection. Let $\boldsymbol{\chi}_m$ be 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 $\boldsymbol{\chi}_m/2$ for each u' and v'.

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

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

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

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

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

2.3.3 Mixing Matrix

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

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

Crossover and mutation are defined in a manner respecting arbitrary 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 $\boldsymbol{\chi}$, then transmission function can be expressed as (see Vose and Wright (1998))

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

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

Operations (+ and ·) acting on elements of $\mathcal{R}$ in this section are component-wise addition and multiplication modulo 2.

The Walsh matrix is defined by

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

where $N^{-1/2}$ is normalization factor and nt is bitwise dot product of binary representation of number n and t.

The matrix is symmetric, i.e.,

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

and it has entries satisfying

$$W_{n,t+k} = N^{1/2} W_{n,t} W_{n,k}$$

The practical importance of this symmetry is that the transform and inverse represent same mathematical operation, hence simplifying the derivation and application of the transform. With the normalized form, *Walsh matrix* is its own inverse, i.e.,

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

In the matrix form, 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 $\widehat{A} = WAW$. If w is a row vector, then w in its Walsh basis $\widehat{w}$ represents wW.

Finite discrete Walsh transform pair on N sampling points, x_t, can be expressed as (see Beauchamp (1975))

$$X_n = \sum_{t=0}^{N-1} x_t W_{n,t} \tag{2.12}$$

$$n = 0, 1, 2...N-1$$

and

$$x_t = \sum_{n=0}^{N-1} X_n W_{n,t}$$

$$t = 0, 1, 2...N-1$$

2.4.1 Fast Walsh Transform

However, computation of discrete Walsh transform given by equation (2.12) takes n^2 operations (addition or subtraction). An algorithm using matrix factorization techniques is found to perform transformation in $n \log_2 n$ operations. This algorithm is Fast Walsh transform (FWT). Shanks (see Shanks (1969)) described FWT algorithm which is analogous to Cooley-Tukey algorithm (see Cooley and Tukey (1965)) for fast Fourier transformation. The algorithm for FWT can be translated into pseudocode as:

Algorithm 1 FWT pseudocode

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

2.4.2 Walsh Transform Adaptation

The Walsh transform has spectacular ability to unravel the intricacies of mixing. And that is why we adapt Walsh transform methods for computing evolutionary trajectories, which have already been established for Vose's haploid model (see Vose and Wright (1998)). Adaptation of Walsh transformation efficiently models infinite diploid population evolution. This adaptation of Walsh transormation helps in making feasible comparisons between finite and infinte diploid population short-term evolutionary behavior. Recalling evolution equation (2.11), without selection, specialized to Vose's infinite population model expressed in mixing matrix's term,

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

where the permutation matrix σ_g is defined by component equations

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

In our model, the Walsh matrix W is defined by component equations

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

where the subscripts u, v (which belong to $\mathcal{R}$) on the left hand side are interpreted on the right hand side as column vectors in $\mathbb{R}^\ell$. Columns of W form the orthonormal basis — the *Walsh basis* — which simultaneously diagonalizes the σ_g.

A change of basis which simultaneously diagonalizes the σ_g unravels the evolution equation (2.11). 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 $p^1 \;=\; \pi(q^1)$. We have little to say about the matrix of the Markov chain corresponding to the mixing transformation $\mathcal{M}$, because it is uncountable; each state is a distribution vector p describing a population. However, that is not an obstacle to computing evolutionary trajectories; (2.15) 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 $\boldsymbol{f}$.

(2.18) computes distance between finite and infinite population efficiently.

2.6 Simplification

The haploid case simplified by equations (2.13) and (2.14) are the consequence of specializing to Vose's infinite population model and computing in the Walsh basis. Time switching between the standard basis and the Walsh basis is negligible; the fast Walsh transform (in dimension n) has complexity $n \log n$ Shanks (1969).

Only one mixing matrix as opposed to 2^ℓ matrices is needed to compute the next generation; evolution equation (2.14) 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 behaviour to that of the infinite diploid population model as described in section 2 (the underlying haploid model for the infinite population case is described in section 2.1).

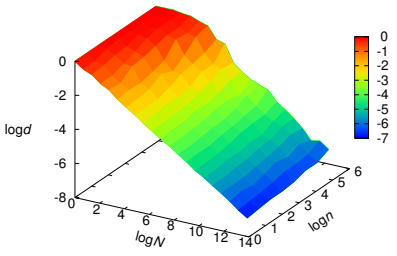
Equations (2.2), (2.13), (2.14), (2.18) were employed to illustrate efficient computation of the distance

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

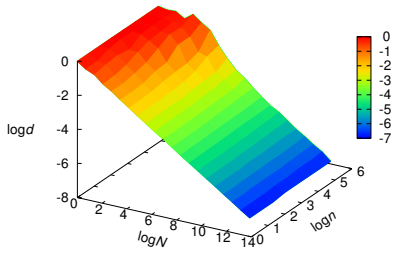
where $\boldsymbol{f}^n$ and $\boldsymbol{q}^n$ represent finite and infinite diploid populations at generation $n \in \{1, 2, 4, 8, 16, 32, 64, 128\}$ respectively, beginning from a random initial population ($\boldsymbol{f}^0 \;=\; \boldsymbol{q}^0$). Genome lengths $\ell \in \{4, 6, 8, 10, 12, 14\}$ and population sizes $N \;=\; 2^i$ for integer $0 \leq i \leq 20$ 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,

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

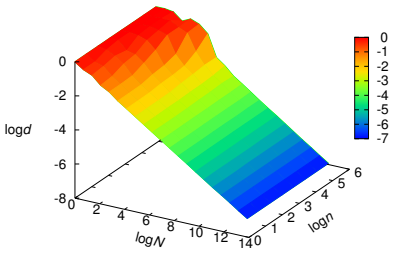
(subscripts above on the left hand side of an equality are interpreted on the right hand side of the equality as column vectors in $\mathbb{R}^\ell$). The finite population case is computed using the itemized procedural definition given in section 2.1; the transmission function (2.10) corresponds to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



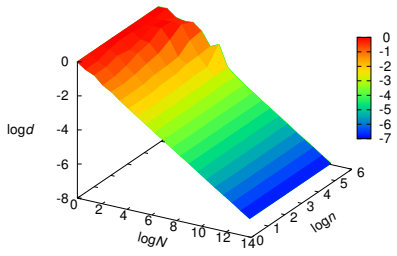
(a) $\ell = 4$.



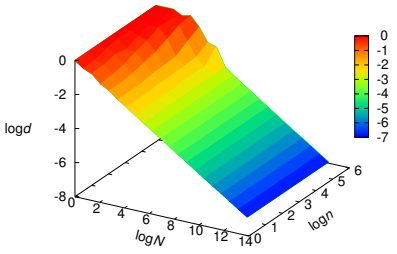
(b) $\ell = 6$.



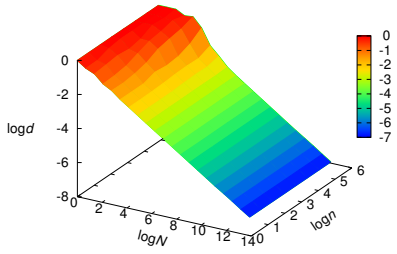
(c) $\ell = 8$.



(d) $\ell = 10$.



(e) $\ell = 12$.



(f) $\ell = 14$.

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

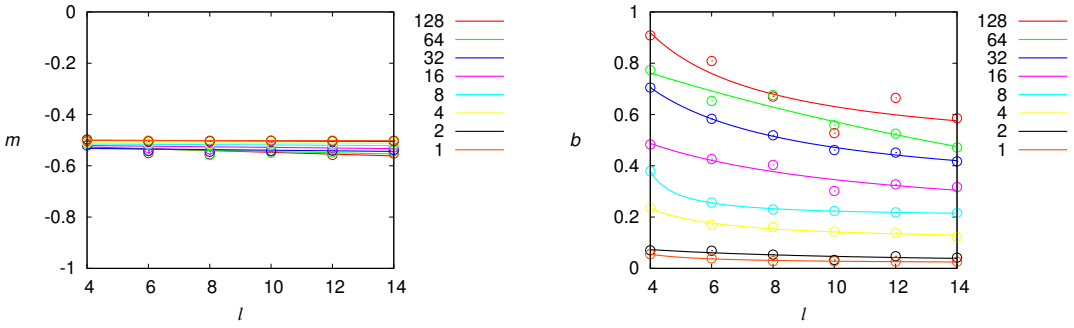
The data, presented in six surface graphs above and organized by genome length, shows a near linear dependence of $\log d$ on $\log N$. As expected, the graphs show

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).

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

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

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

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

Vose (see Vose (1999)) calculated variance of next generation population with respect to expected population as

$$\mathcal{E}(\|\boldsymbol{q} - \mathcal{G}(\boldsymbol{p})\|^2) = (1 - \|\mathcal{G}(\boldsymbol{p})\|^2)/\boldsymbol{N}$$

where $\boldsymbol{q}$ is actual population and $\mathcal{G}(\boldsymbol{p})$ is expected population. Let x be the random variable $\|\boldsymbol{q} - \mathcal{G}(\boldsymbol{p})\|$. Let ϕ be the function $\phi(x) = x^2$ which is convex function. Then $\mathcal{E}(\|\boldsymbol{q} - \mathcal{G}(\boldsymbol{p})\|^2)$ becomes $\mathcal{E}(\phi(x))$. From Jensen's Inequality (see Wikipedia (2016)), if ϕ is a convex function, then

$$\begin{aligned} \phi(\mathcal{E}(x))) &\leq \mathcal{E}(\phi(x)) \\ \mathcal{E}(x) &\leq \sqrt{\mathcal{E}(x^2)} \end{aligned}$$

Substituting original variables,

$$\mathcal{E}(\|\boldsymbol{q} - \mathcal{G}(\boldsymbol{p})\|) \leq \sqrt{(1 - \|\mathcal{G}(\boldsymbol{p})\|^2)/\boldsymbol{N}} \tag{2.21}$$

Equation 2.21 shows the expected rate of convergence for the single-step haploid case; the distance is inversely proportional to square root of population size. And equation 2.20 agrees with equation 2.21.

The consistent convergence rate across multiple generations is somewhat surprising, simulation results above indicate it may persist to generation $n = 128$.

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

2.8 Summary

In this chapter, we began with description of simple diploid Markov model under mutation and crossover with no selective pressure. With reduction to haploid and specialization using masked base recombination operators, we showed Vose's infinite population model, which is a haploid model, can be extended to diploid case. Using computational benefits of this reduction to haploid model and Walsh transform, we showed via experiment and regression of resulting data that distance between finite diploid population and infinite diploid population decreases like $1/N_{th}$ which is the single step-step haploid case convergence behavior predicted by Vose's infinite population model.

Chapter 3

Evolutionary Limits

This chapter investigates evolutionary limits predicted by Vose using infinite population model under no selective pressure. It uses computation of predicted limits of infinite population and discusse necessary and sufficient conditions stated by Vose for population to converge in to periodic orbits. We investigate predicting the convergence of finite population short-term behavior to infinite population evolutionary limits under no selective pressure. Then it studies behavior of infinite and finite population in case of violation in the necessary and sufficient conditions for population to converge periodic orbits.

3.1 Limits

Vose states under mild assumptions on mutations (considered later), populations converge under repeated application of $\mathcal{M}$. Vose mentions that in general case, periodic orbits are possible but populations converge under repeated application of $\mathcal{M}^2$ and 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.

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

Following Vose's theorem, let $S_g = g\mathcal{R}/\{\mathbf{0}, g\}$, and let $|g|$ be the number of non zero bits in g.

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

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,

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

With above notations, limits can be expressed in Walsh basis by recursive equations

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

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

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 $|g|$ and performing complete indcution on $|g|$. If $|g| = 0$ then $g = 0$. Since $\widehat{\boldsymbol{p}^*}_0 = 2^{-\ell/2}$ for all distributions $\boldsymbol{p}$, the $\mathbf{0}$the components of the sequence $\mathcal{M}^n(\boldsymbol{p})$ are identical in the Walsh basis. Since $|x_0| = 2$ ($x_g = 2\widehat{\mathcal{M}}_{g,0}$ and $\widehat{\mathcal{M}}_{0,0} = 1$), $\widehat{\boldsymbol{p}^*}_g = \widehat{\boldsymbol{q}^*}_g = 2^{-\ell/2}$. Next, consider $|g| = 1$. $y_g = 0$ for $|g| = 0$ (noted from above). These two cases $|g| < 2|$ are base cases for complete induction on $|g|$. The inductive

hypothesis given by Vose is that for $|k| < |g|$, the kth component of $\mathcal{M}^n(\boldsymbol{p})$ in the Walsh basis converges to $\widehat{\boldsymbol{p}^*}_k$ or $\widehat{\boldsymbol{q}^*}_k$ as $n \to \infty$ through even or odd values respectively, and if $x_k \neq -1$ for all such k, then $\widehat{\boldsymbol{p}^*}_k = \widehat{\boldsymbol{q}^*}_k$. And computation of $y_g(z)$ involves only the kth components of z where $|k| < |g|$.

Vose gives a necessary and sufficient condition for the sequence

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

to converge to a periodic orbit as that for some g

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

3.2 Computation of Mutation and Crossover Distribution

Following algorithm installs values of mutation and crossover distributions that satisfies condition described by equation (3.3) for evolutionary sequence to converge in periodic orbits. Operations ($+$ and $\cdot$) acting on elements of $\mathcal{R}$ in this section below are component-wise addition and multiplication modulo 2. Let $\boldsymbol{\mu}_j$ and $\boldsymbol{\chi}_k$ represent mutation and crossover distributions respectively where $j, k \in \mathcal{R}$ and $U01()$ be random number between 0 and 1. For any g where $g \in \mathcal{R}$ and $g \neq 0$. For all $j \in \mathcal{R}$,

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

This installs some random values in some specific positions in $\boldsymbol{\mu}$ distribution array according to value of g and others set to 0. Normalization of $\boldsymbol{\mu}_j$ gives values for $\boldsymbol{\mu}$ distribution

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

such that

$$\sum_{j \in \mathcal{R}} \boldsymbol{\mu}_j = 1.$$

The values $\boldsymbol{\mu}_j$ satisfy condition (3.3) for $\boldsymbol{\mu}$ distribution.

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

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

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

$$\begin{aligned} \bar{g}k &= \bar{g}\bar{g}i \\ \bar{g}k &= \bar{g}i \\ \bar{g}k &= k \end{aligned}$$

For all $k \in \mathcal{R}$,

$$\begin{aligned} \boldsymbol{\chi}_k &= U01() \\ \boldsymbol{\chi}_{k+g} &= U01() \end{aligned}$$

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

$$\boldsymbol{\chi}_k = 0$$

for other values of k.

This installs some random values in some specific positions in array of $\boldsymbol{\chi}$ according to value of g and others set to 0. Normalization of $\boldsymbol{\chi}_k$ gives values for *chi* distribution

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

such that

$$\sum_{k \in \mathcal{R}} \boldsymbol{\chi}_k = 1.$$

The values $\boldsymbol{\chi}_k$ satisfy condition (3.3) for $\boldsymbol{\chi}$ distribution.

3.3 Initial Population

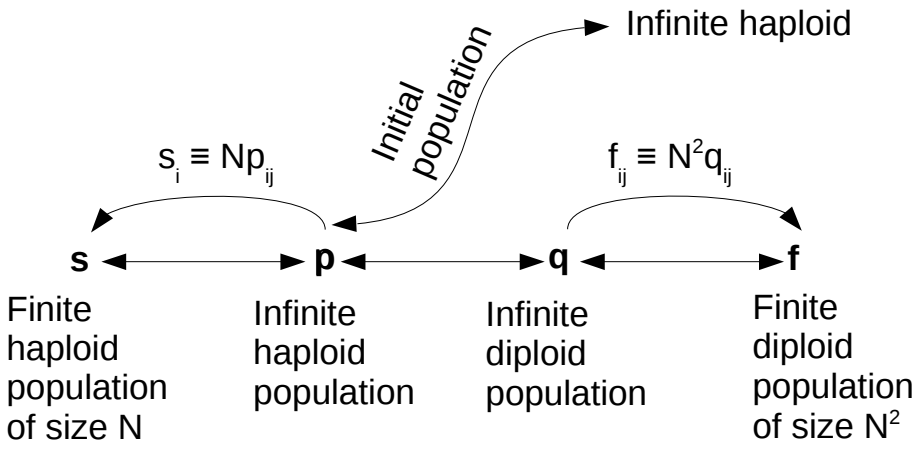


Figure 3.1: Initial population computation

Let finite haploid population $\boldsymbol{s}^n$, finite diploid population $\boldsymbol{f}^n$, infinite haploid population $\boldsymbol{p}^n$ and infinite haploid population $\boldsymbol{q}^n$ be considered with initial population $\boldsymbol{s}^0$, $\boldsymbol{f}^0$, $\boldsymbol{p}^0$, $\boldsymbol{q}^0$ respectively. To investigate oscillating behavior of infinite population evolutionary limits and finite population, same initial population is desired.

For a genome length ℓ, let $x = 2^\ell$ be number of possible strings in finite haploid population array $\boldsymbol{t}$ of population size N. Possible strings $\boldsymbol{t}_i$ are $0, 1, .., x-1$ where $i = 0, 1, .., N-1$. An arbitrary vector $\boldsymbol{f}$ of size x was considered where

$$\boldsymbol{r}_i = U01(); \qquad i = 0, 1, .., x-1$$

and U01() is random number between 0 and 1. Let $\boldsymbol{t}$ represent finite haploid population strings array.

$$\boldsymbol{t}_j = randp(\boldsymbol{r}); \qquad j = 0, .., N-1$$

where $\boldsymbol{t}_j$ is j^{th} population member and $randp(\boldsymbol{r})$ returns random index i in array $\boldsymbol{r}$ with probability $\boldsymbol{r}_i$.

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

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

Then infinite population vector $\boldsymbol{p}$ is calculated as

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

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

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

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

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

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

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

$$\boldsymbol{q}_{i,j} = \boldsymbol{p}_i \cdot \boldsymbol{p}_j; \qquad i = 0, .., x-1; \qquad j = 0, .., x-1.$$

Let $\boldsymbol{v}$ represent finite diploid population member array of size N^2 and $\boldsymbol{d}_{i,j}$ represent count of diploid member $\langle i, j \rangle$ in $\boldsymbol{v}$. Then $\boldsymbol{v}$ can be filled with population member to match initial population vector $\boldsymbol{p}$ generating diploid members such that

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

Finite diploid population vector $\boldsymbol{f}$ can be obtained from finite diploid population member array $\boldsymbol{v}$ using

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

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

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

3.4 Oscillation

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

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 $\boldsymbol{q}_d^*$ as

$$\begin{aligned} \boldsymbol{p}^*_{d\langle\gamma_0,\gamma_1\rangle} &= \boldsymbol{p}^*_{h\gamma_0}\boldsymbol{p}^*_{h\gamma_1} \\ \boldsymbol{q}^*_{d\langle\gamma_0,\gamma_1\rangle} &= \boldsymbol{q}^*_{h\gamma_0}\boldsymbol{q}^*_{h\gamma_1} \end{aligned}$$

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

For a genome length ℓ, same initial population (calculated as described in (3.3)) was used for infinite population and all sizes of finite population conisdered. Genome lengths $\ell = \{8, 10, 12, 14\}$ were used. Base population size of $N_0 = 64$ was considered for finite haploid case and different population sizes $N = \{1N_0^2, 10N_0^2, 20N_0^2\}$ were considered for plotting graphs. 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. Distance data of finite population to infinite population were also plotted.

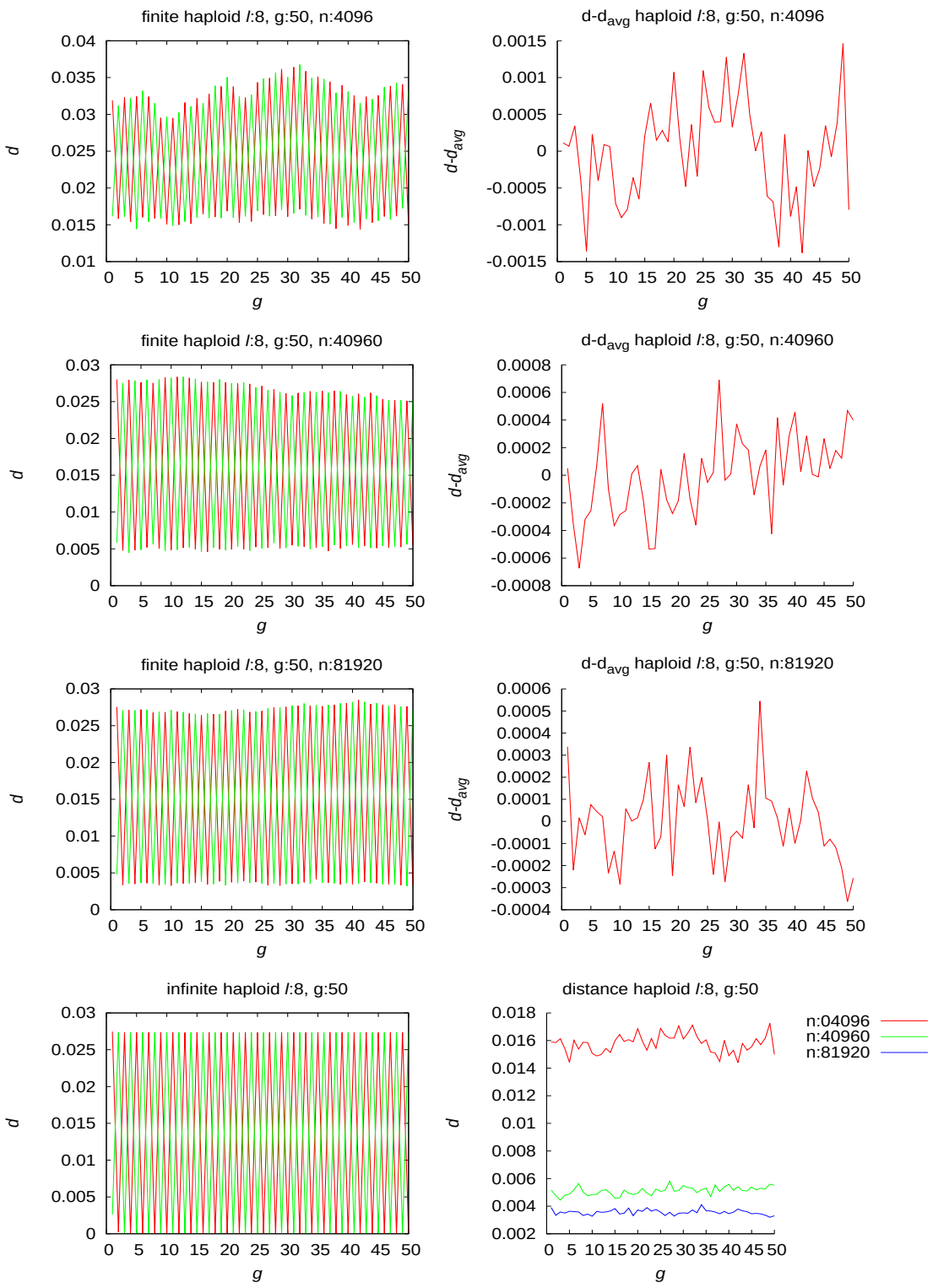


Figure 3.2: Infinite and finite haploid population oscillation behavior for genome length $\ell = 8$ (bits): In left column, d is distance of finite population of size n or infinite population to limits for g generations. In right column, d is distance of finite population to infinite population for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

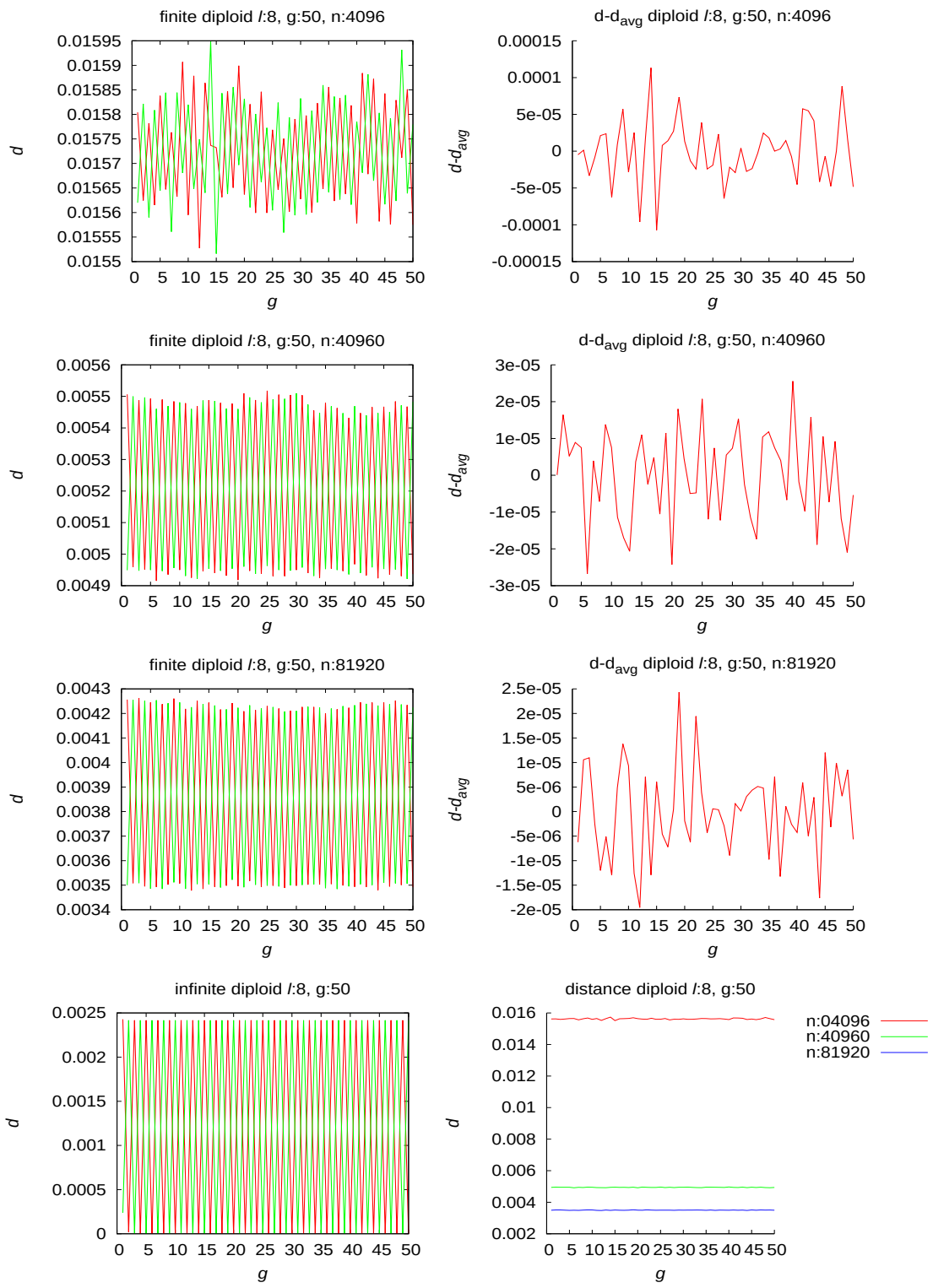


Figure 3.3: Infinite and finite diploid population oscillation behavior for genome length $\ell = 8$ (bits): In left column, d is distance of finite population of size n or infinite population to limits for g generations. In right column, d is distance of finite population to infinite population for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

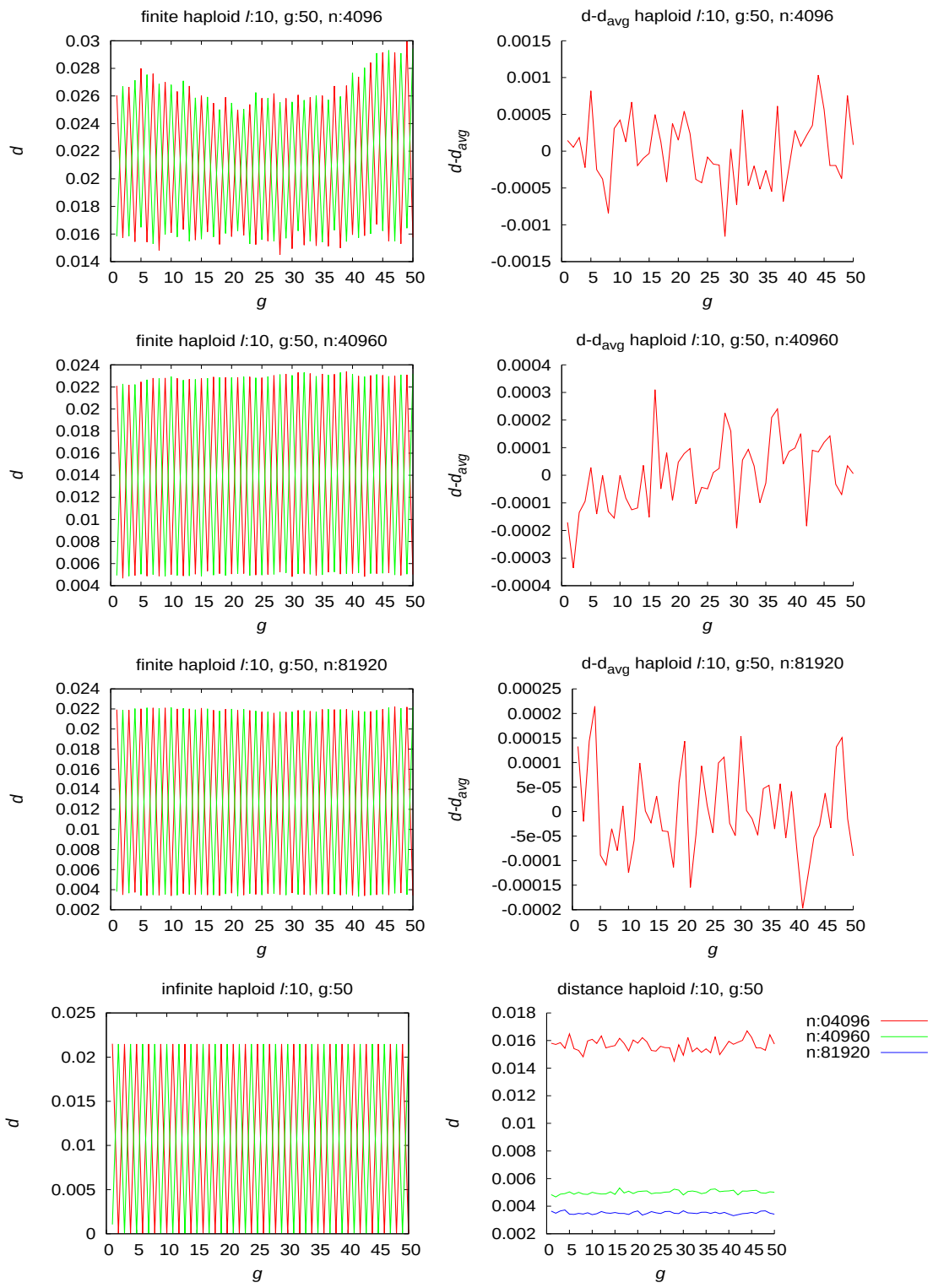


Figure 3.4: Infinite and finite haploid population oscillation behavior for genome length $\ell = 10$ (bits): In left column, d is distance of finite population of size n or infinite population to limits for g generations. In right column, d is distance of finite population to infinite population for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

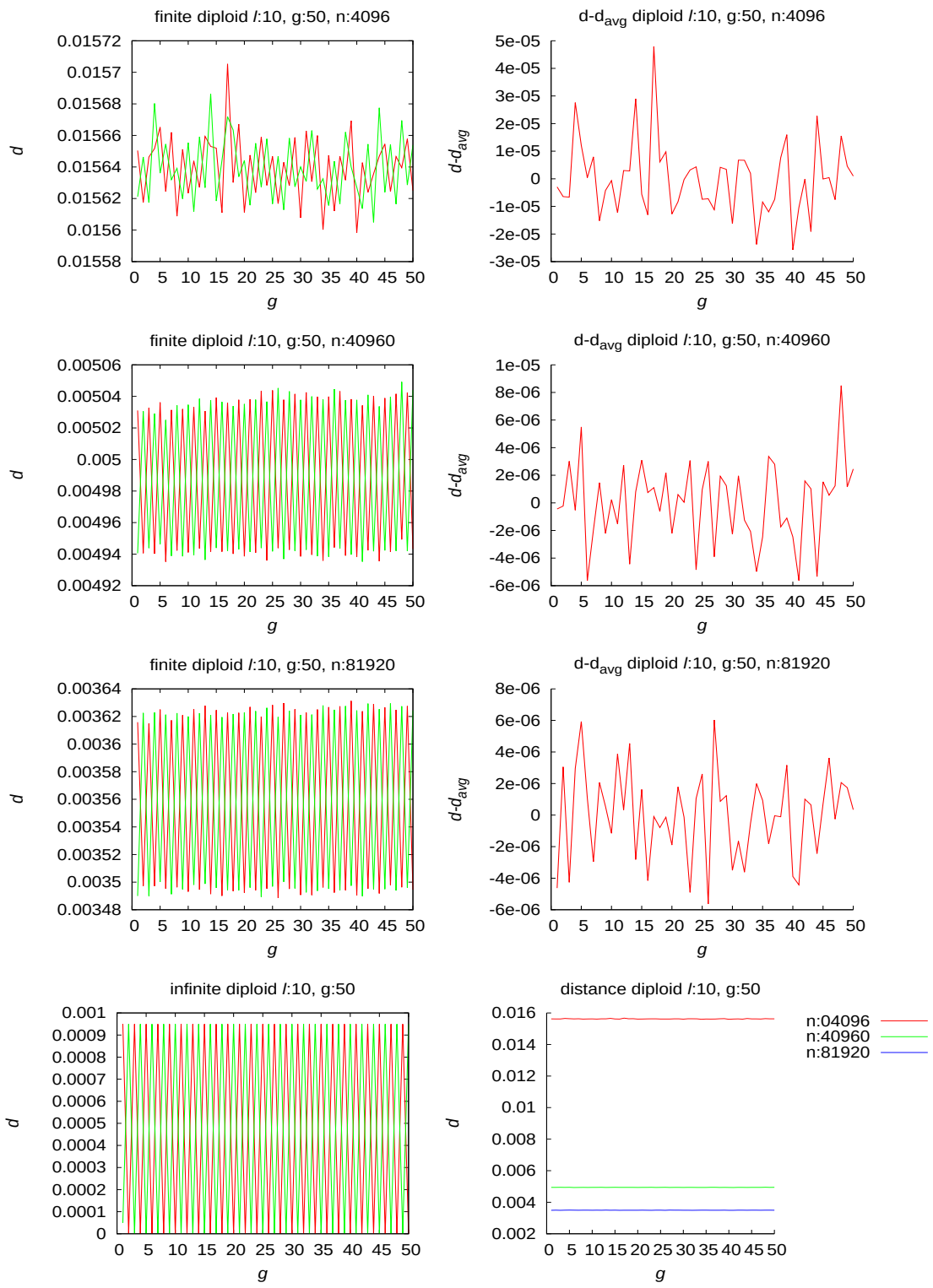


Figure 3.5: Infinite and finite population oscillation behavior for genome length $\ell = 10$ (bits): In left column, d is distance of finite population of size n or infinite population to limits for g generations. In right column, d is distance of finite population to infinite population for g generations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

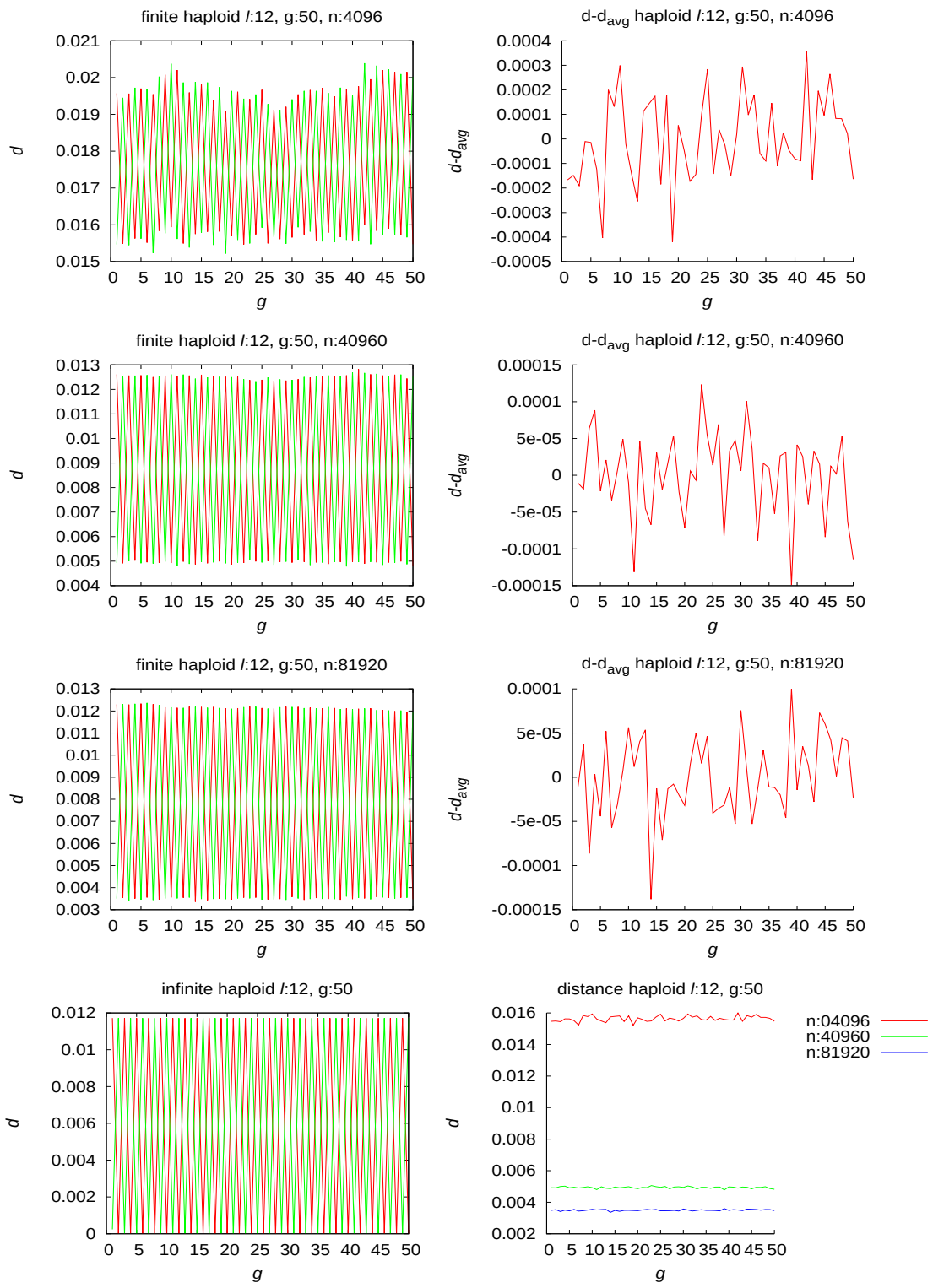


Figure 3.6: Infinite and finite haploid population oscillation behavior for genome length $\ell = 12$ (bits): In left column, d is distance of finite population of size n or infinite population to limits for g generations. In right column, d is distance of finite population to infinite population for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

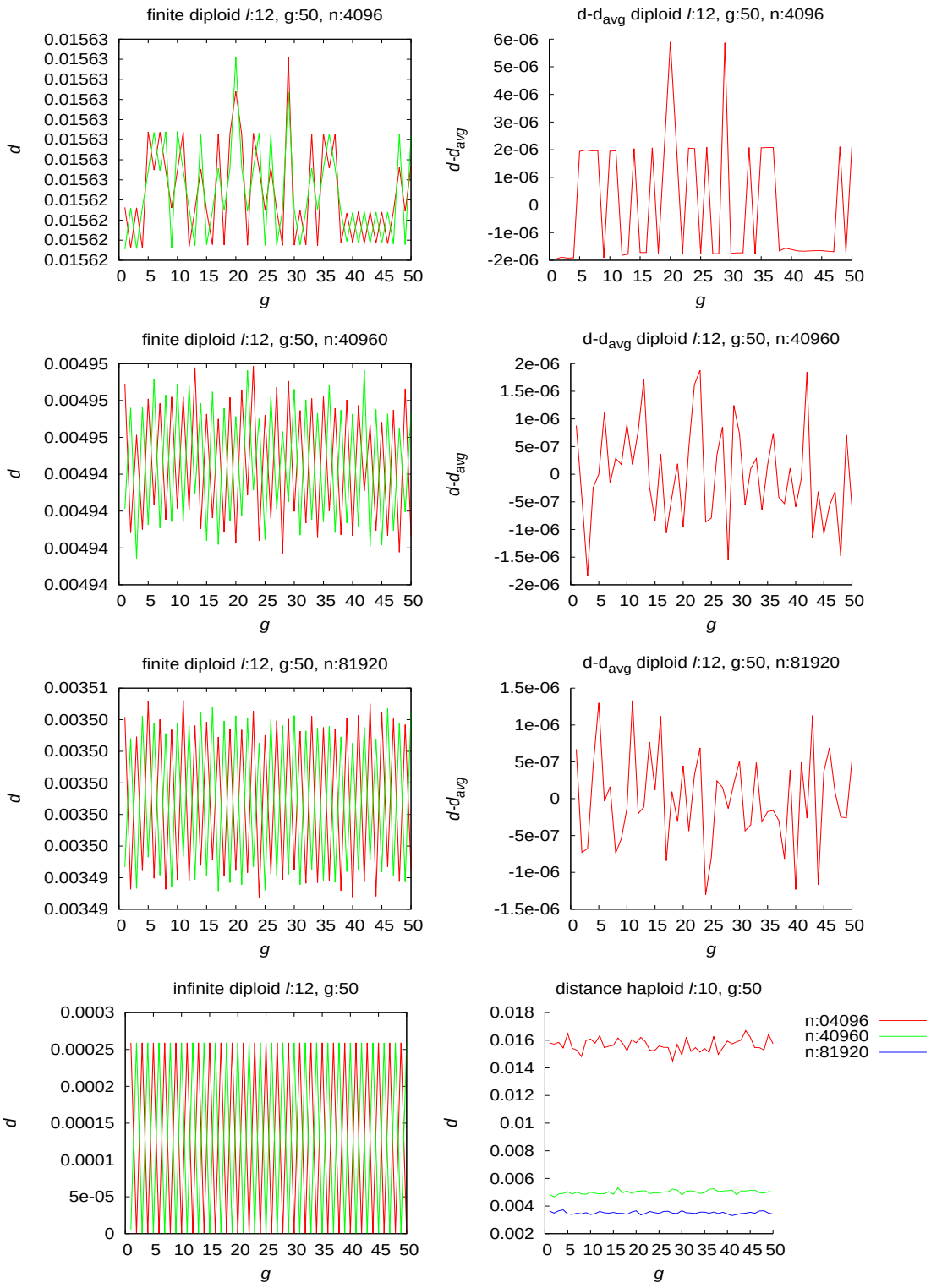


Figure 3.7: Infinite and finite diploid population oscillation behavior for genome length $\ell = 12$ (bits): In left column, d is distance of finite population of size n or infinite population to limits for g generations. In right column, d is distance of finite population to infinite population for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

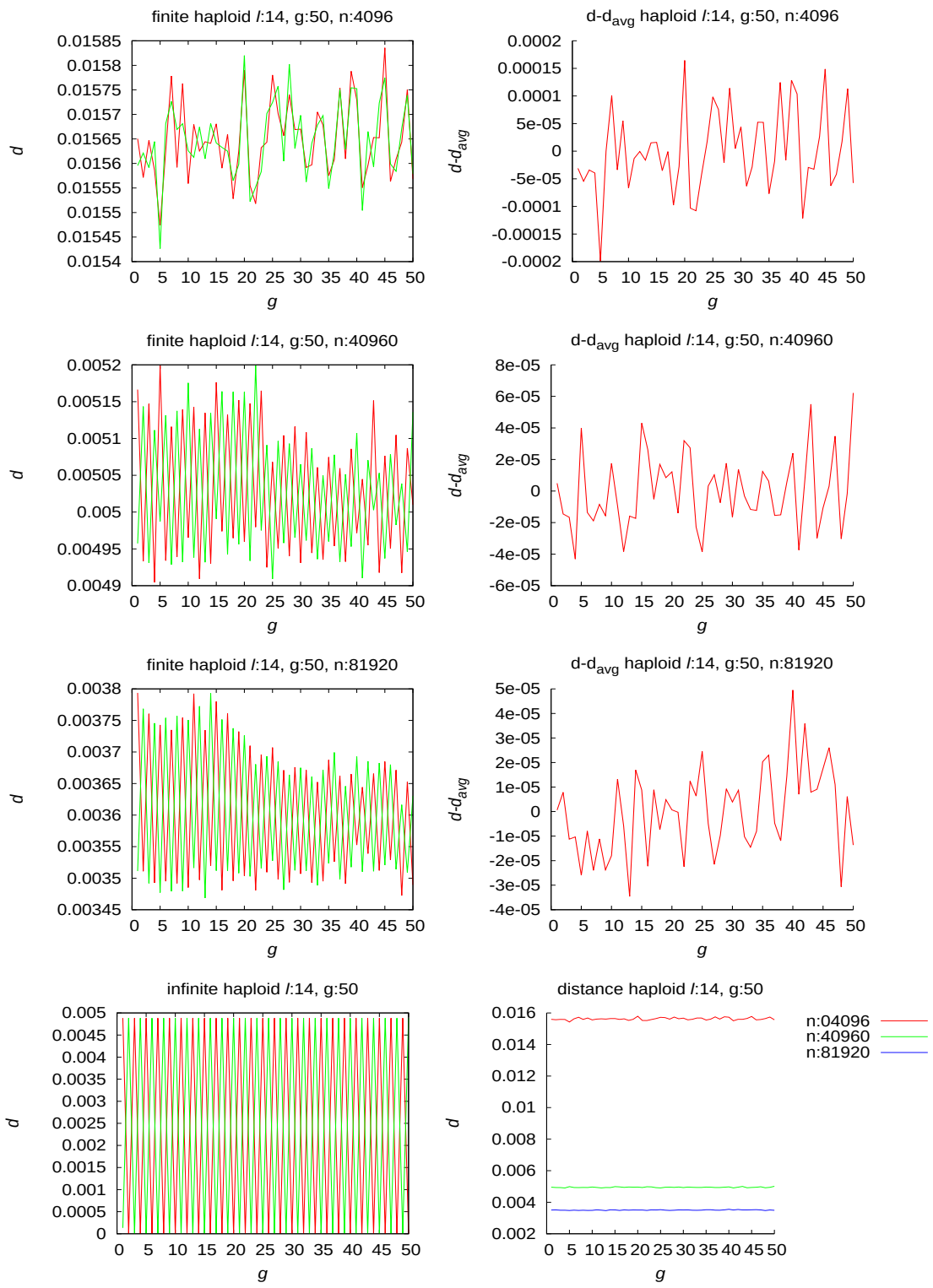


Figure 3.8: Infinite and finite haploid population oscillation behavior for genome length $\ell = 14$ (bits): In left column, d is distance of finite population of size n or infinite population to limits for g generations. In right column, d is distance of finite population to infinite population for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

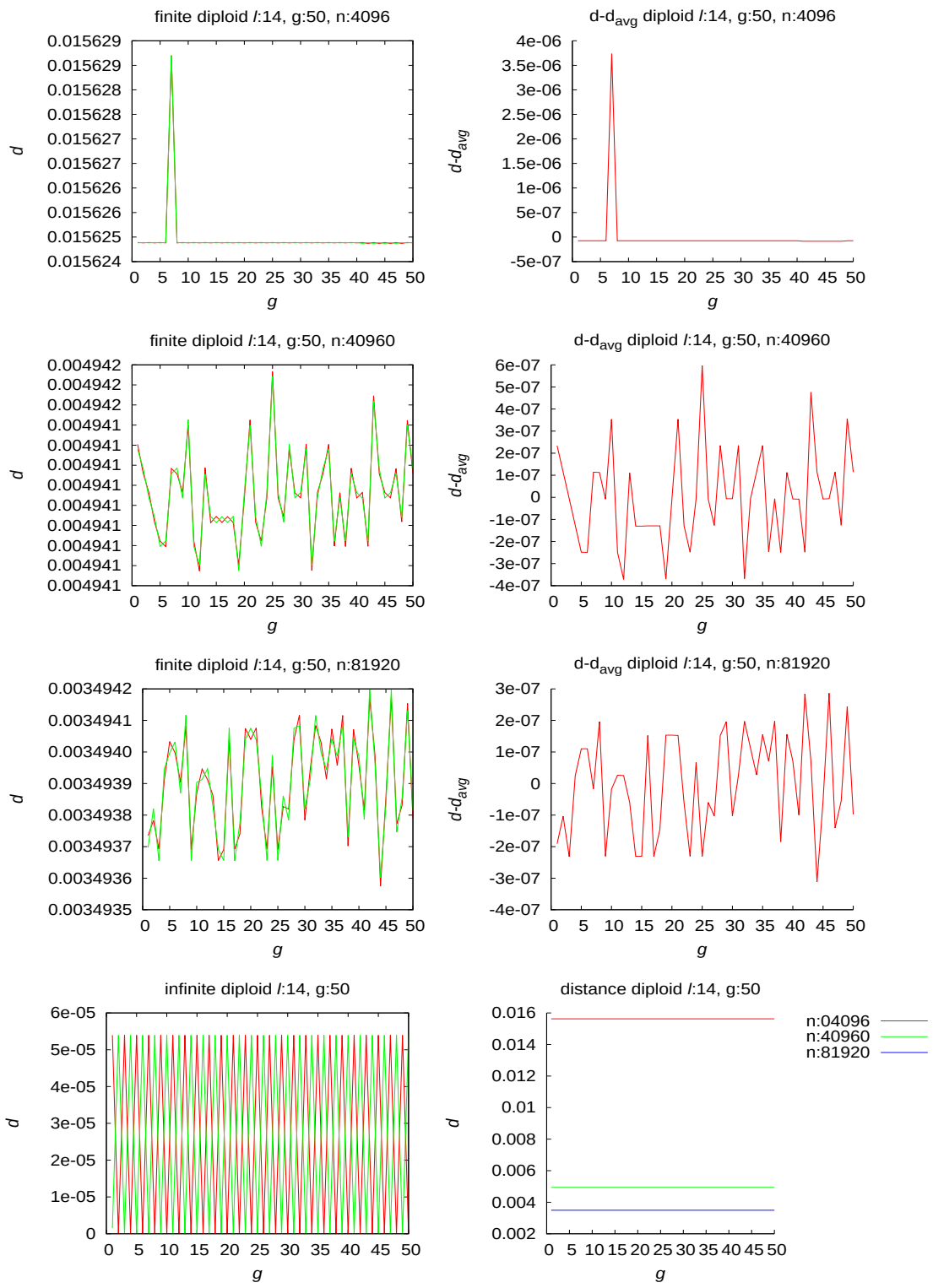


Figure 3.9: Infinite and finite diploid population oscillation behavior for genome length $\ell = 14$ (bits): In left column, d is distance of finite population of size n or infinite population to limits for g generations. In right column, d is distance of finite population to infinite population for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

Figures 3.2, 3.3, 3.4, 3.5, 3.6, 3.7, 3.8 and 3.9 arranged by genome length ℓ and sub-figures within each figures arranged by population size (N) for finite population for haploid and diploid population depicts oscillating behavior of both infinite and finite population when necessary and sufficient condition 3.3 is met. Oscillation in finite population in both haploid and diploid case simulation became sharper with increased population size as expected. As size of finite population increased, oscillation behavior depicted by finite population grew analogous to oscillation behavior depicted by infinite population.

Graphs of difference in distance (d) and average distance (d_{avg}) were plotted for both haploid and diploid cas on right side of oscillation graphs in above figures where d is distance of finite population to infinite population and d_{avg} is average of distance from 1^{st} to 50^{th} generation. For a given genome length ℓ and haploid or diploid case, single graph for distances of different finite population sizes ($N = 1N_0^2, 10N_0^2, 20N_0^2$) were plotted. The resulting graphs showed distance decreased as population size increased which is expected and in congruence with results from section 2.1. Graphs show curve of distance of finite population to infinite population smoothens as population size increased. The graphs of $d - d_{avg}$ shows decrease in amplitude of ripples as population size increases.

Numerator $\sqrt{(1 - \|\mathcal{G}(\boldsymbol{p})\|^2)}$ in equation 2.21 is calculated using initial population and is ≈ 1. So from 2.21, the expected single step distance between finite and infinite population, d, can be approximated as

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

where N is size of finite population. Then mathematical expection of single step distance for finite population size $N = \{4096, 40960, 81920\}$ considered for plotting is shown in following table.

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

N	4096	40960	81920
d	0.015625	0.004941	0.003494

Data obtained from distances of finite population of size $N = \{4096, 40960, 81920\}$ to infinite population obtained from simulation for oscillation for both haploid and dioploid case with genome length $\ell = \{8, 10, 12, 14\}$ are tabulated below.

Table 3.2: Experimental distance measured for oscillation: N is finite population size, ℓ is genome length and $\{d', d'', d'''\}$ are distances to infinite population from finite population of sizes $\{4096, 40960, 81920\}$

case	ℓ	d'	d''	d'''
haploid	8	0.015785	0.005123	0.003559
	10	0.015663	0.005007	0.003511
	12	0.015638	0.004933	0.003497
	14	0.015626	0.004948	0.003503
diploid	8	0.015623	0.004942	0.003498
	10	0.015624	0.004941	0.003494
	12	0.015625	0.004941	0.003494
	14	0.015625	0.004941	0.003494

From table 3.2, averaged distances obtained for finite population size $N = \{4096, 40960, 81920\}$ are $\{0.015651, 0.004972, 0.003506\}$ respectively. The results from table 3.2 shows the distance between finite population and infinite population, although measured over number of generations, follows closely the expected single step distance between finite and infinite population given by table 3.1 and the distance decreased as $1/\sqrt{N}$.

3.5 Violation

The results showed when $\boldsymbol{\chi}$ and $\boldsymbol{\mu}$ distributions satisfies (3.3), oscillation occurs in both infinite and finite population. Error $\boldsymbol{\epsilon}$ was introduced to $\boldsymbol{\mu}$ distribution and $\boldsymbol{\chi}$ distribution such that (3.3) did not satisfy anymore and $x_g \neq 1$ for all g (x_g and g defined in 3.1) so that $\boldsymbol{p}^* = \boldsymbol{q}^*$.

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

$\boldsymbol{\mu}$ distribution was treated with $\boldsymbol{\epsilon}$ such that

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

So that sum of $\boldsymbol{\mu}$ distribution becomes,

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

Then set

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

$\boldsymbol{\chi}$ distribution was treated with $\boldsymbol{\epsilon}$ such that

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

So that

$$\boldsymbol{\chi}_i + \boldsymbol{\chi}_{i+g} = 1 - \boldsymbol{\epsilon}; \qquad g \text{ is defined in section 3.1}$$

Then j is chosen where $\boldsymbol{\chi}_j = 0$ and set $\boldsymbol{\chi}_j = \boldsymbol{\epsilon}$.

Simulations were run again with violations in (3.3) implemented. Genome lengths $\ell = \{8, 10, 12, 14\}$ were considered. Different finite hapliod population sizes $N = \{1N_0^2, 10N_0^2, 20N_0^2\}$ were considered.

Let $\boldsymbol{p}1^*$ and $\boldsymbol{q}1^*$ be evolutionary limits with violation, then $\boldsymbol{p}1^* = \boldsymbol{q}1^* = \boldsymbol{z}^*$; $\boldsymbol{z}^*$ is limit when there is violation. The distances of p^n and $\boldsymbol{s}^n$ to $\boldsymbol{z}^*$ were plotted and the distances of $\boldsymbol{q}^n$ and $\boldsymbol{f}^n$ to $\boldsymbol{z}^*$ were plotted from 1^{st} to 50^{th} generations. The distances of population to evolutionary limits that would be without violation in $\boldsymbol{\mu}$ and $\boldsymbol{\chi}$ were also plotted from 1^{st} to 50^{th} generations.

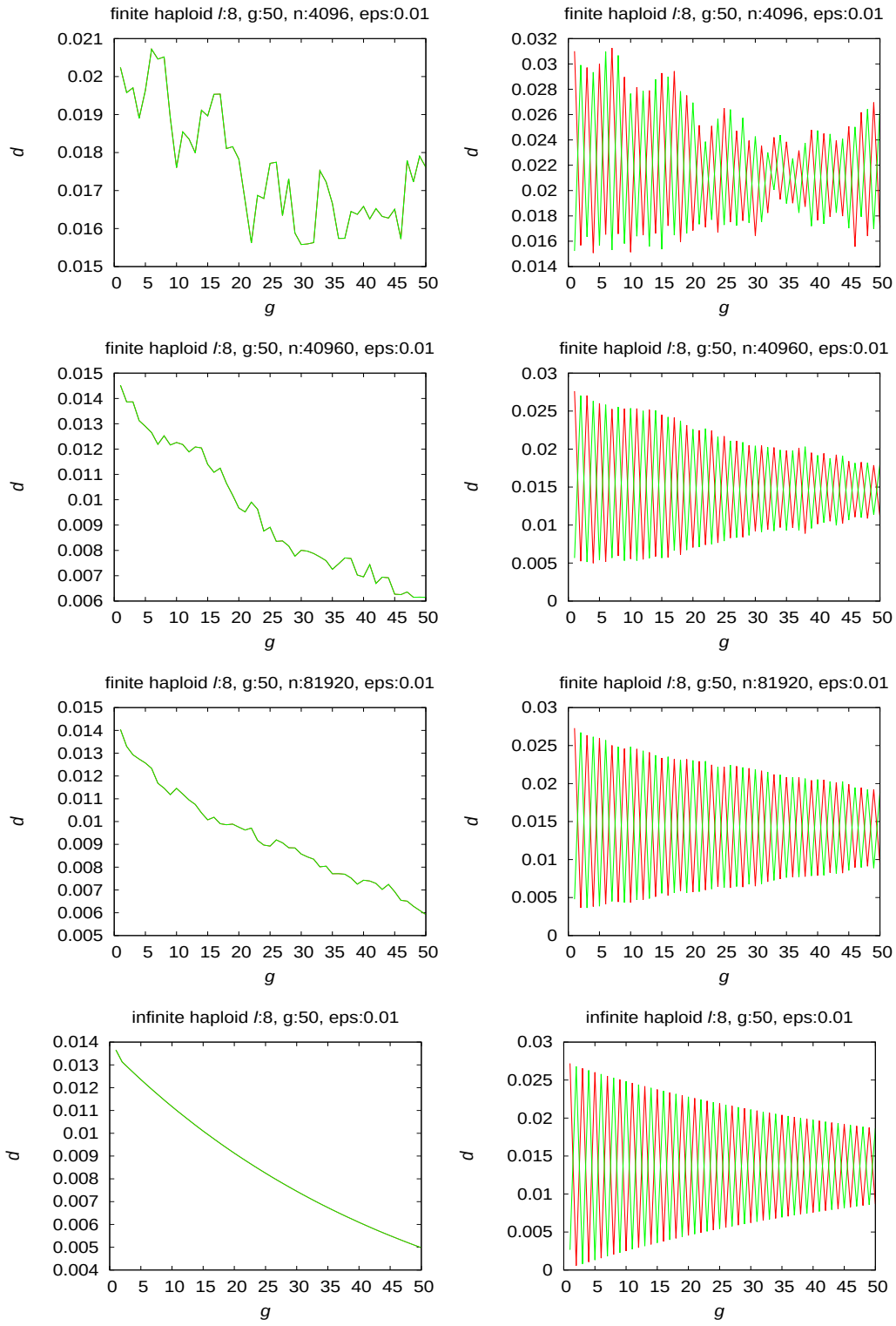


Figure 3.10: Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 8$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite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

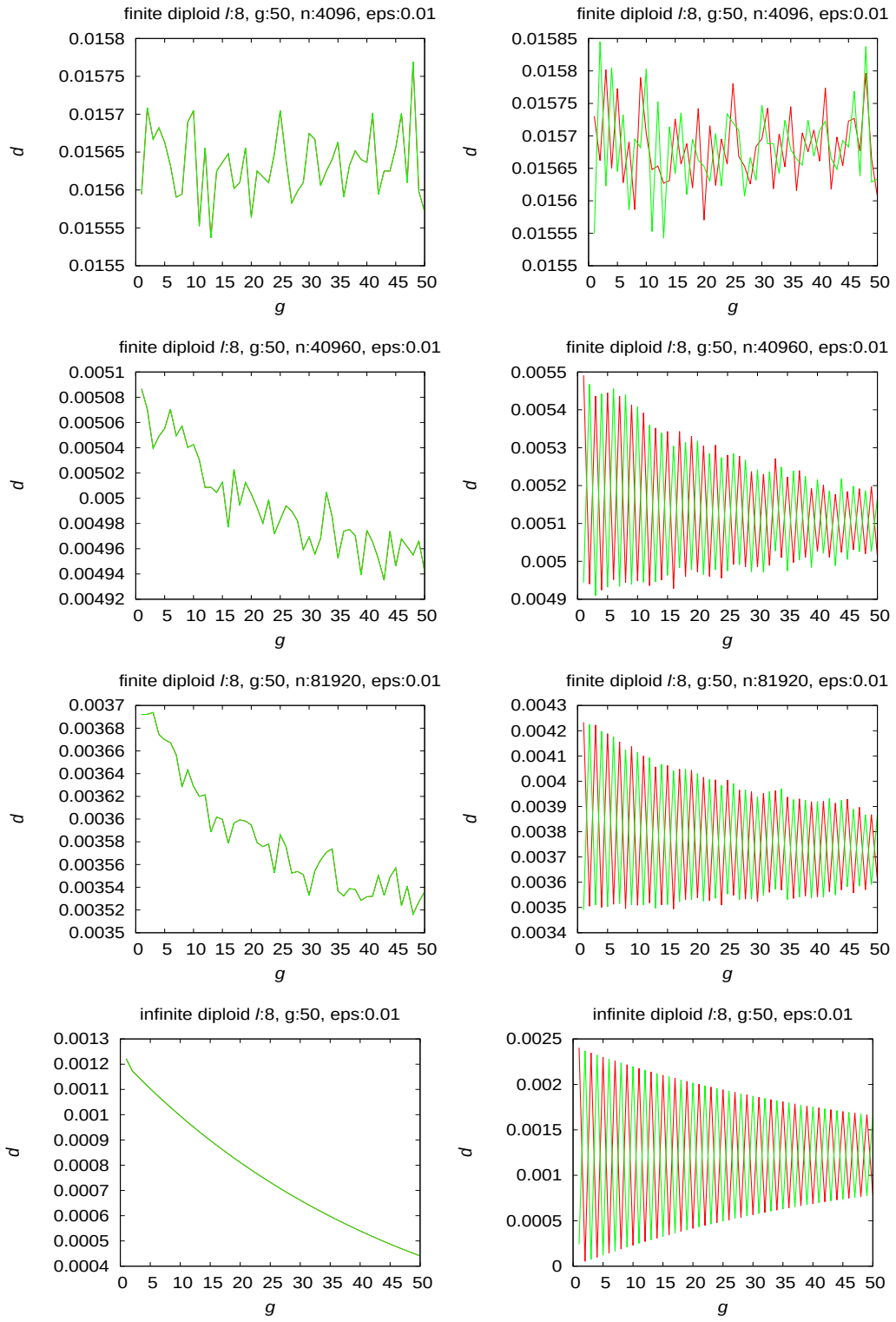


Figure 3.11: Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 8$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

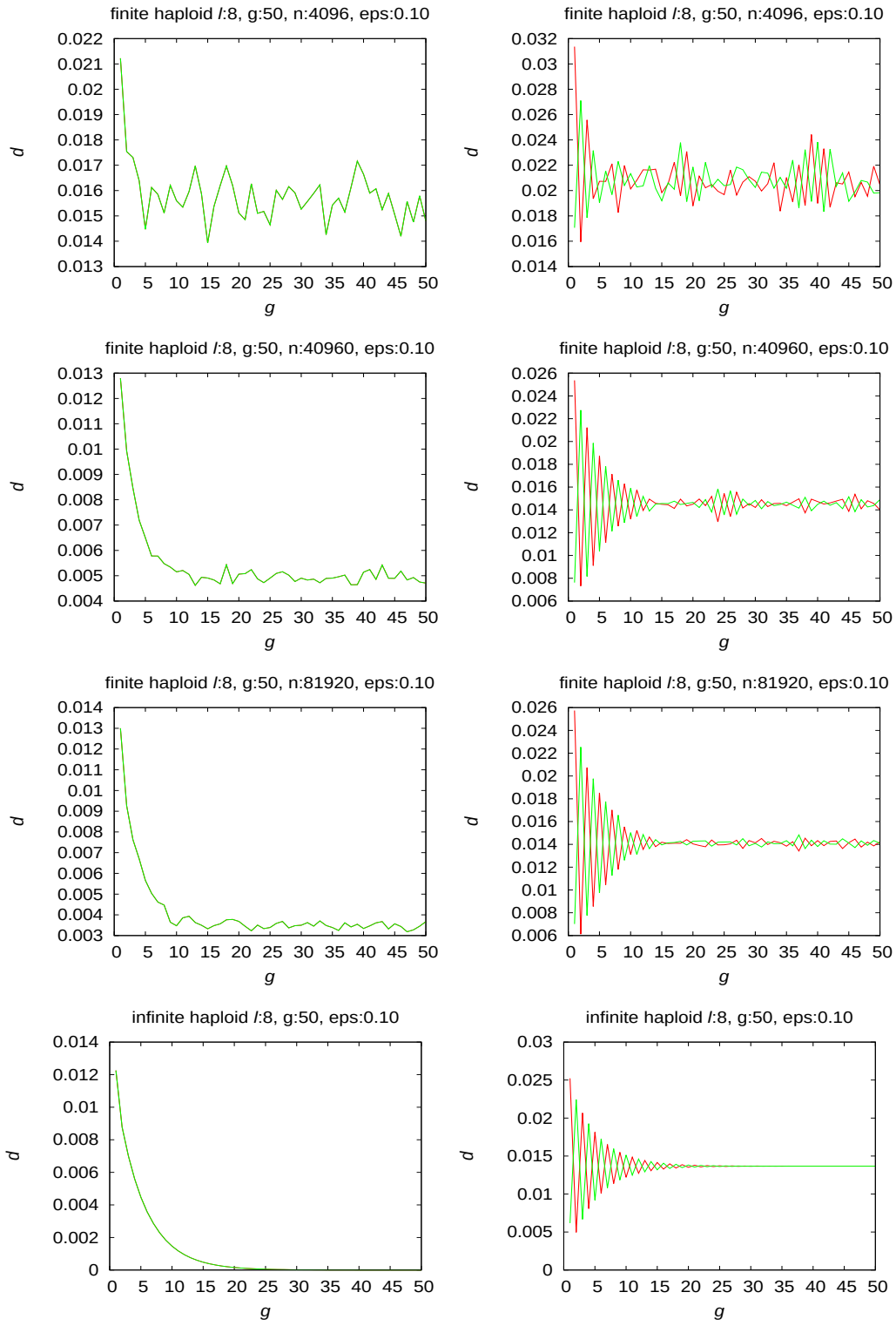


Figure 3.12: Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 8$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

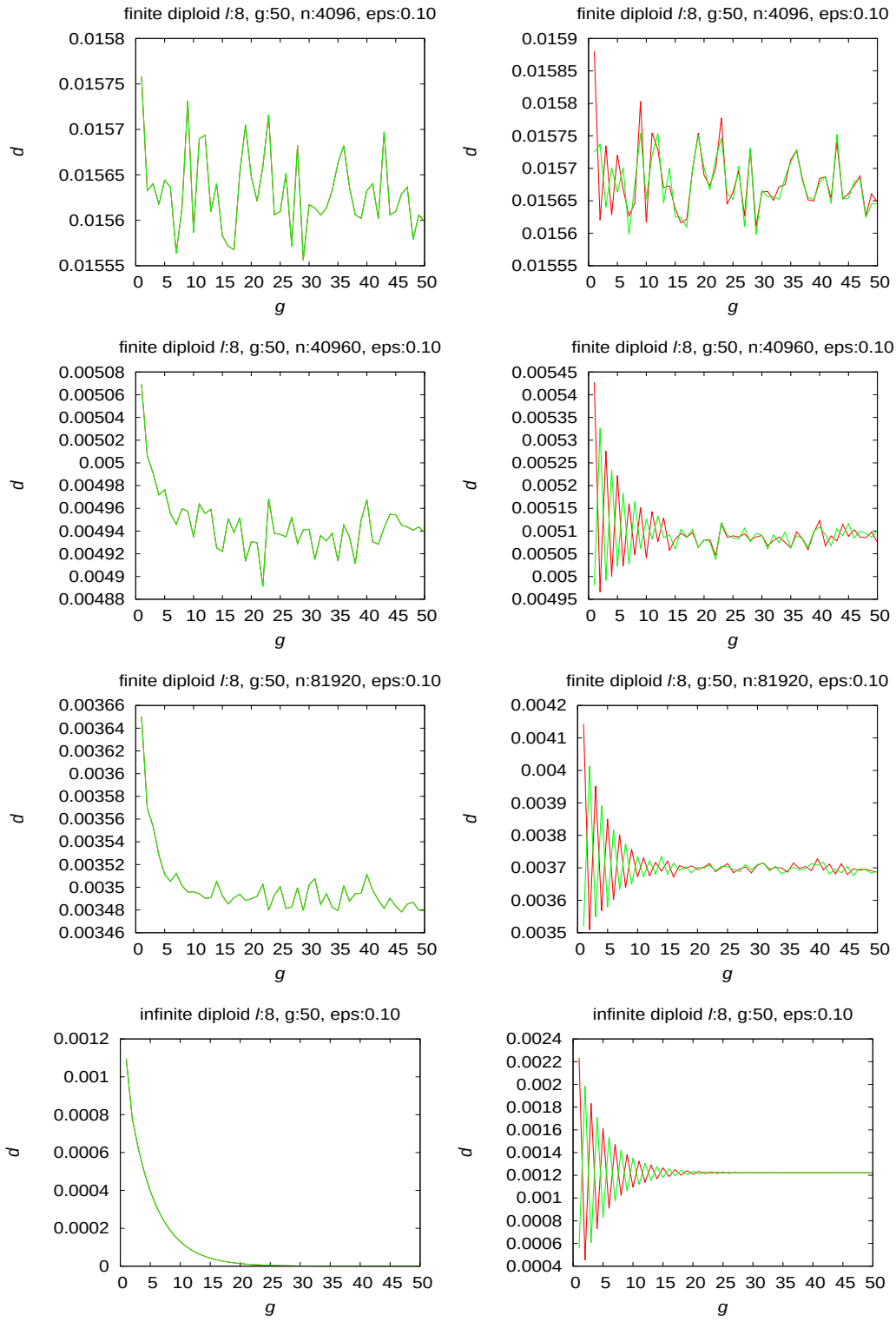


Figure 3.13: Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 8$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

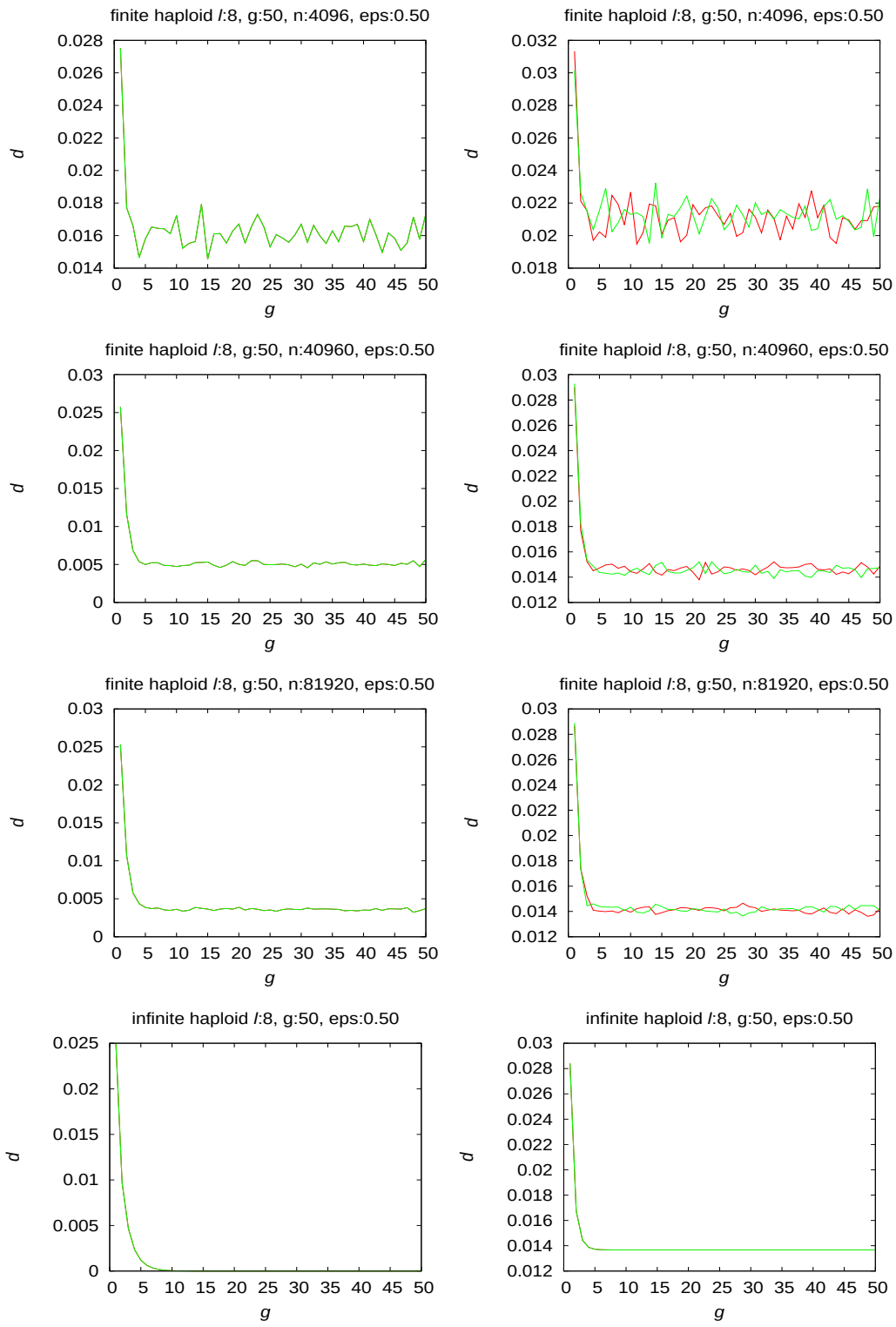


Figure 3.14: Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 8$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

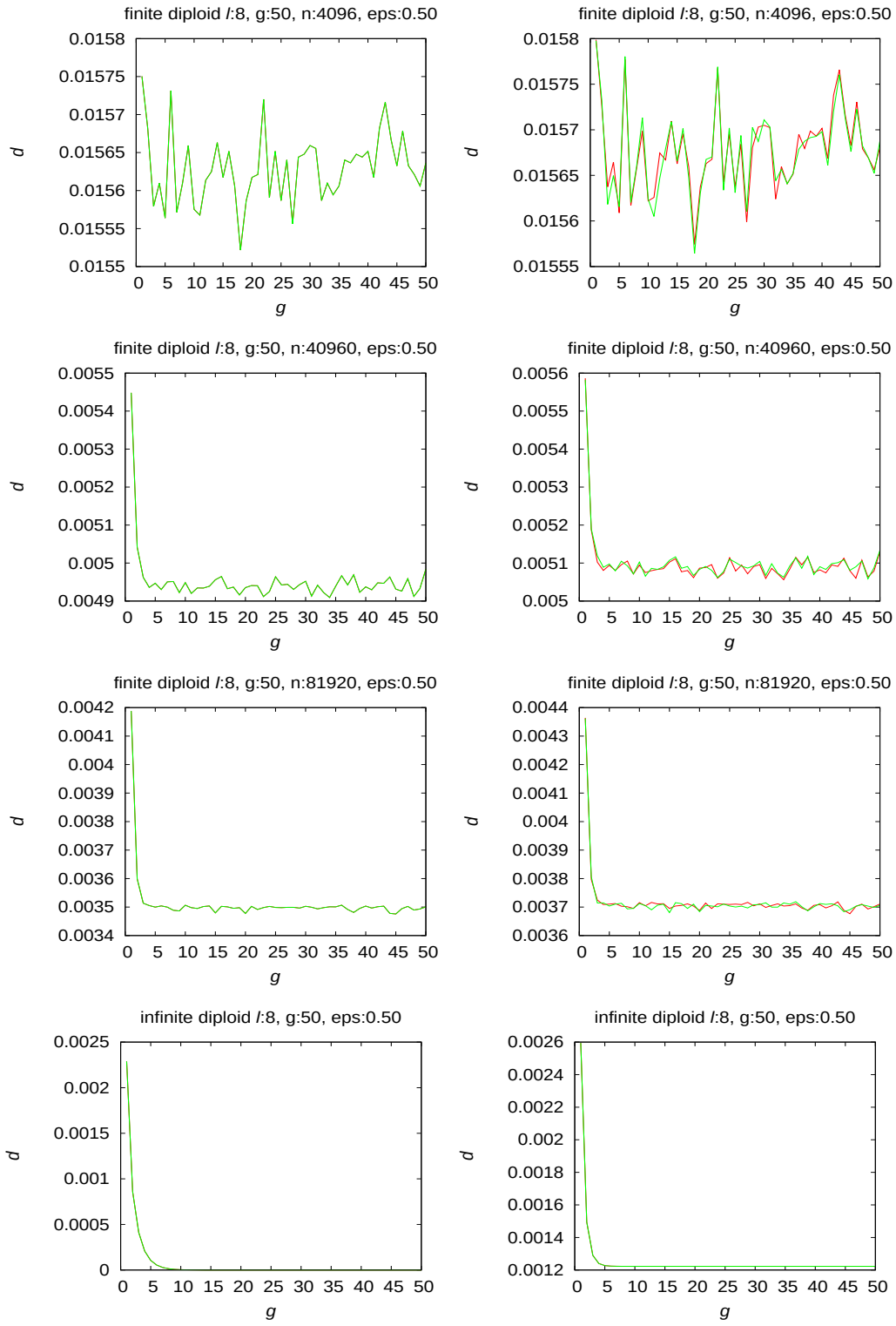


Figure 3.15: Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 8$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

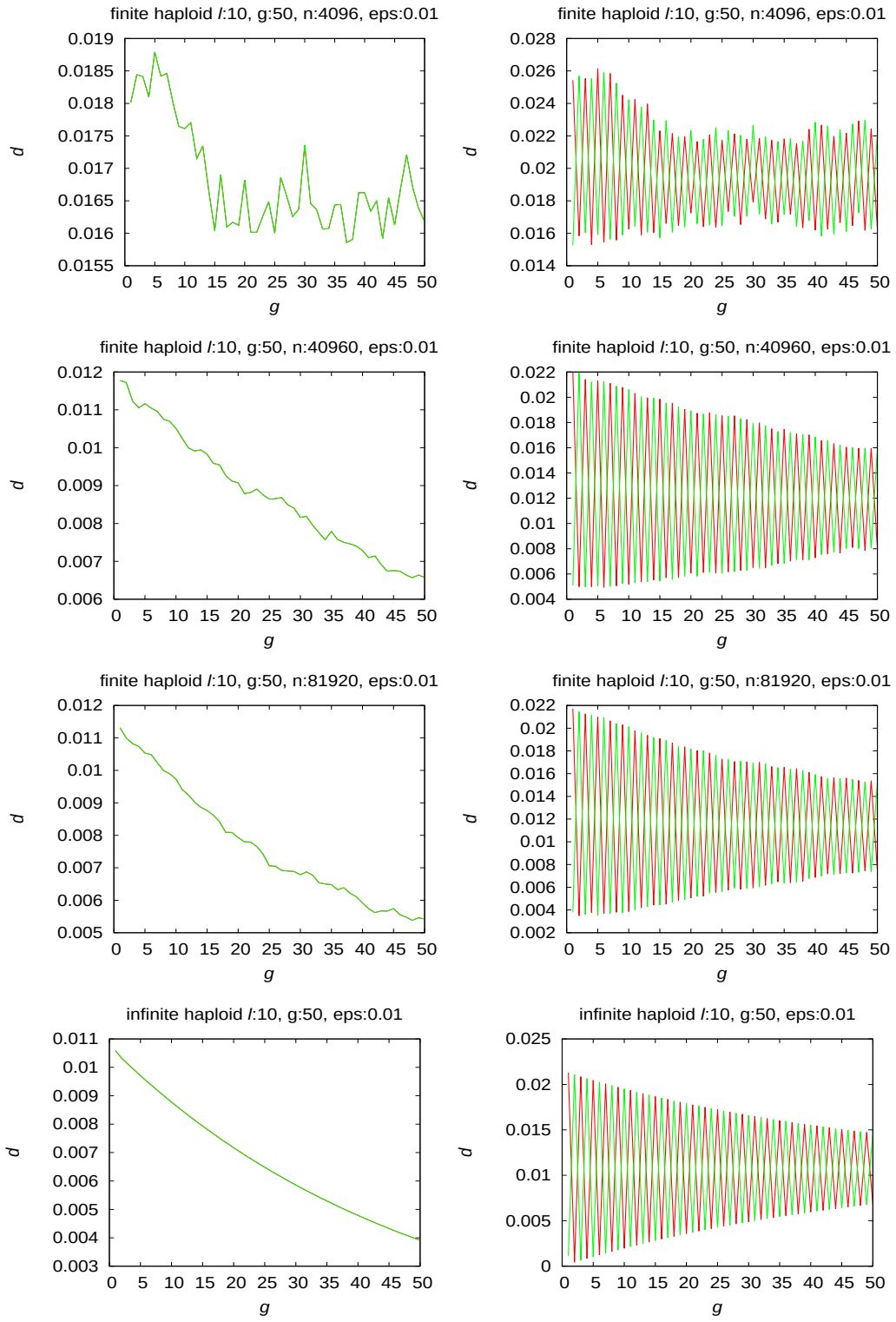


Figure 3.16: Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 10$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite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

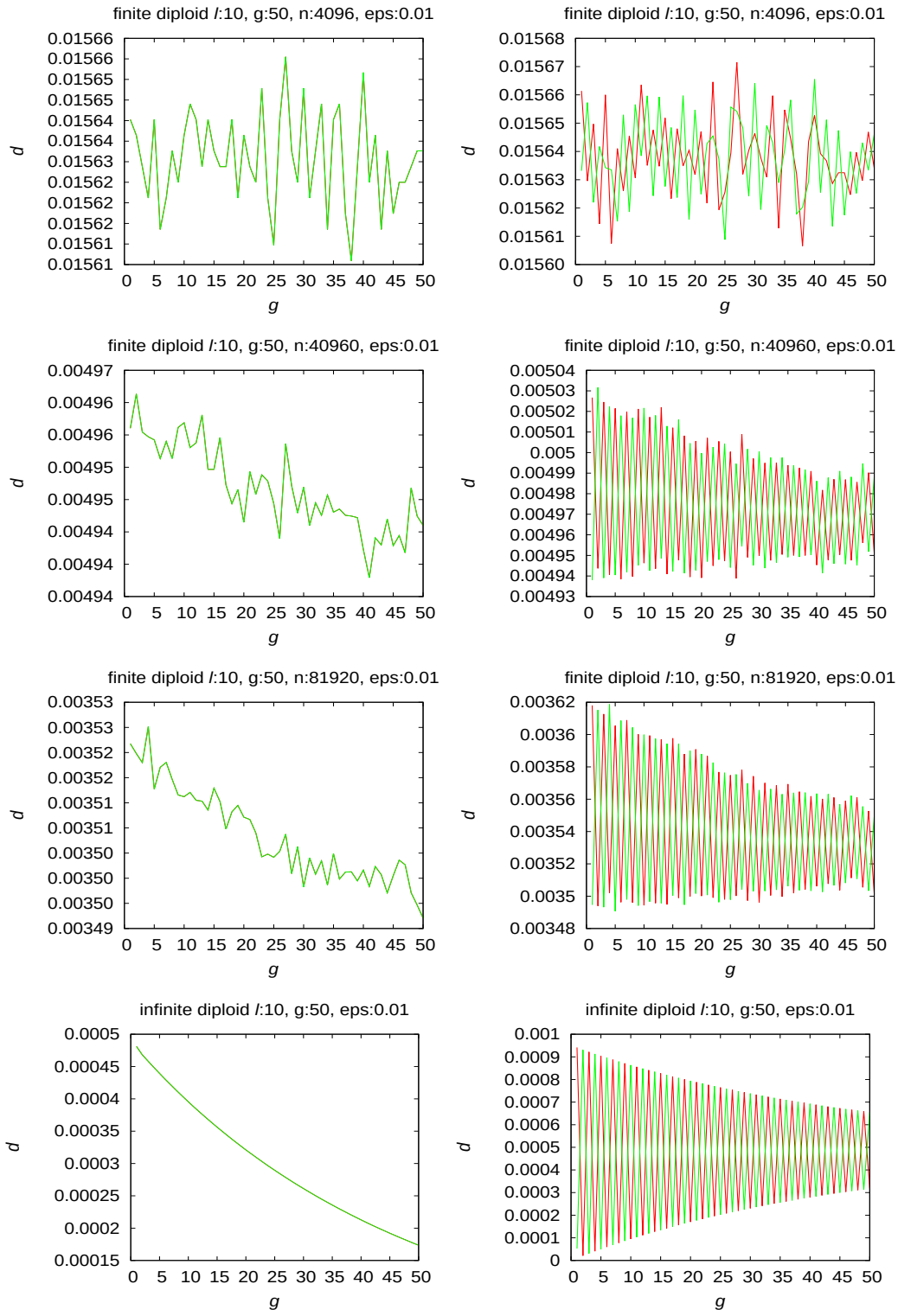


Figure 3.17: Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 10$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

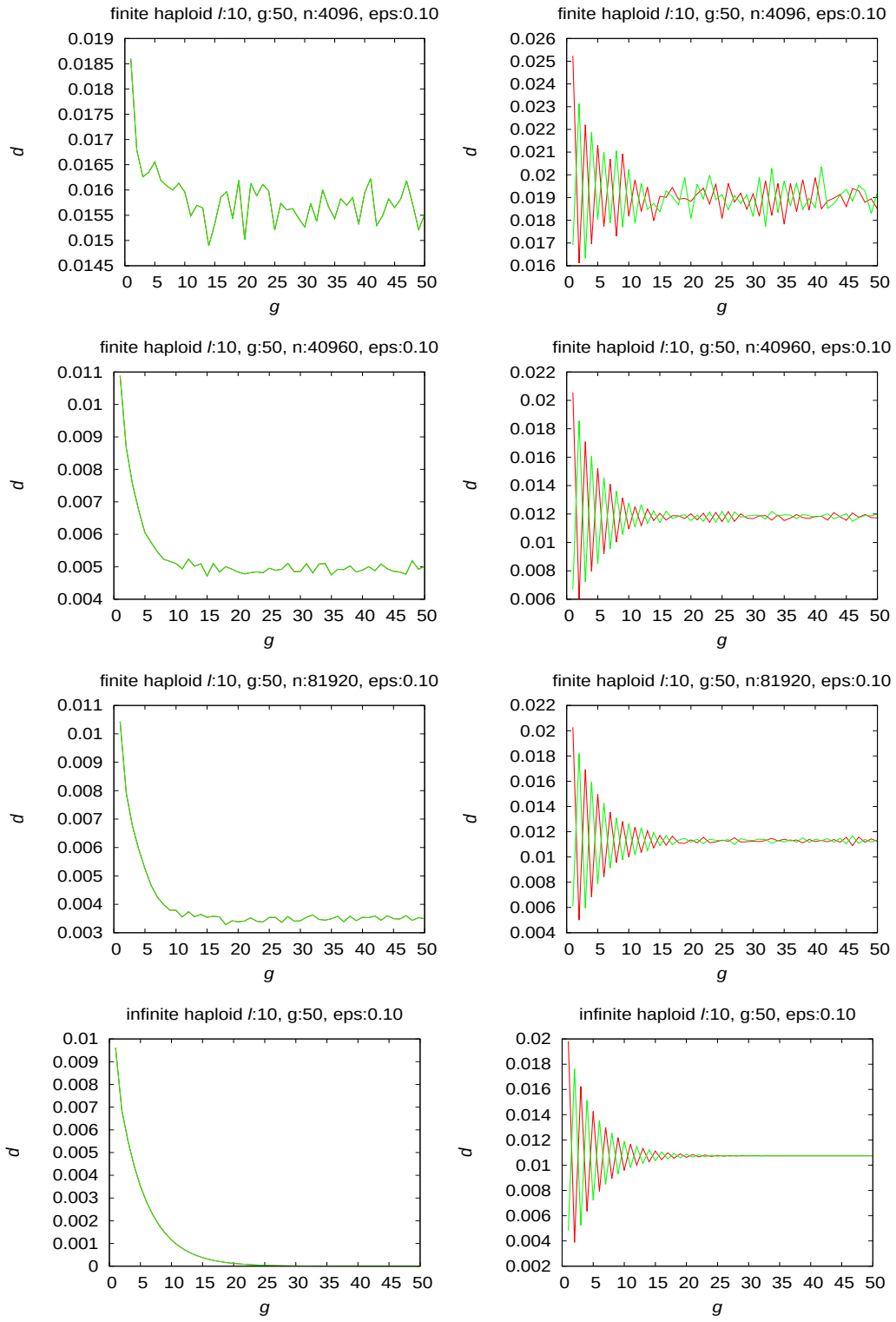


Figure 3.18: Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 10$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

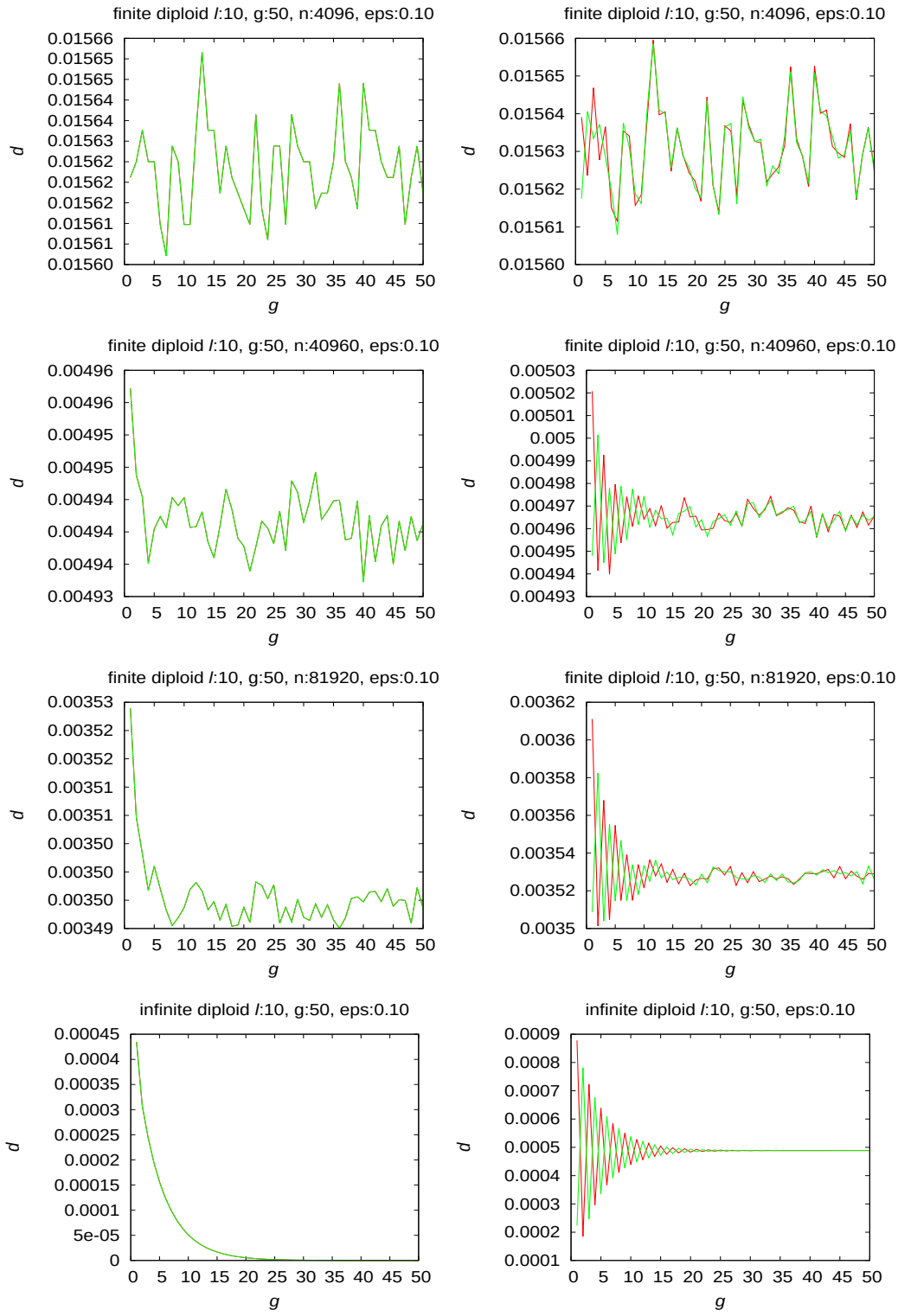


Figure 3.19: Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 10$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

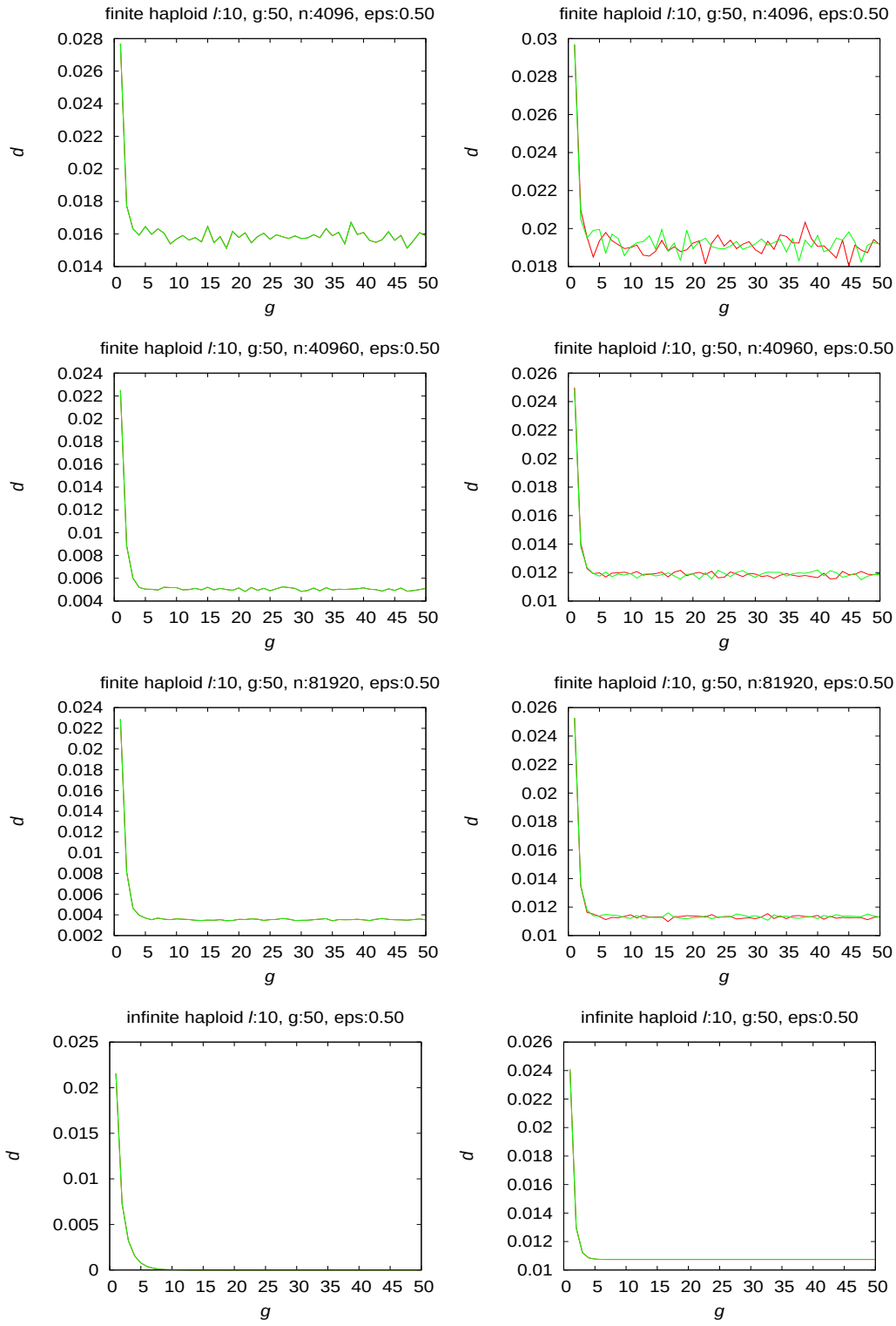


Figure 3.20: Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 10$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

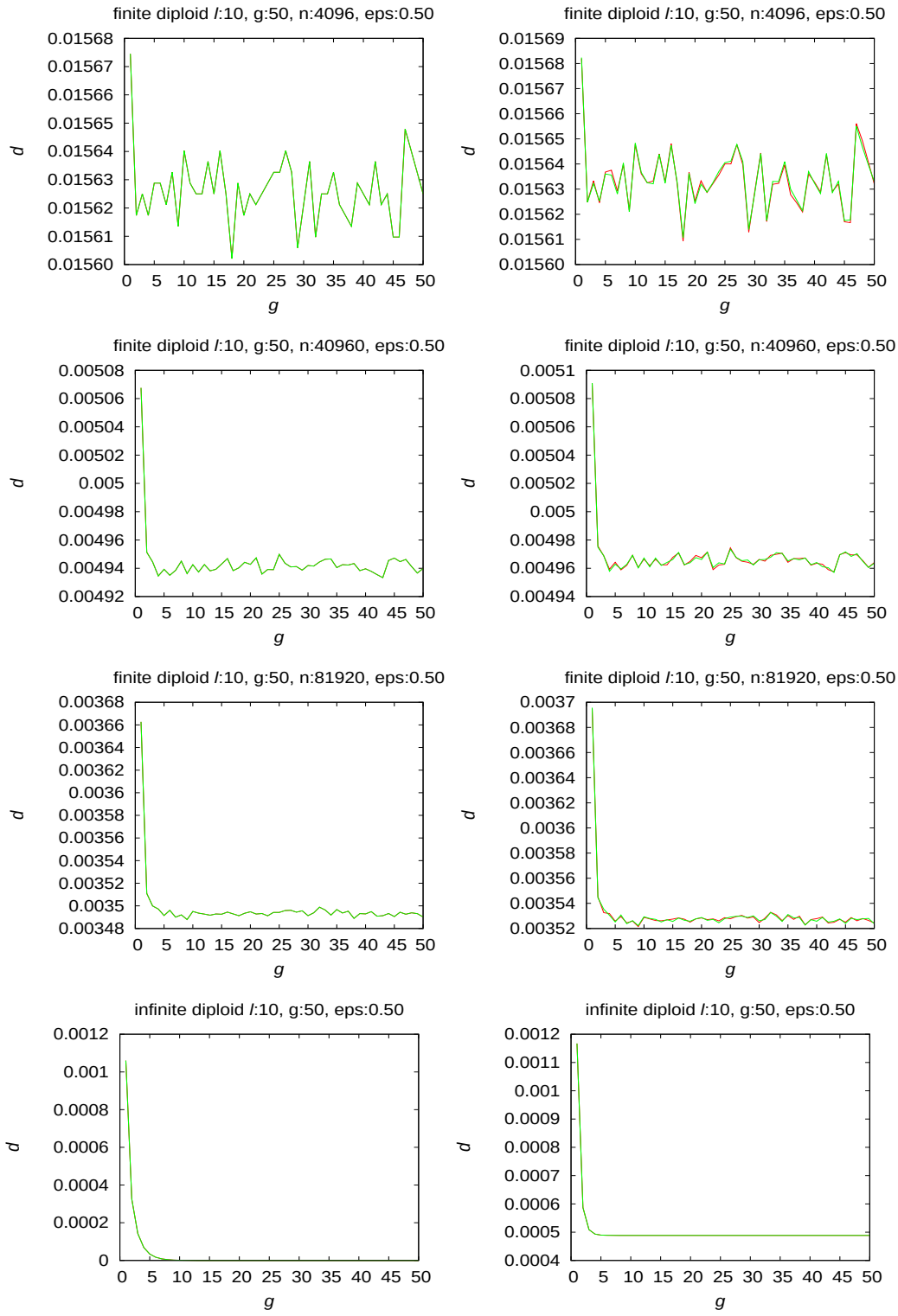


Figure 3.21: Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 10$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

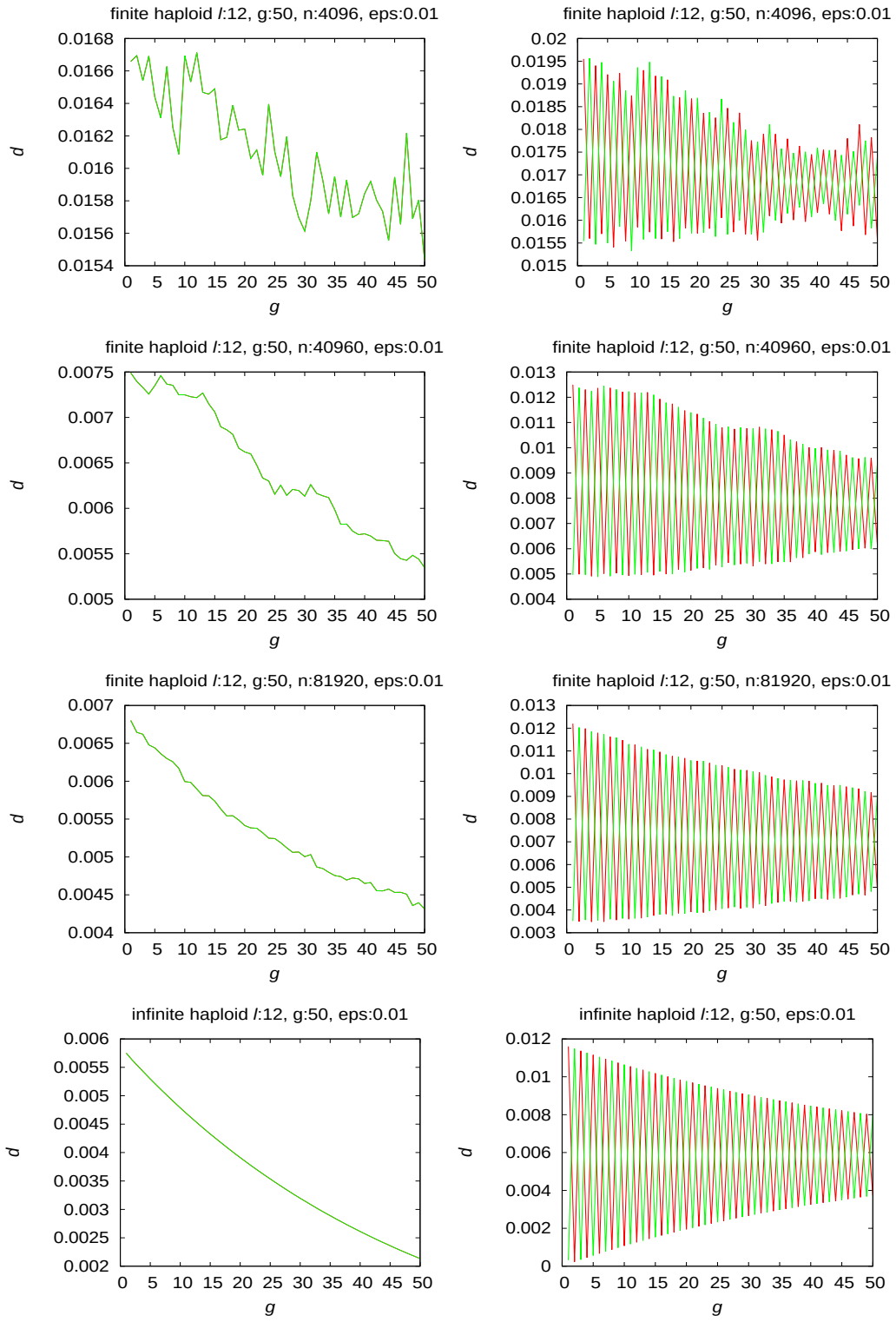


Figure 3.22: Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 12$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

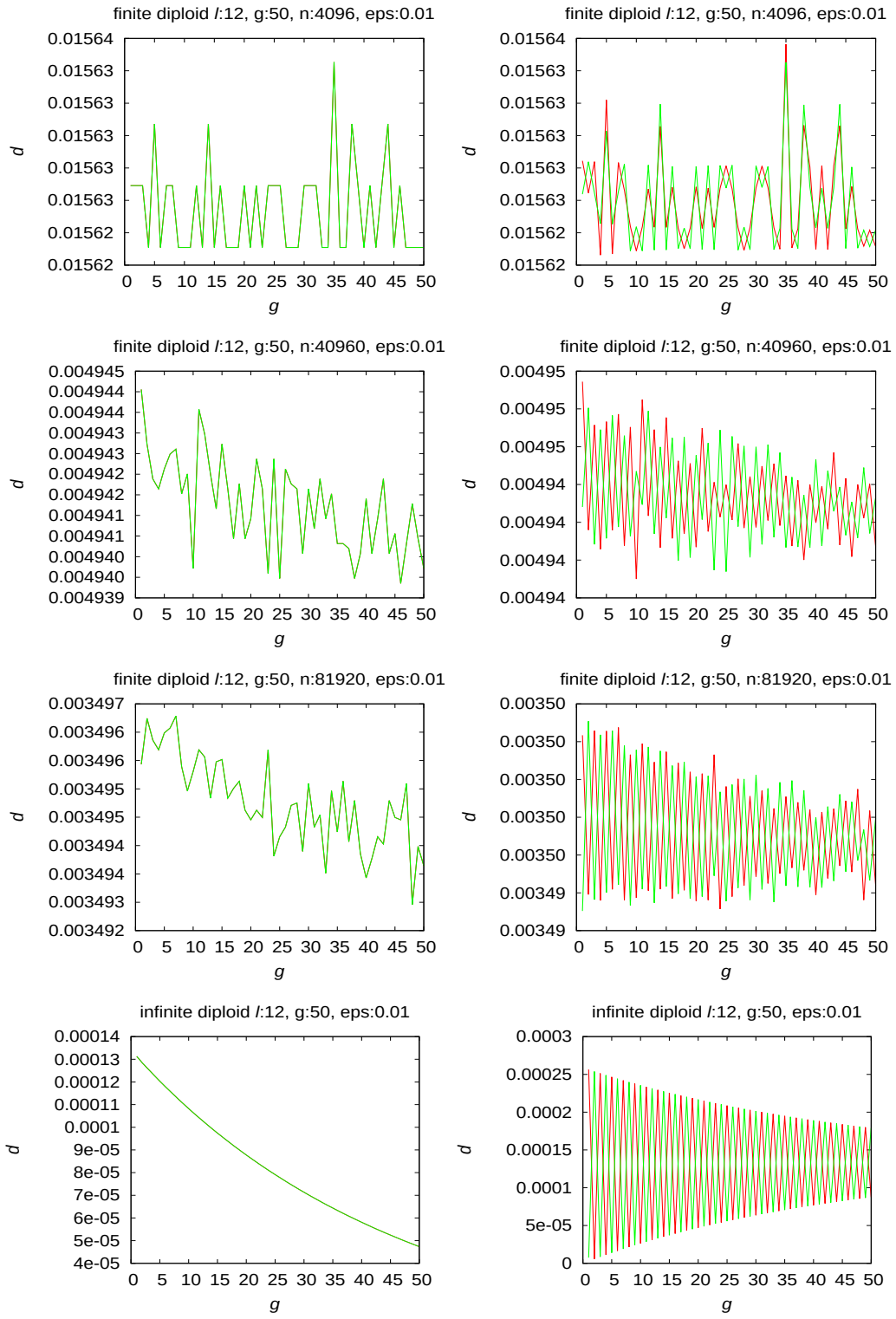


Figure 3.23: Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 12$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

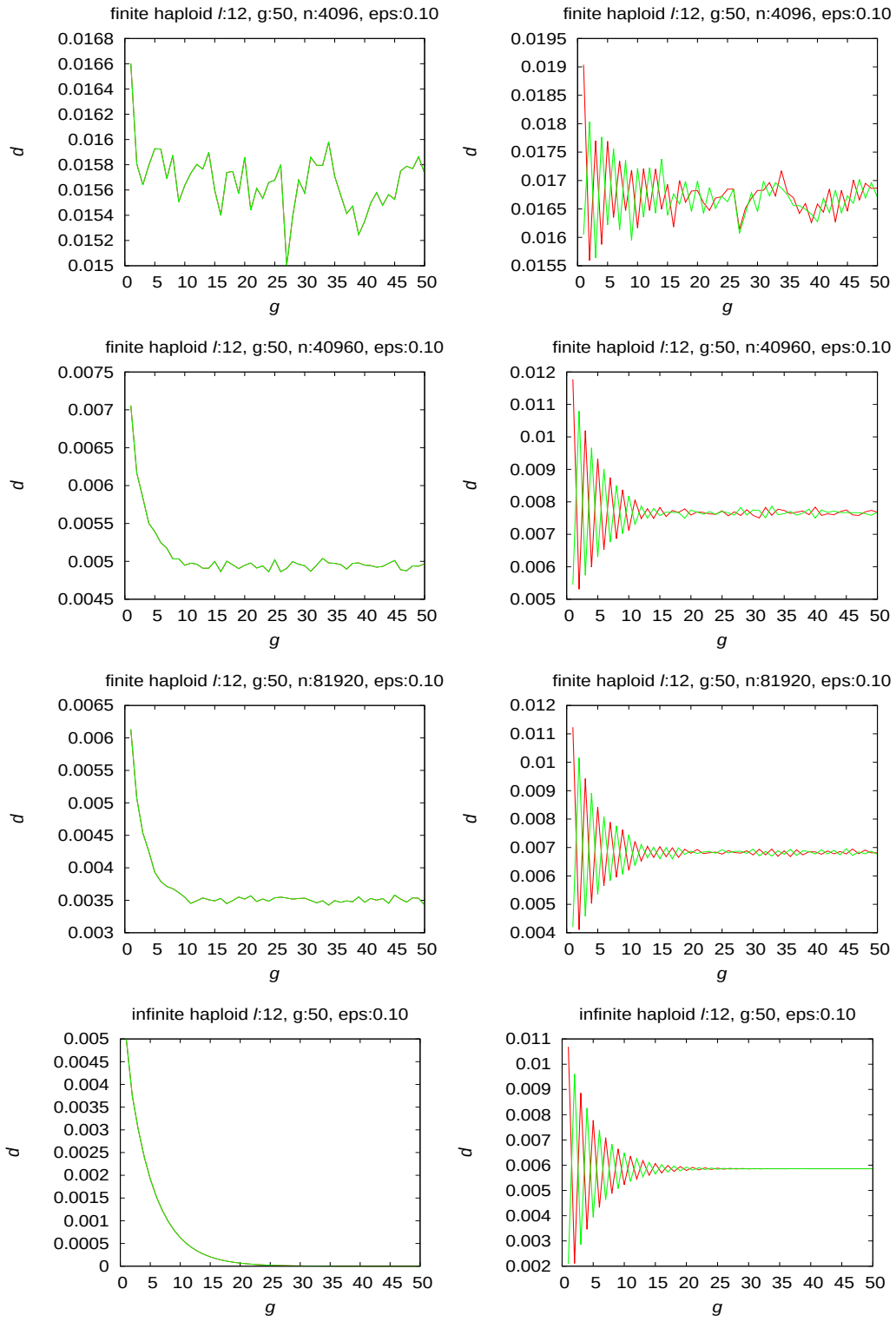


Figure 3.24: Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 12$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

Figure 3.25: Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 12$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

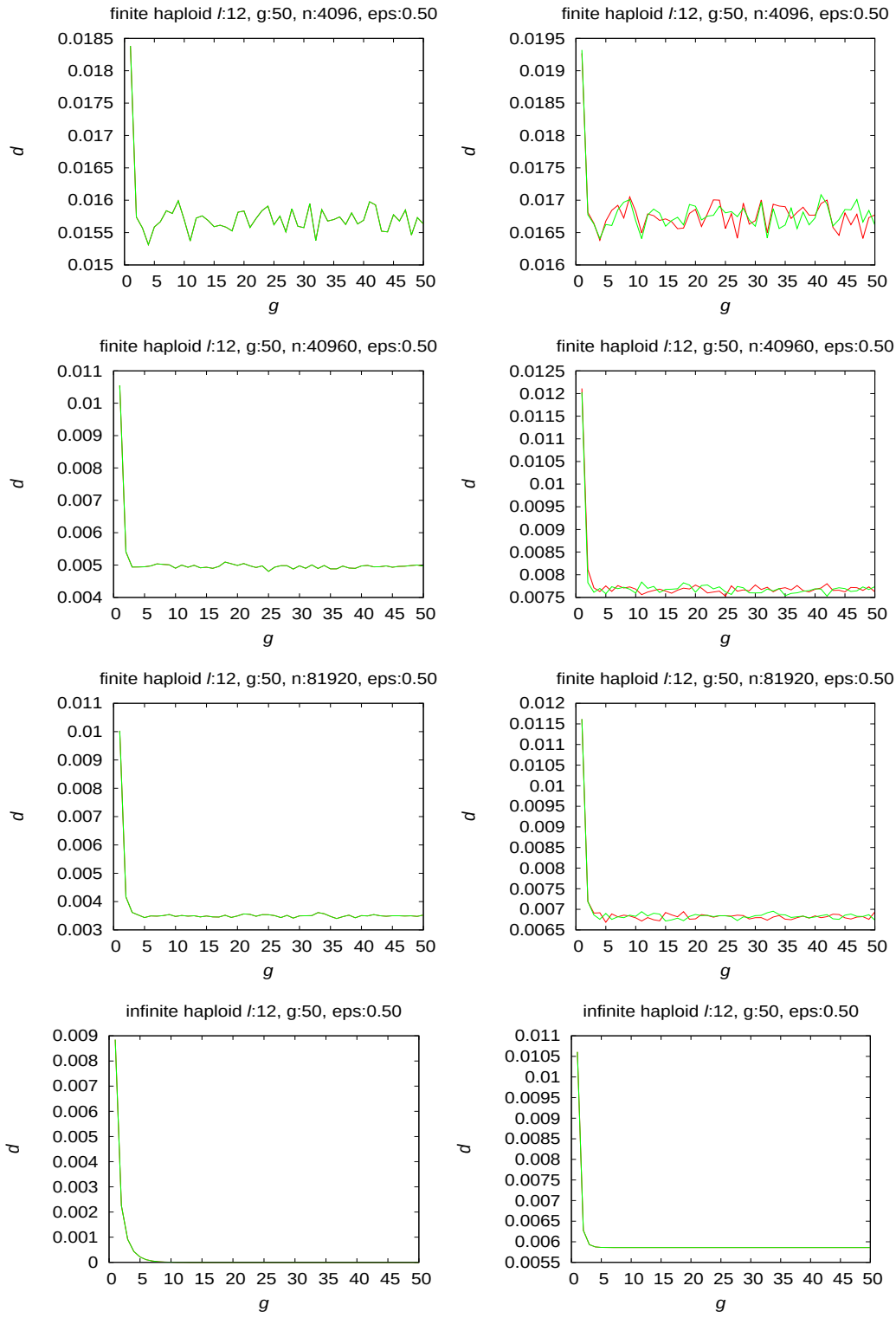


Figure 3.26: Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 12$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

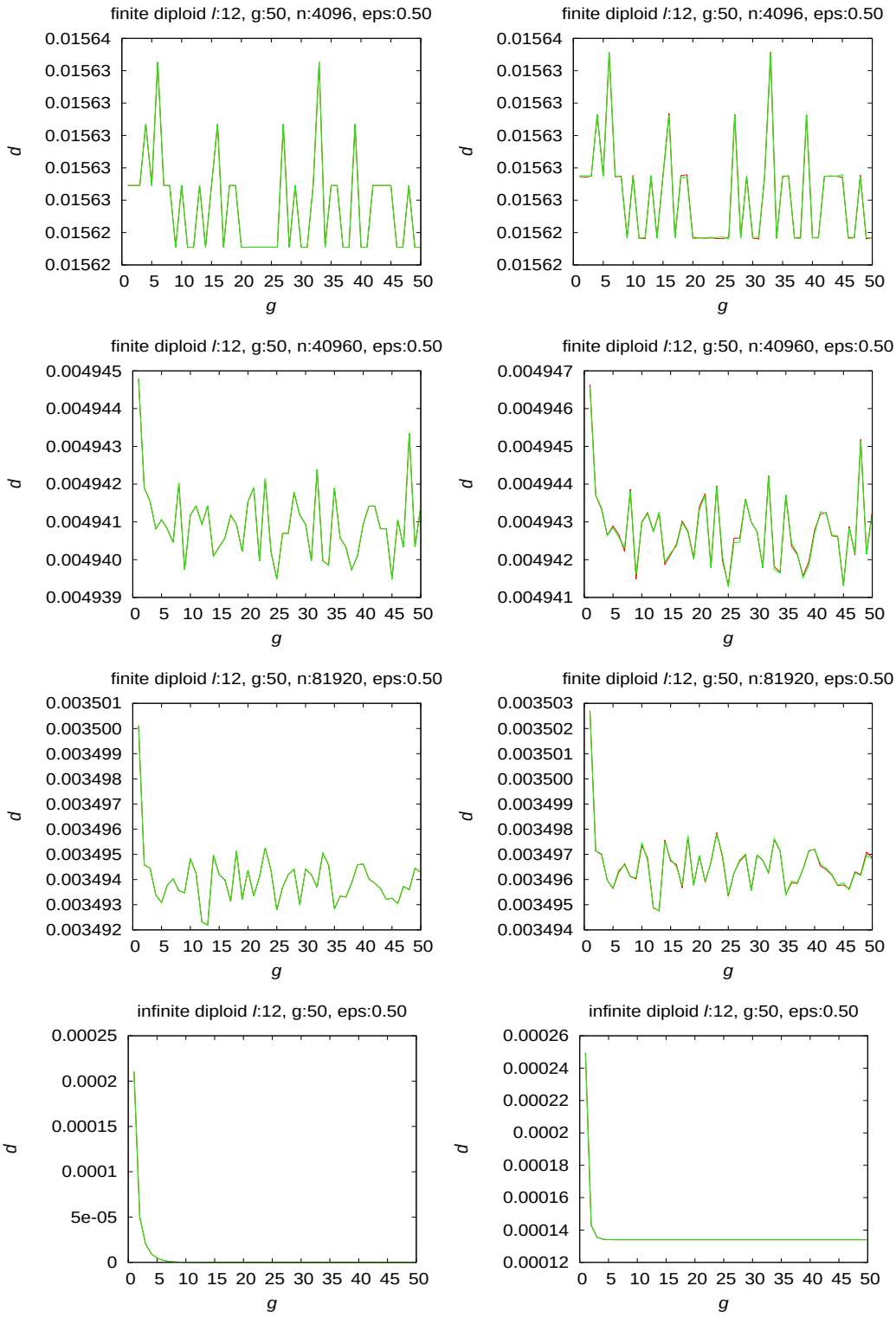


Figure 3.27: Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 12$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite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

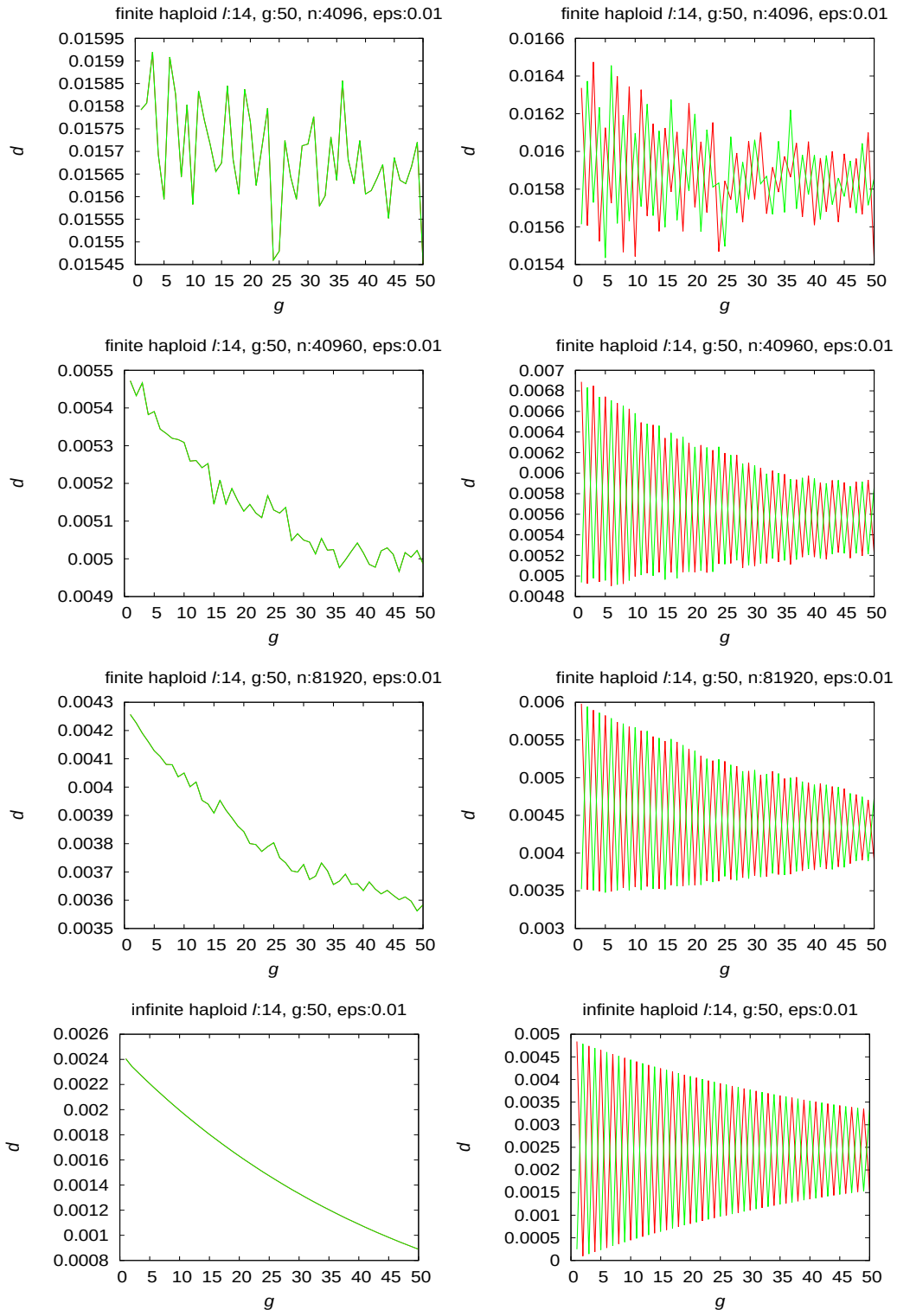


Figure 3.28: Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 14$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

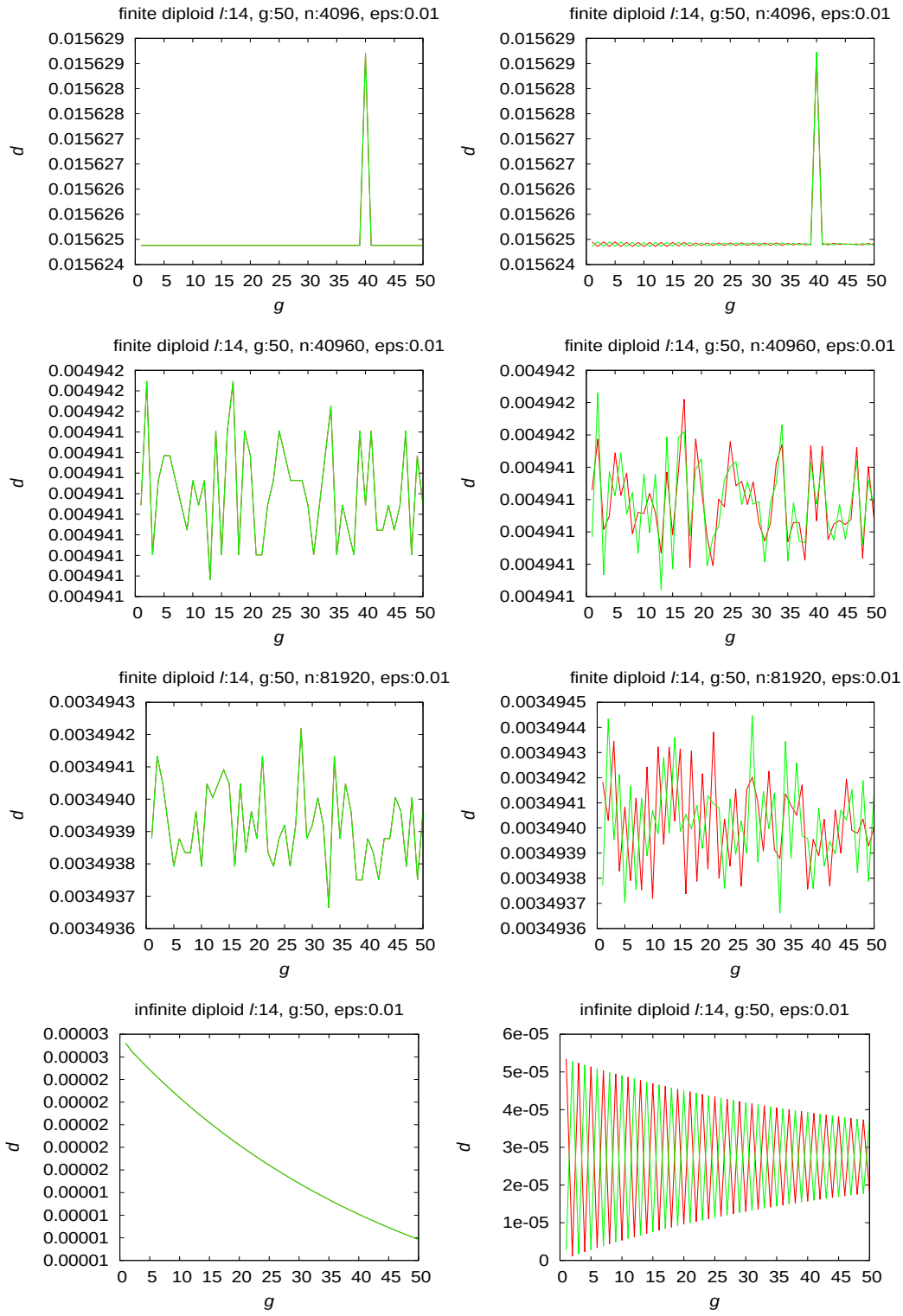


Figure 3.29: Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 14$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

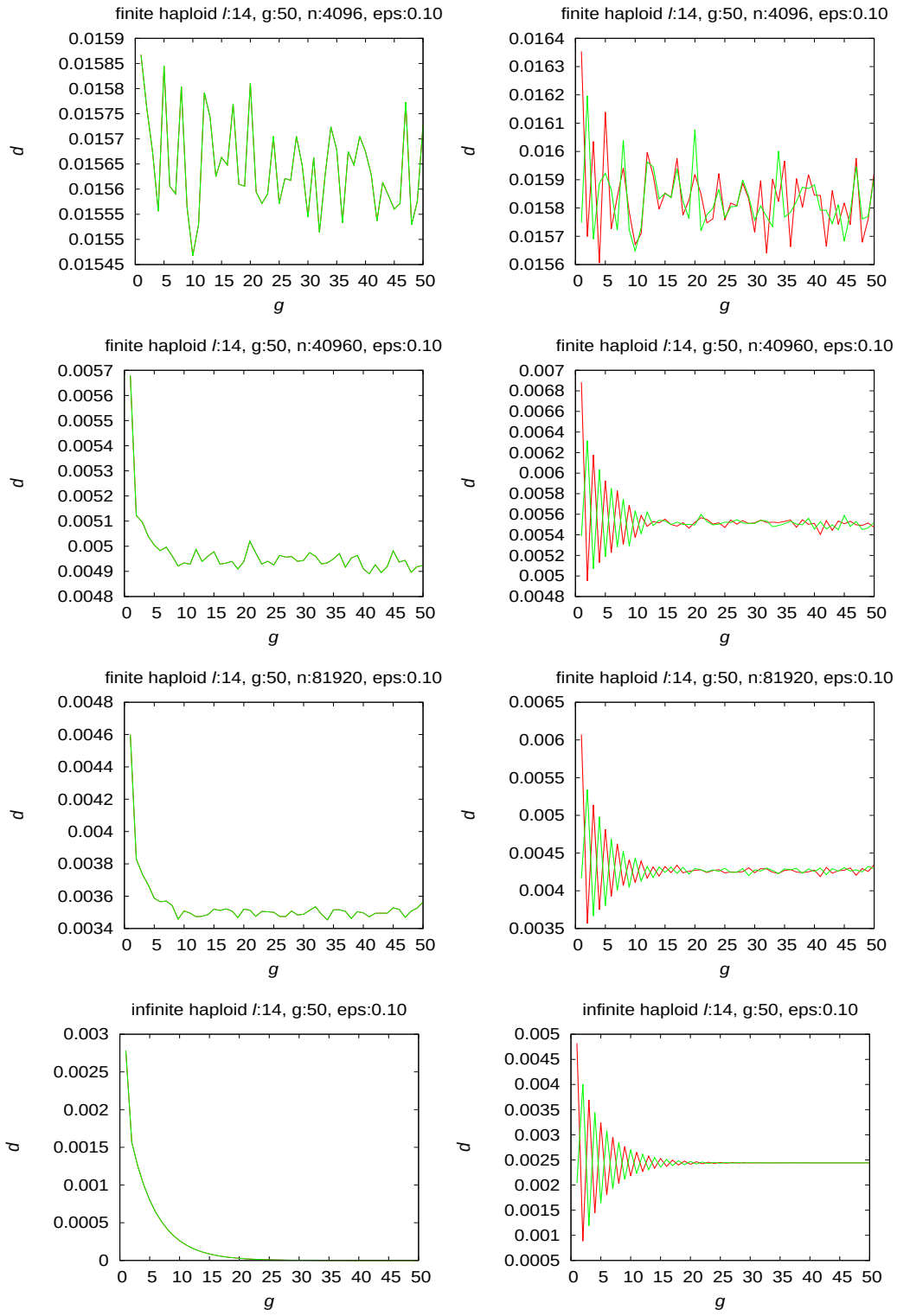


Figure 3.30: Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 14$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

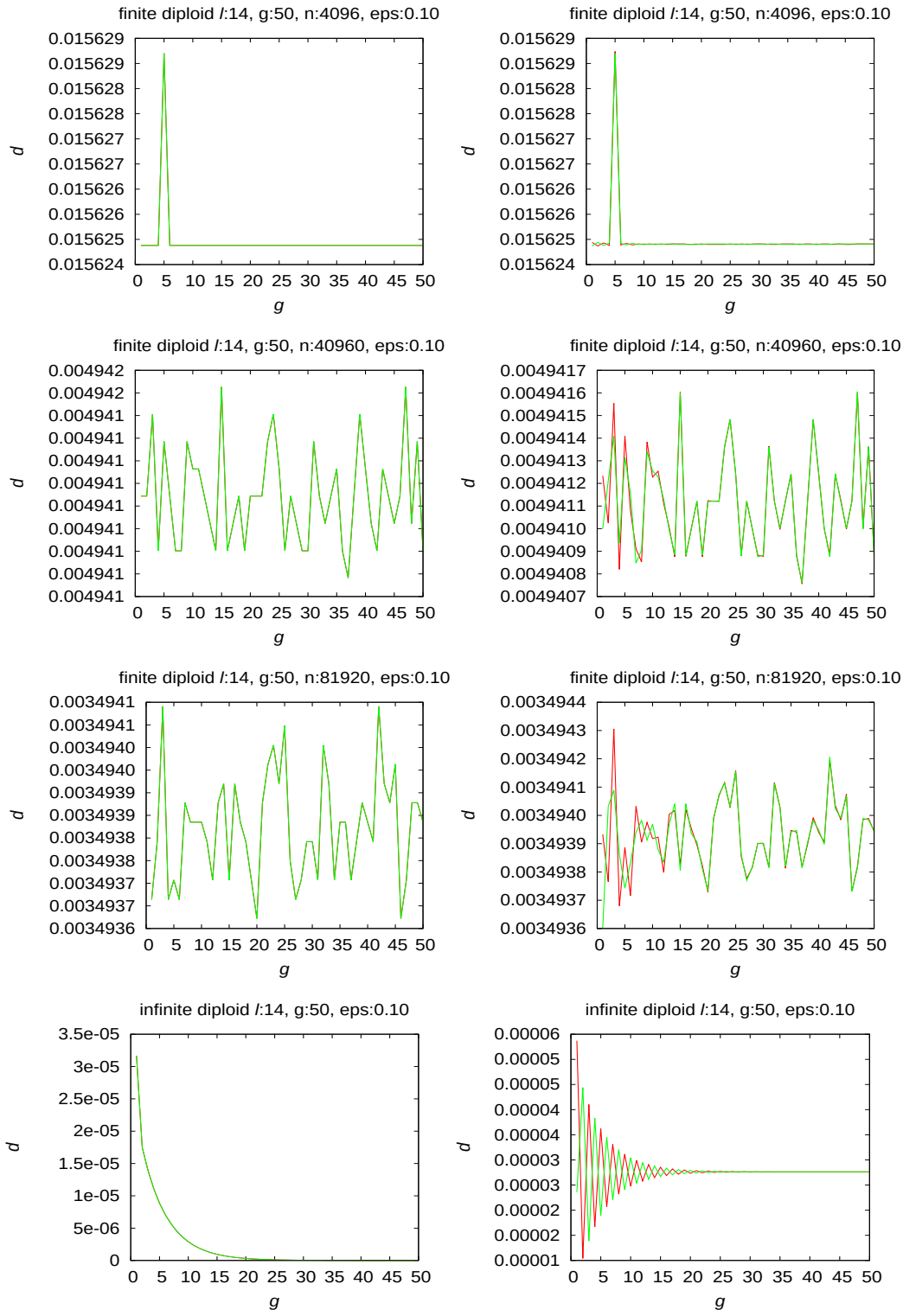


Figure 3.31: Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 14$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite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

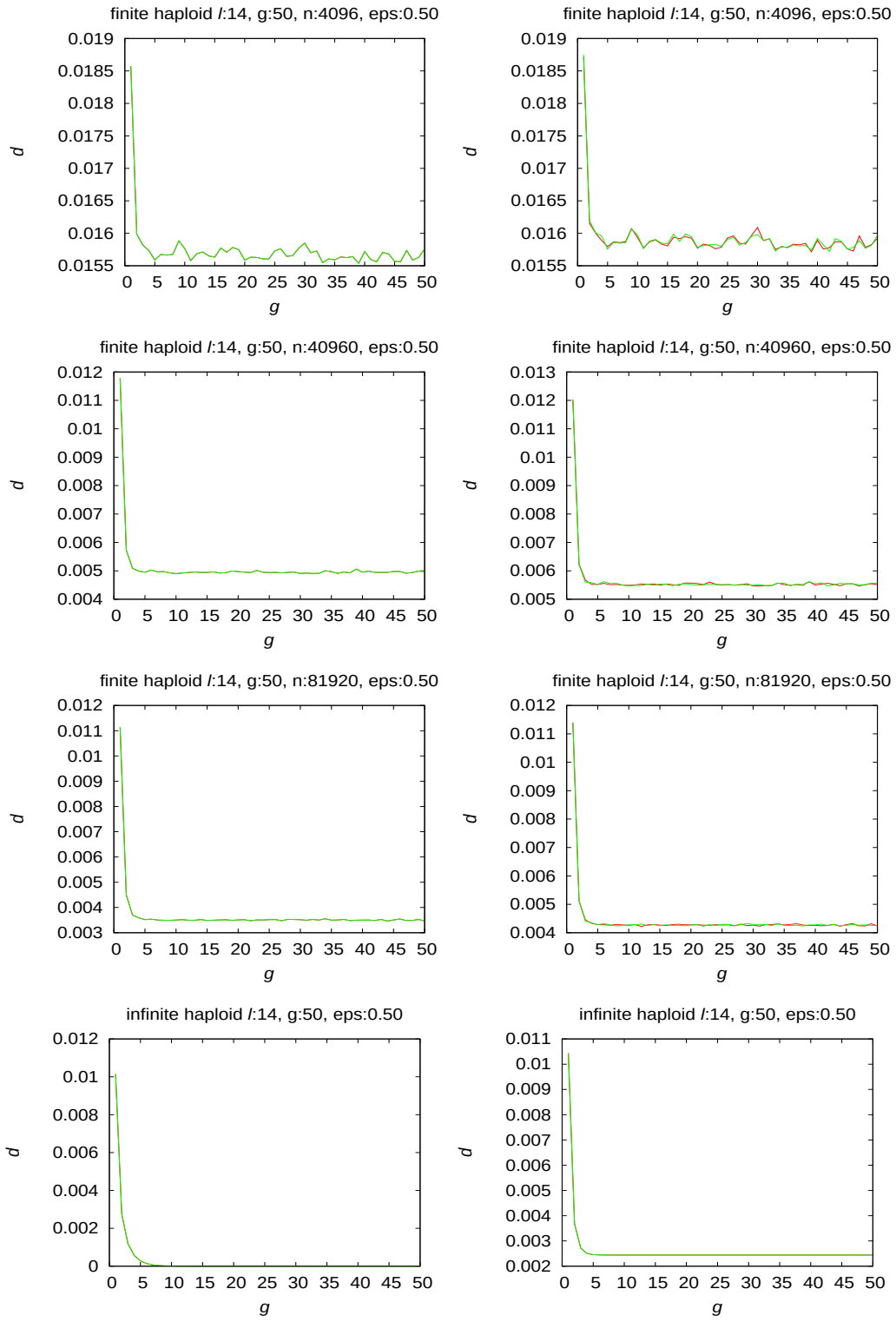


Figure 3.32: Infinite and finite haploid population oscillation behavior in case of violation in μ for genome length $\ell = 14$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

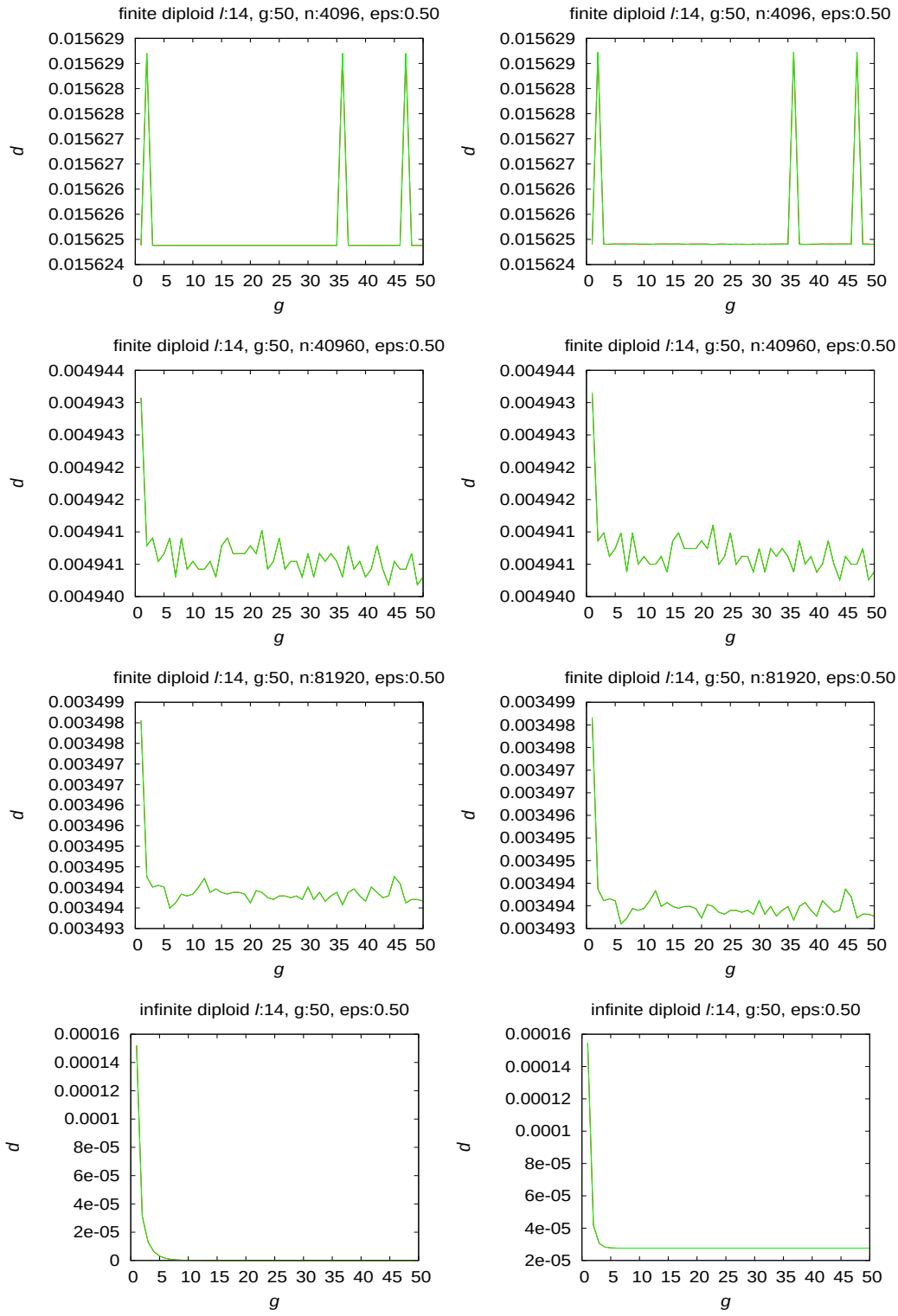


Figure 3.33: Infinite and finite diploid population oscillation behavior in case of violation in μ for genome length $\ell = 14$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite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

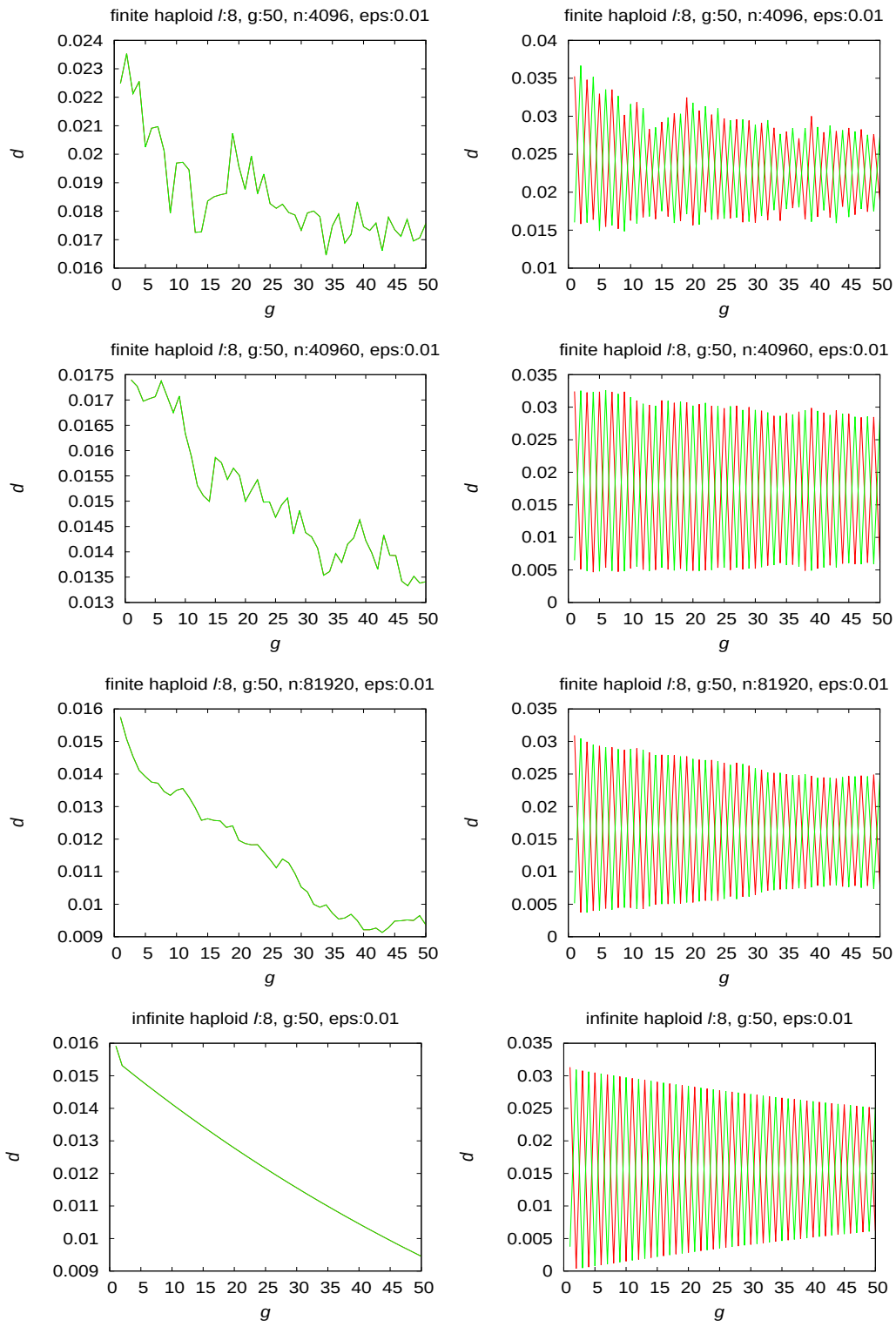


Figure 3.34: Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 8$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite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

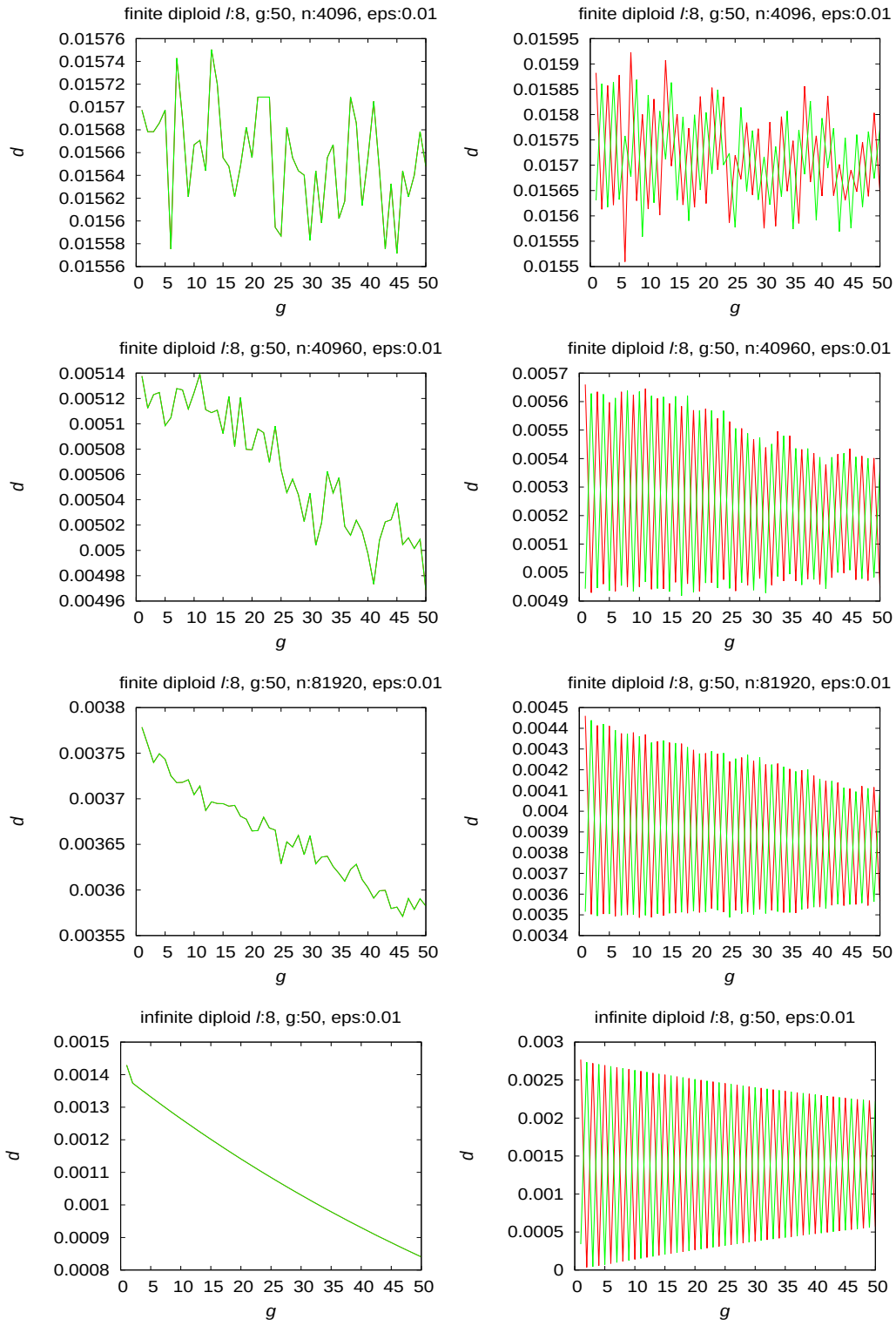


Figure 3.35: Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 8$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

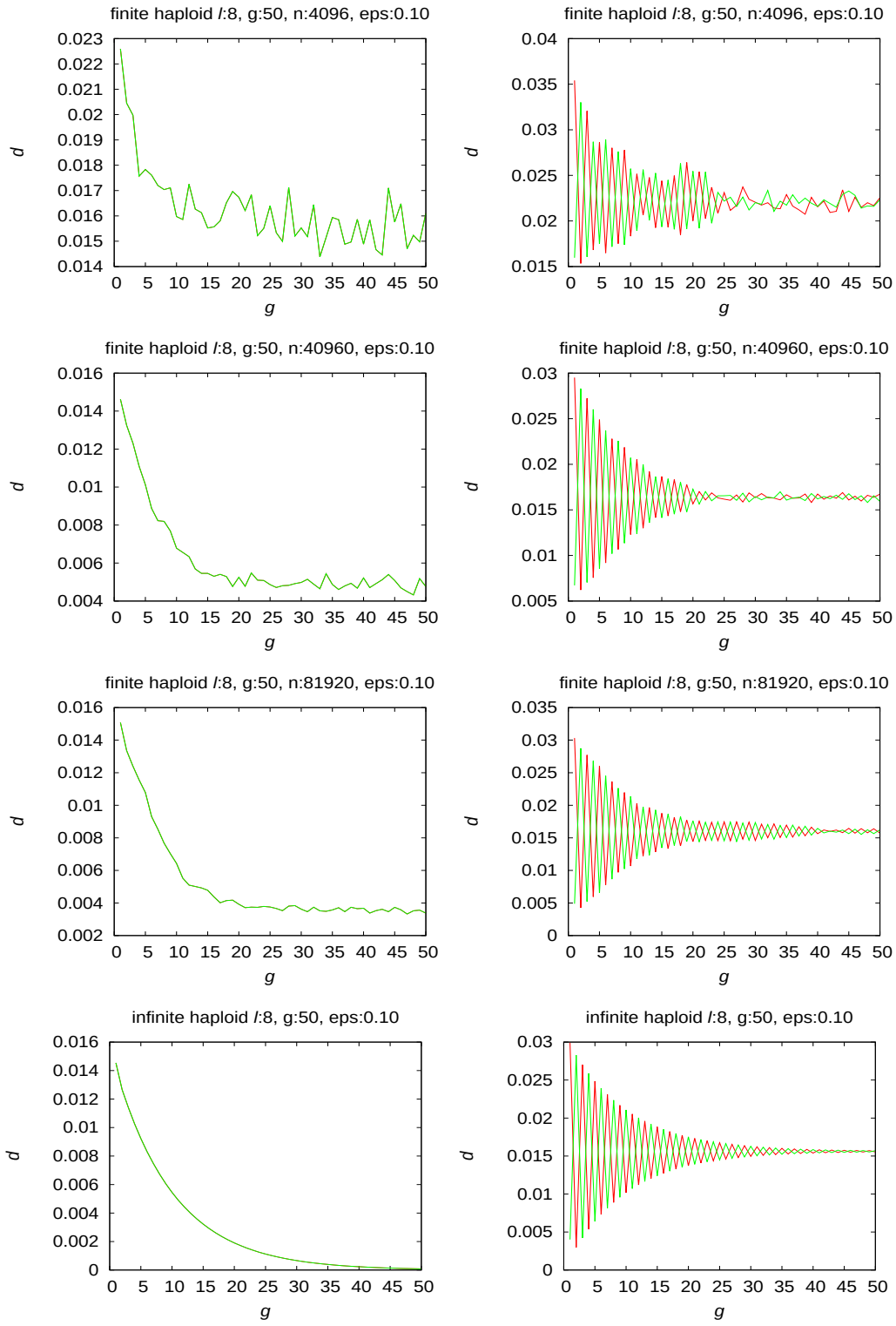


Figure 3.36: Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 8$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite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

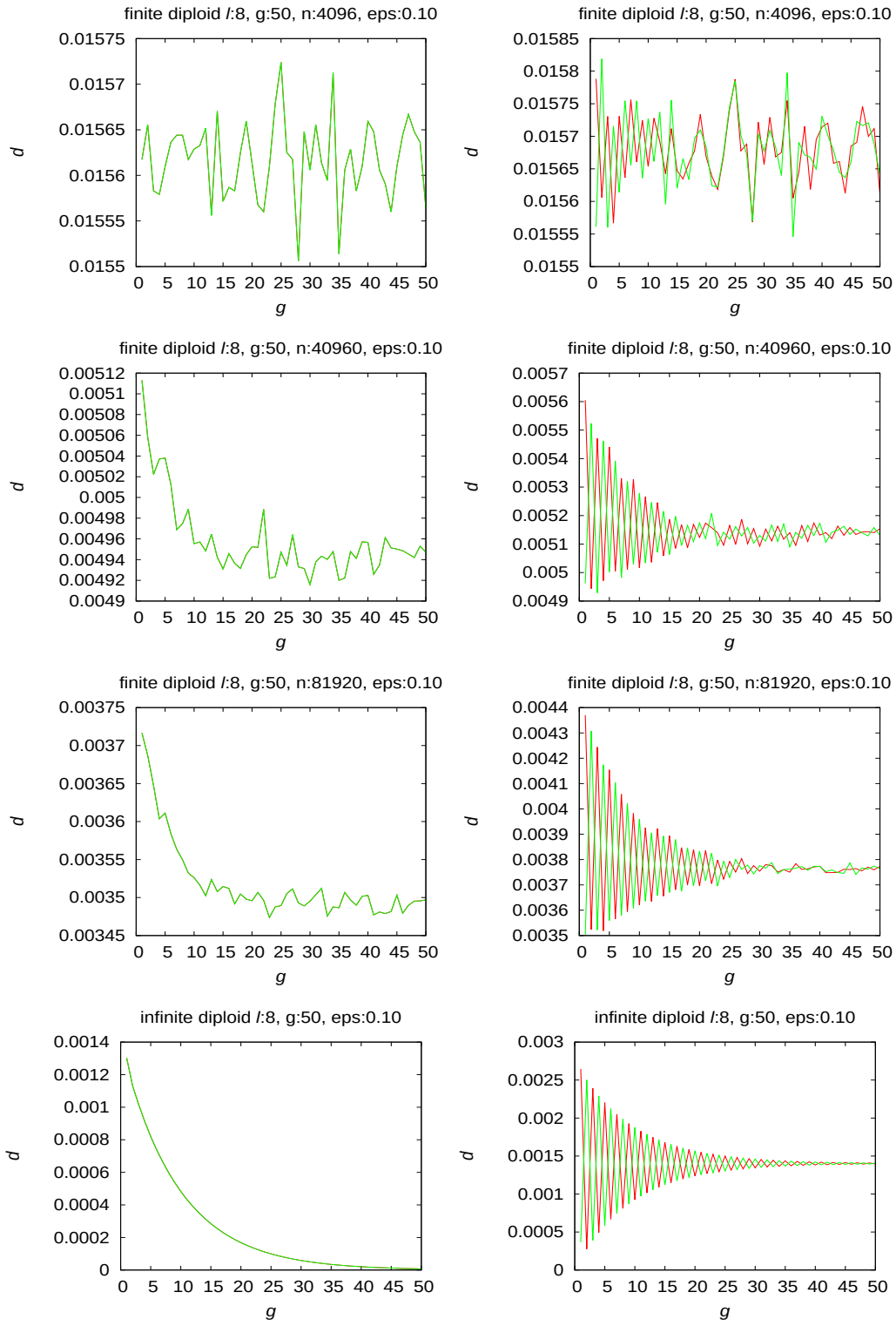


Figure 3.37: Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 8$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

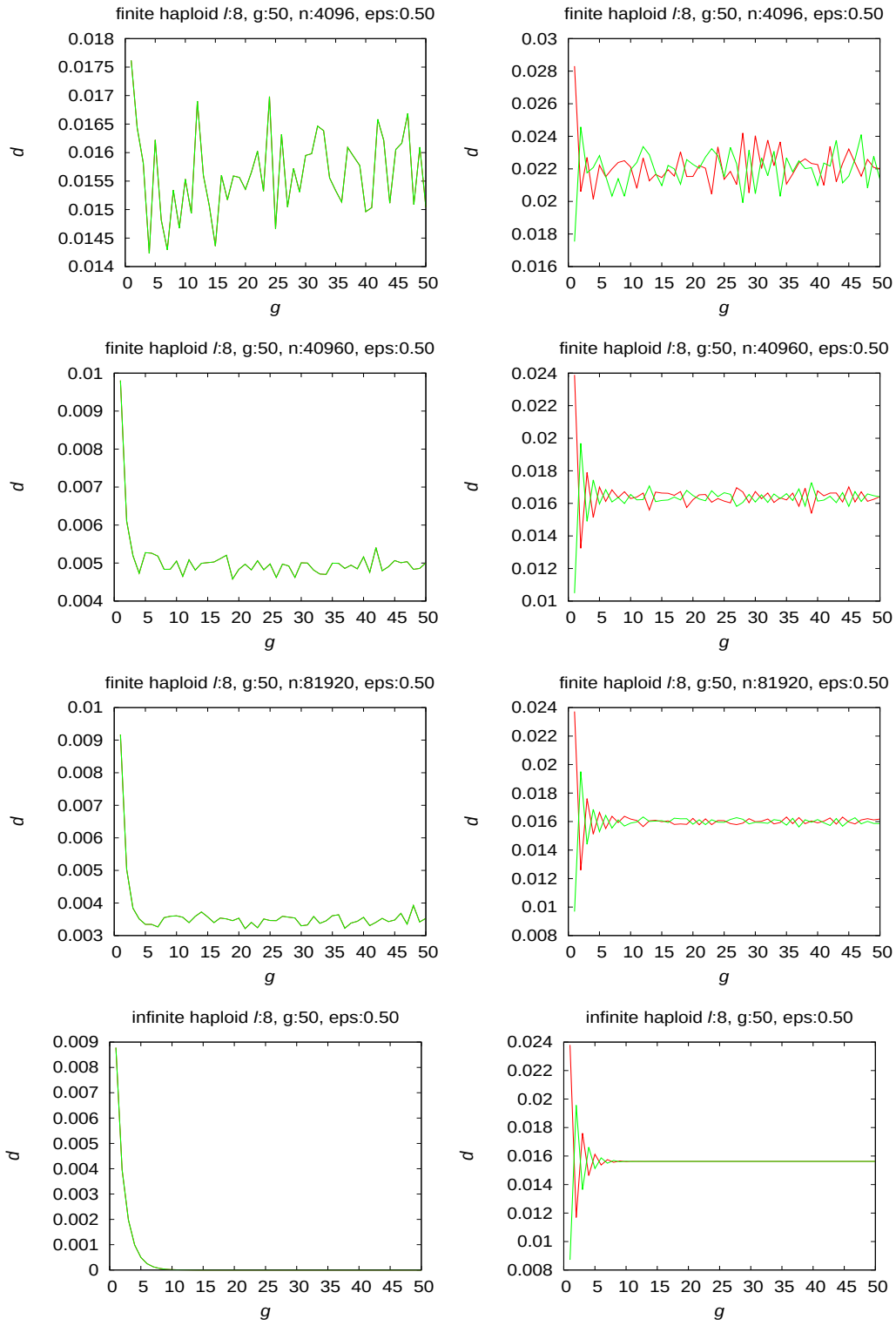


Figure 3.38: Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 8$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

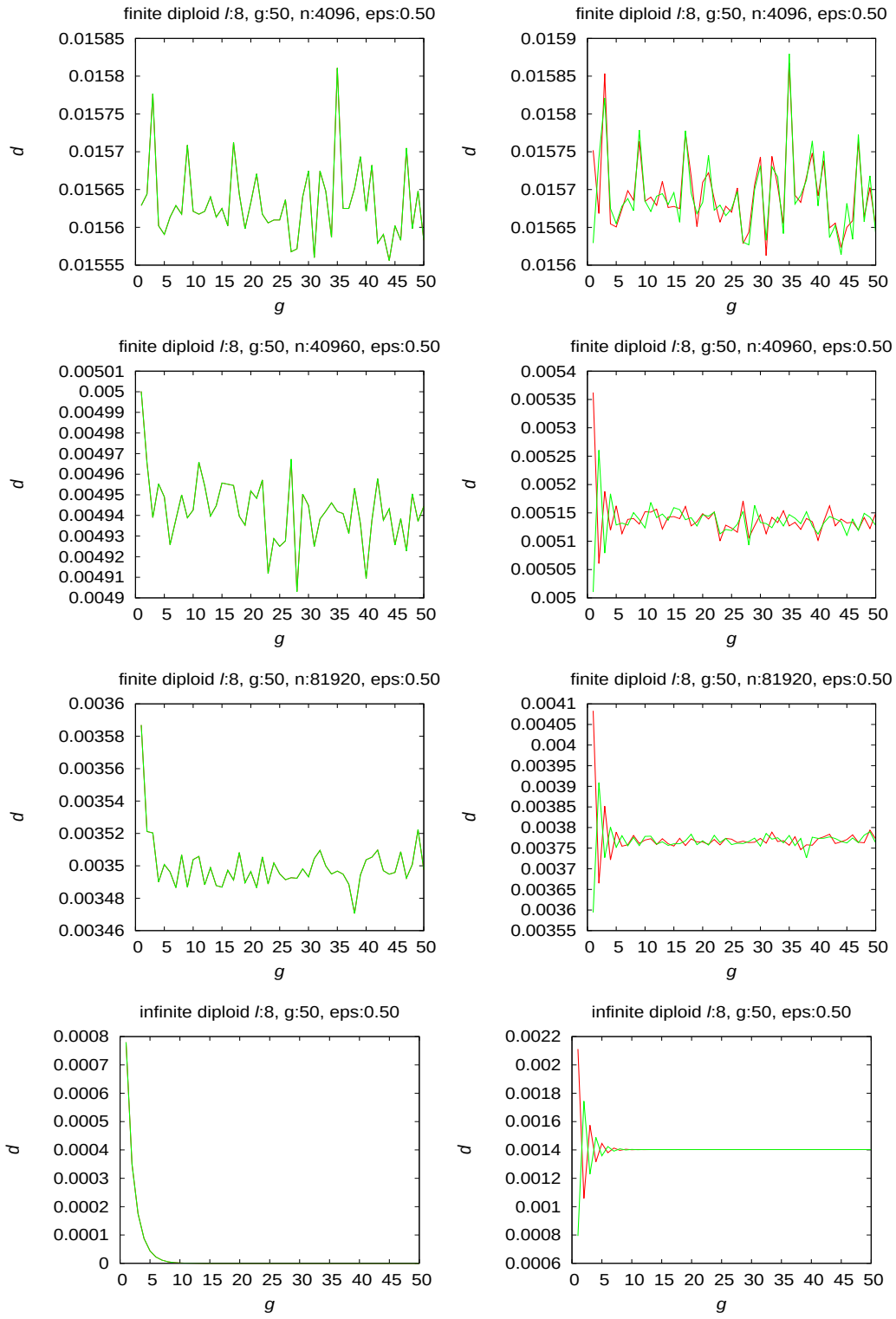


Figure 3.39: Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 8$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

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

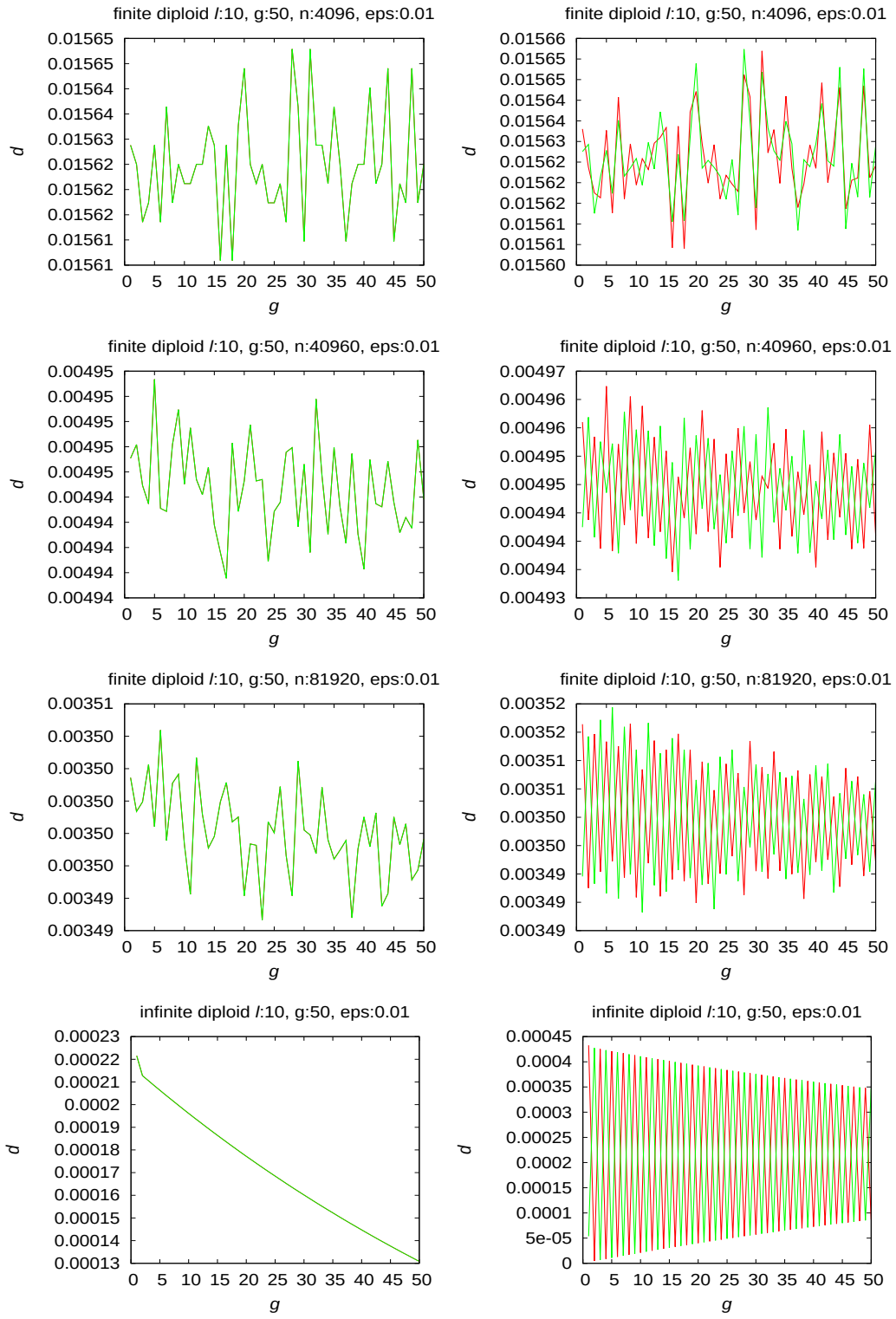


Figure 3.41: Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 10$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

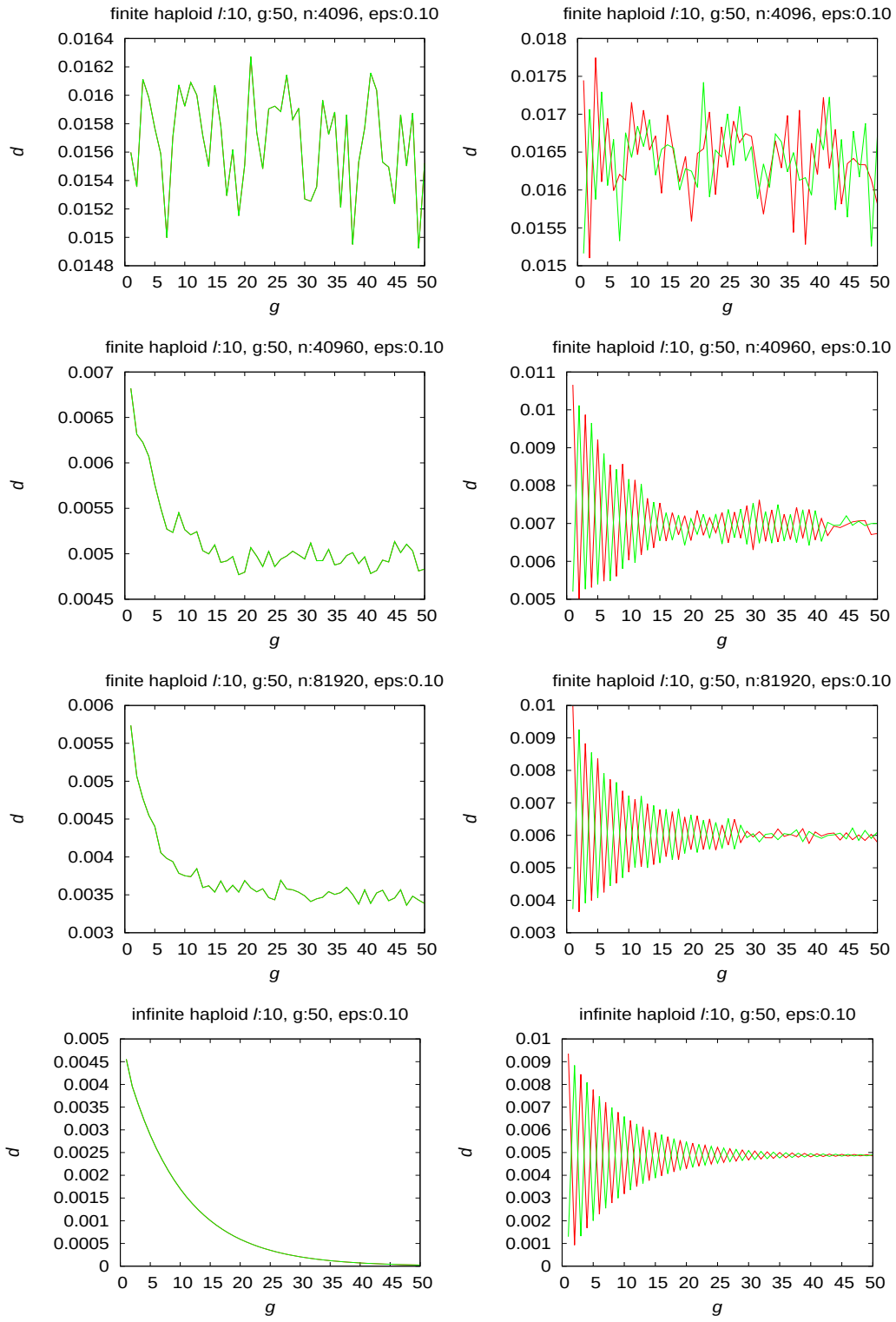


Figure 3.42: Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 10$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

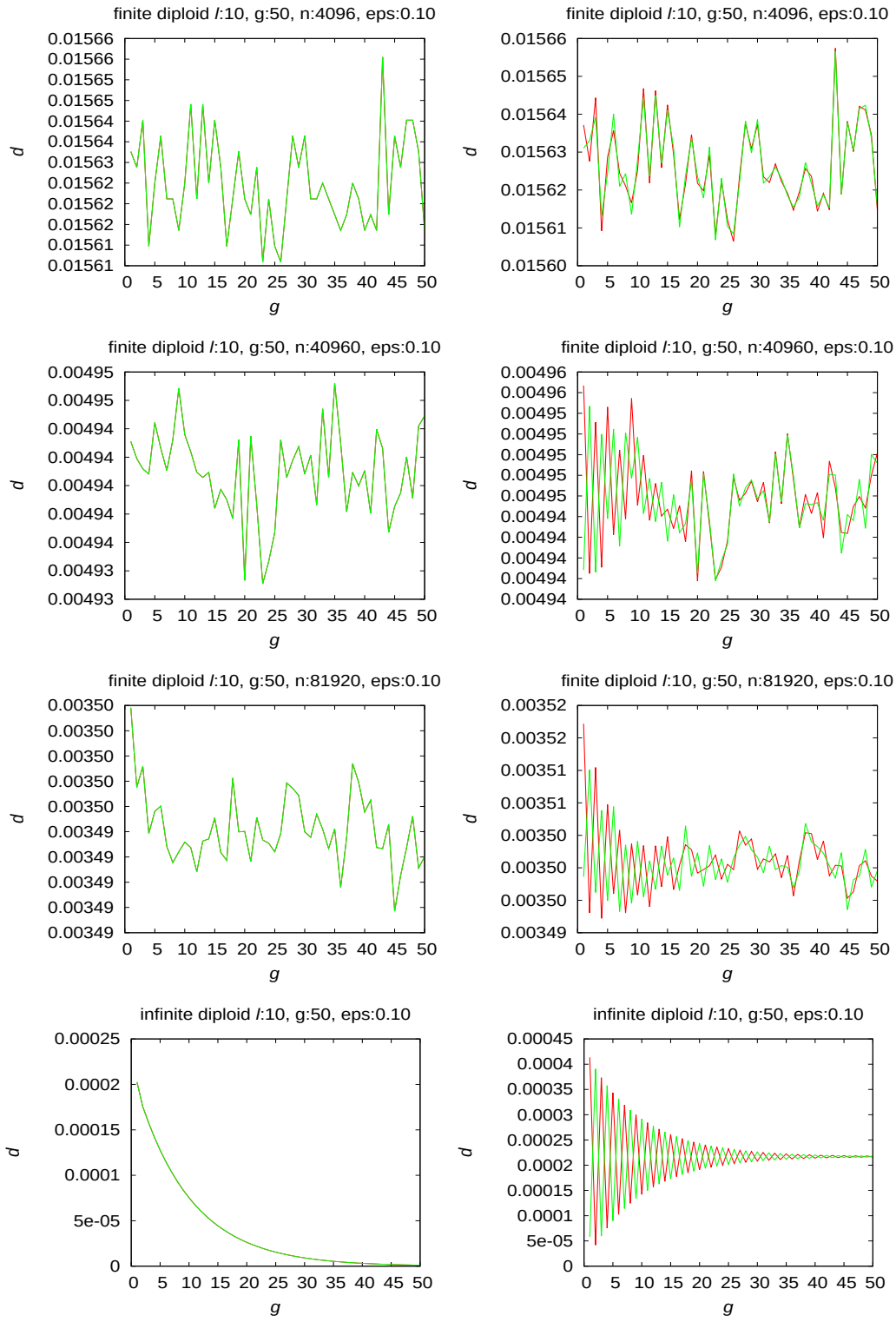


Figure 3.43: Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 10$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

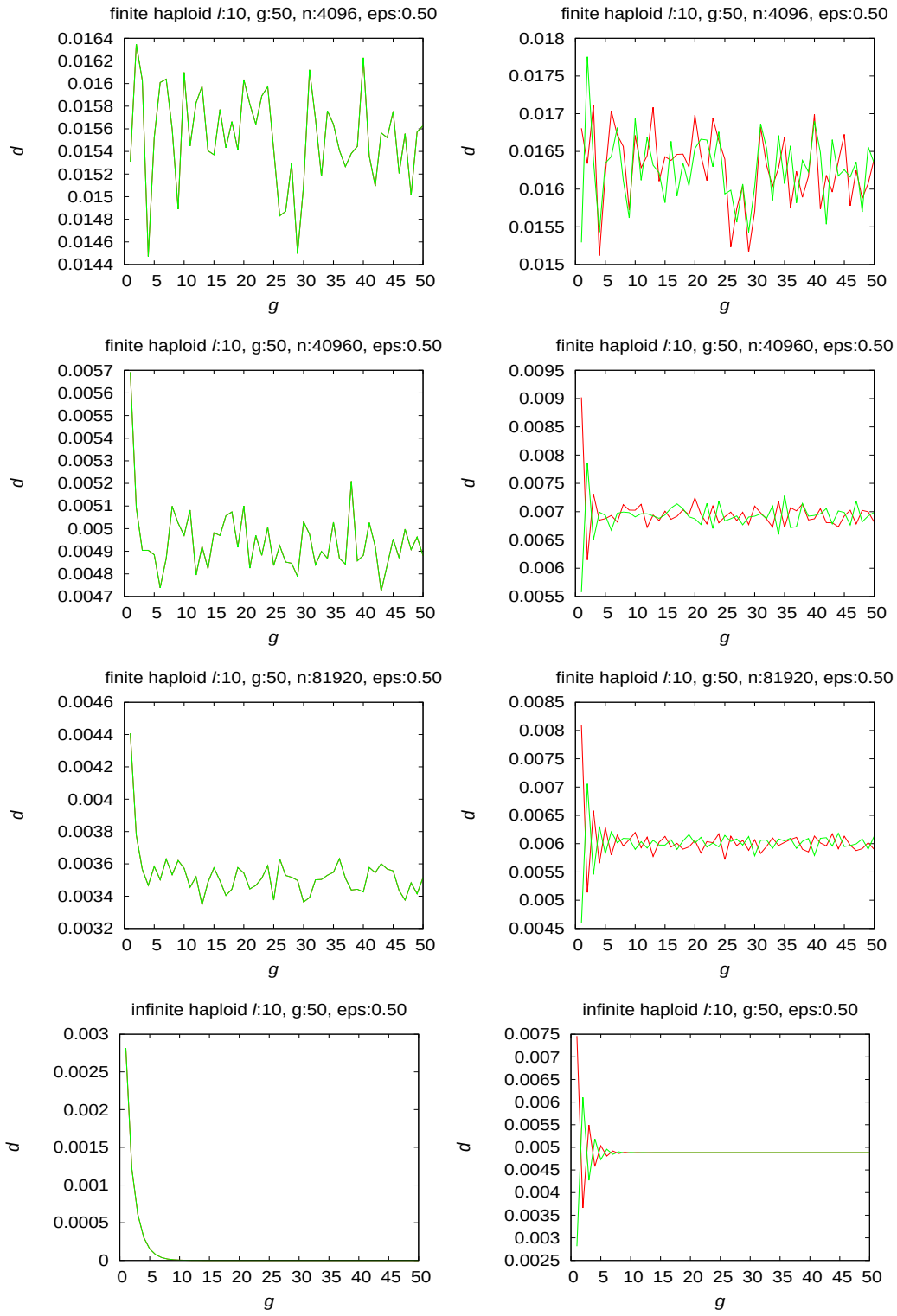


Figure 3.44: Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 10$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

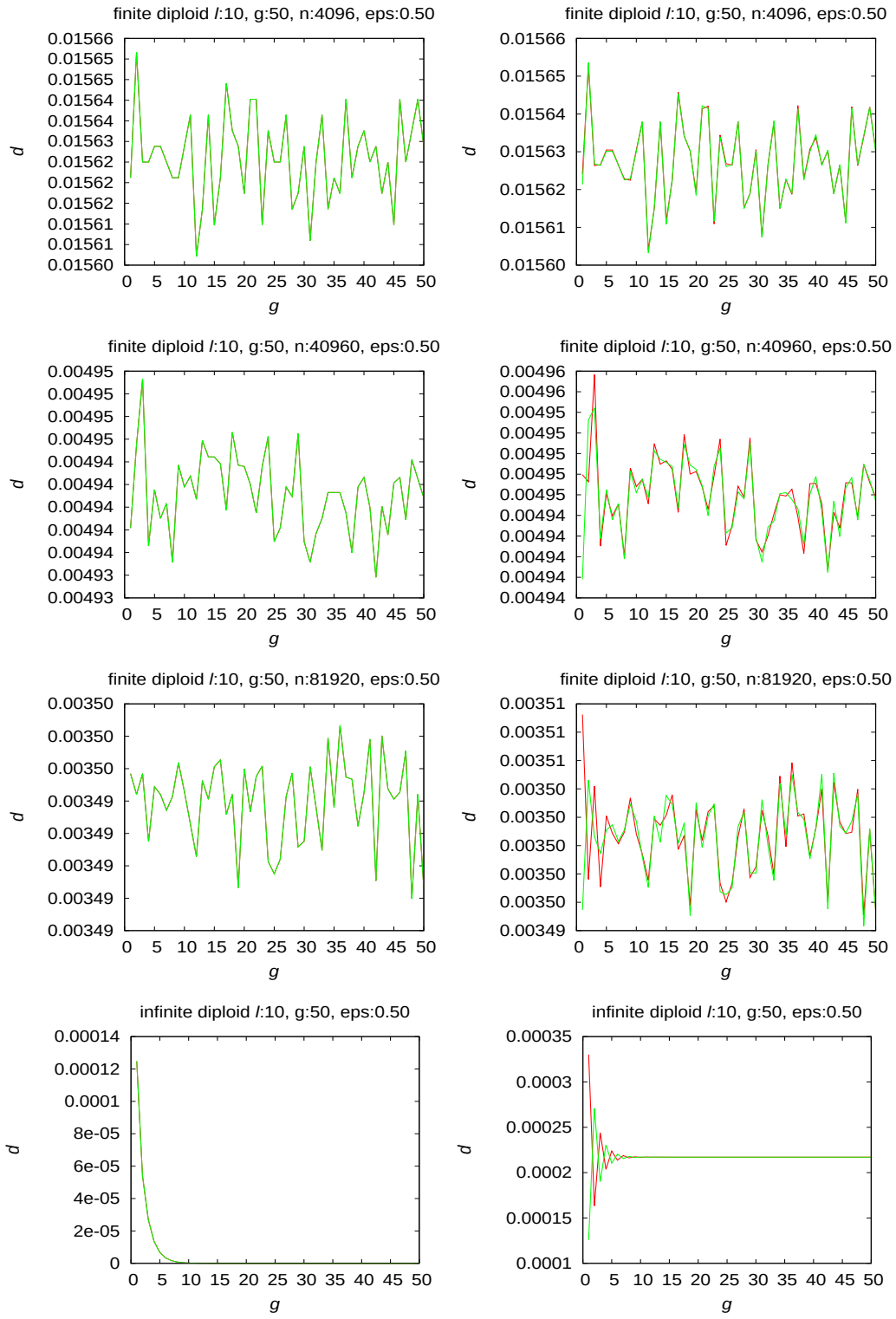


Figure 3.45: Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 10$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

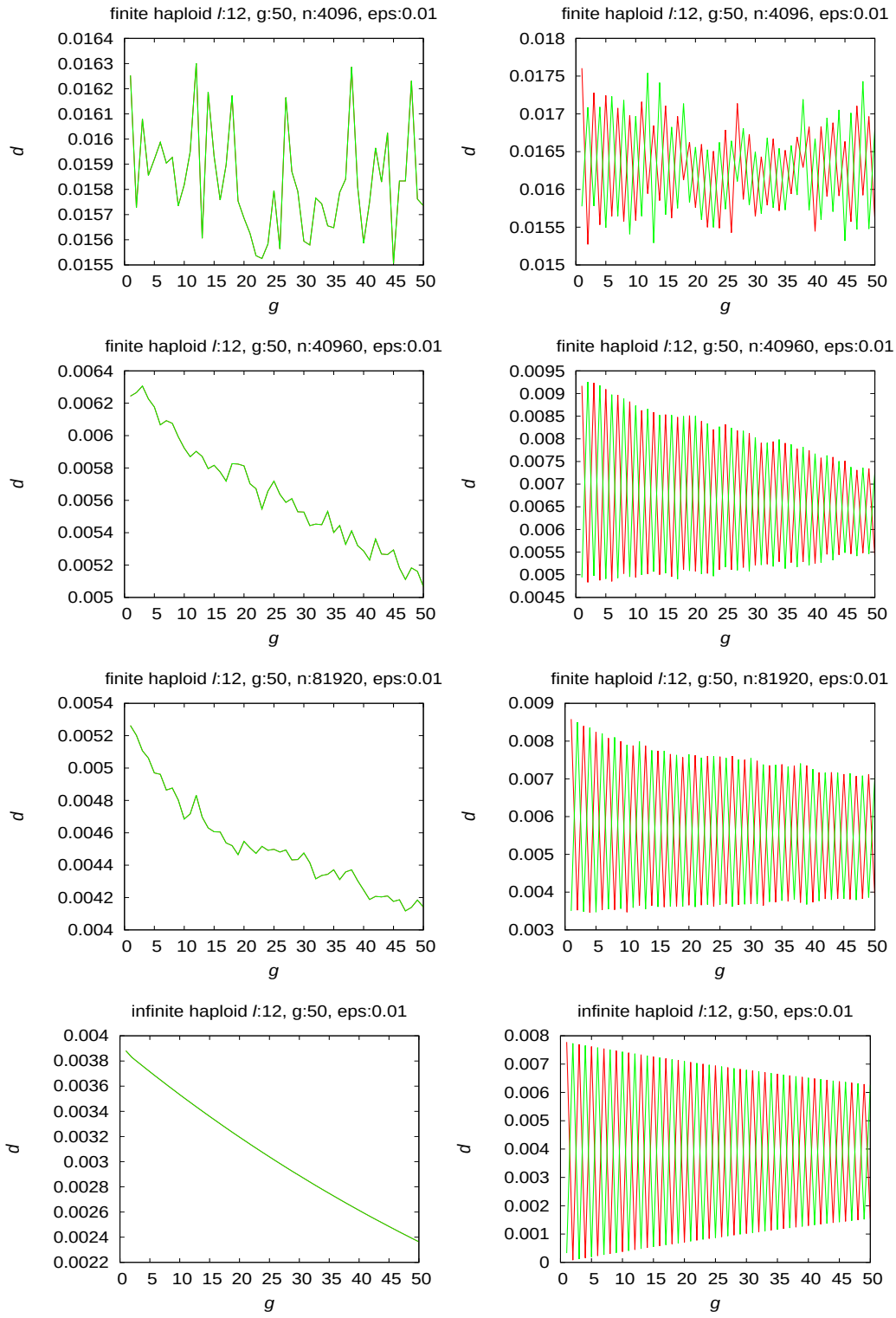


Figure 3.46: Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 12$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

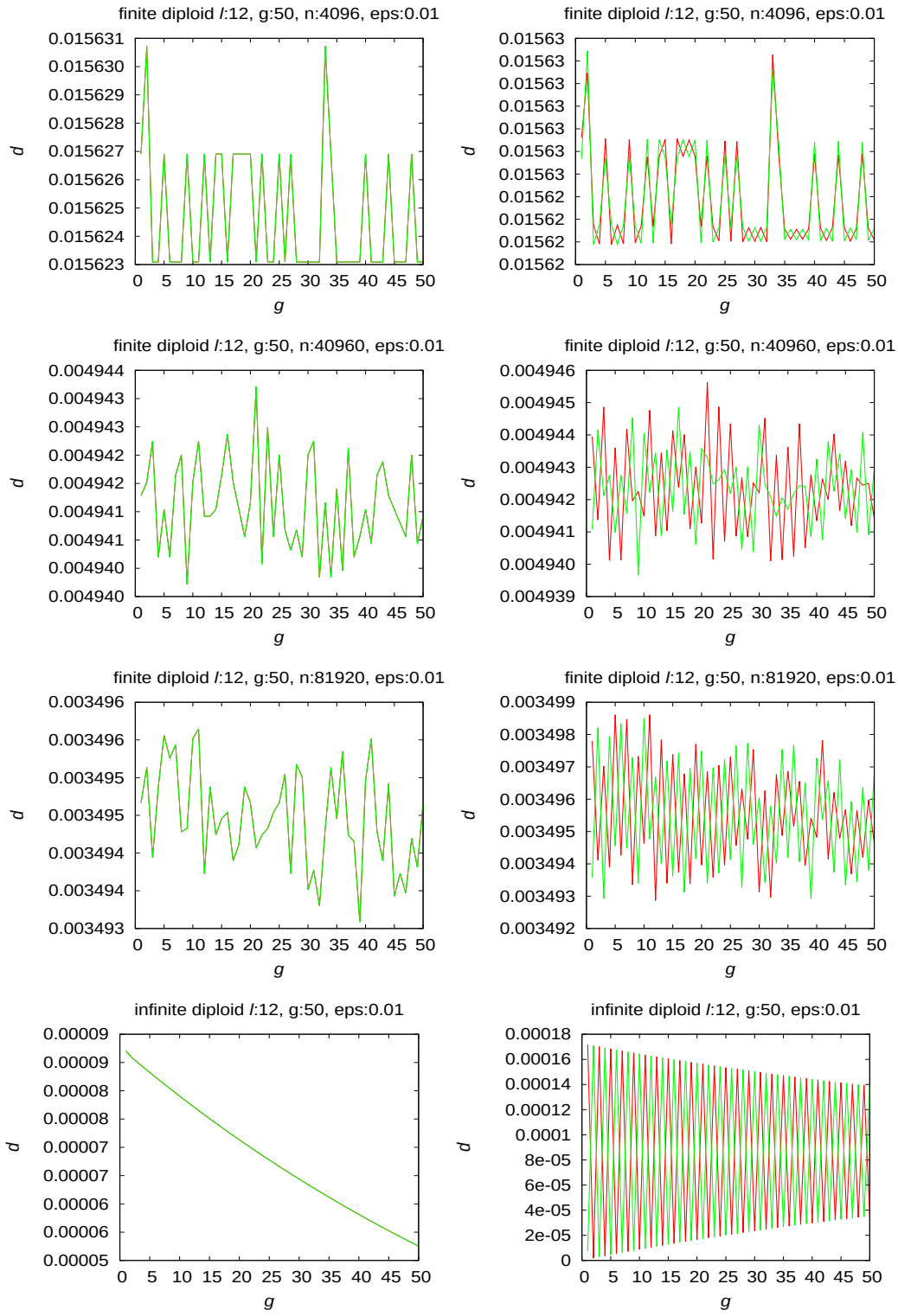


Figure 3.47: Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 12$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

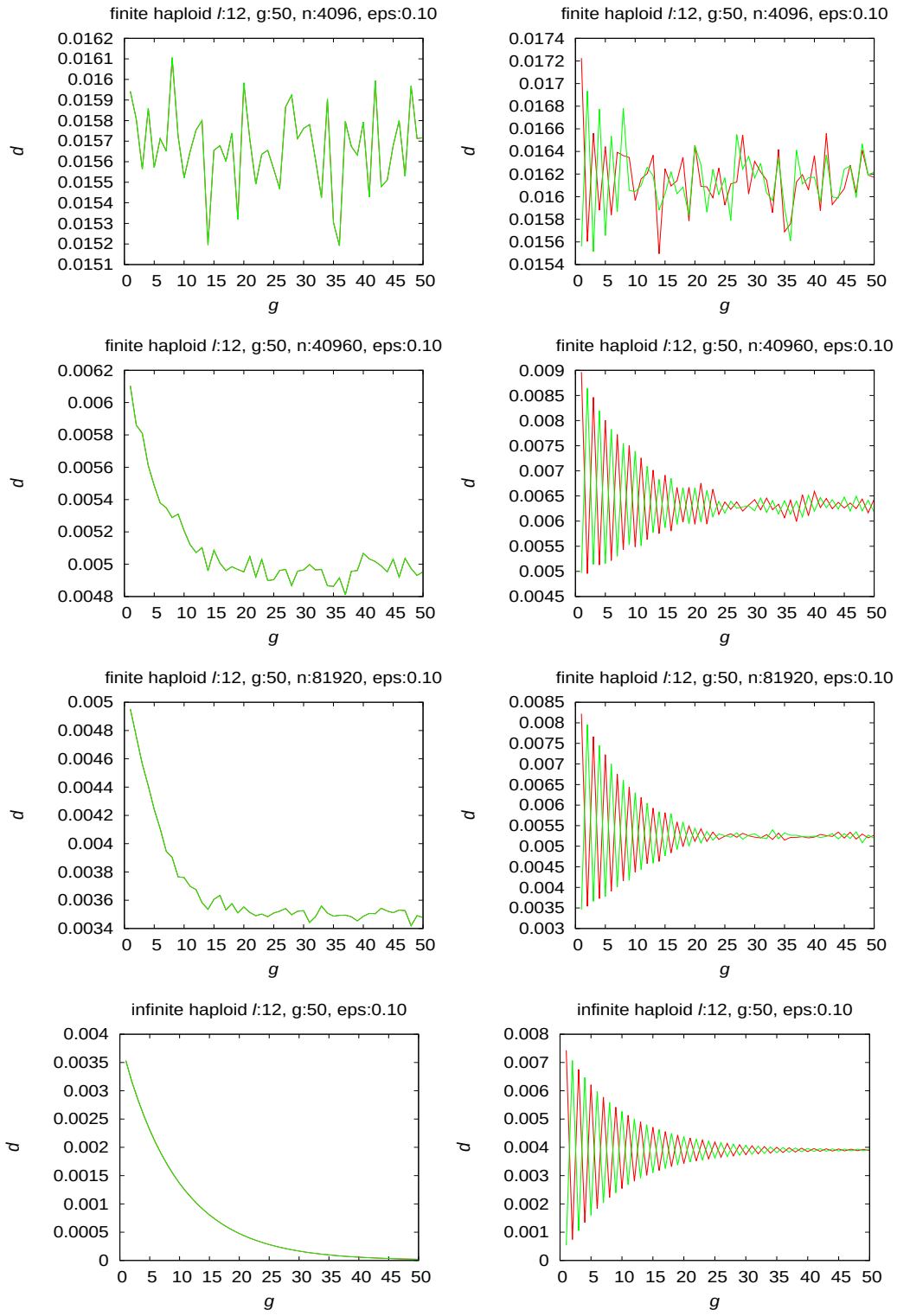


Figure 3.48: Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 12$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

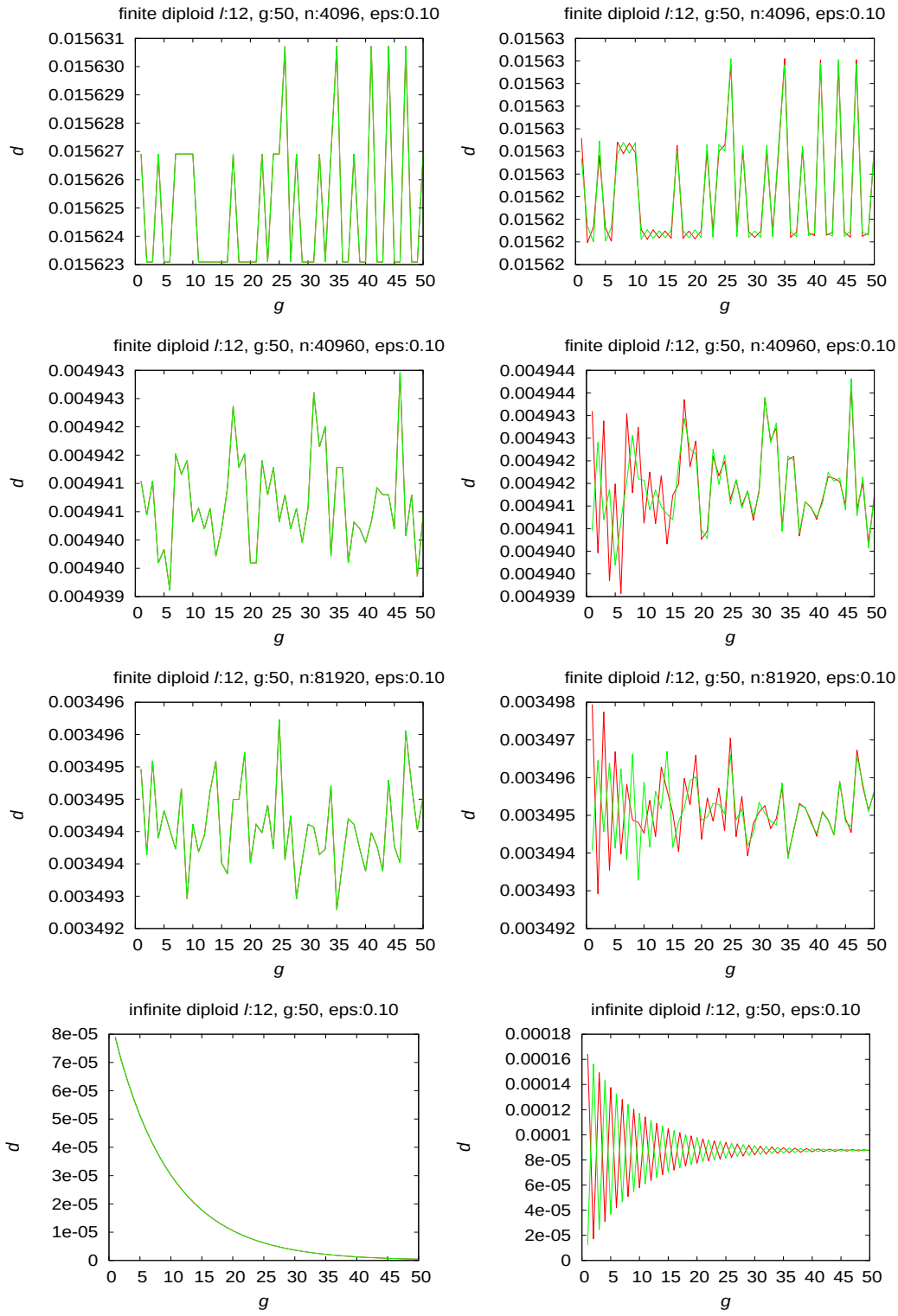


Figure 3.49: Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 12$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

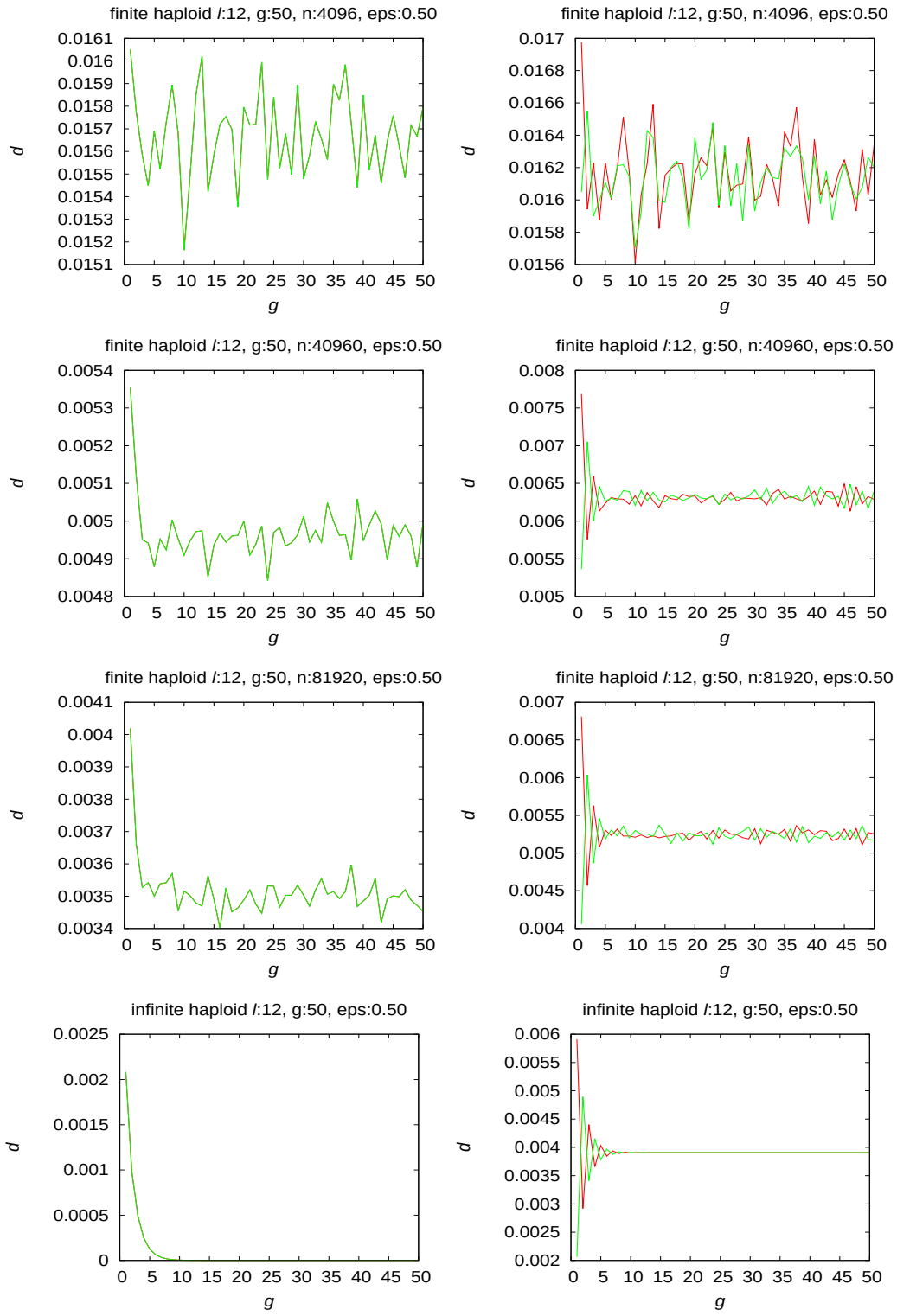


Figure 3.50: Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 12$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

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

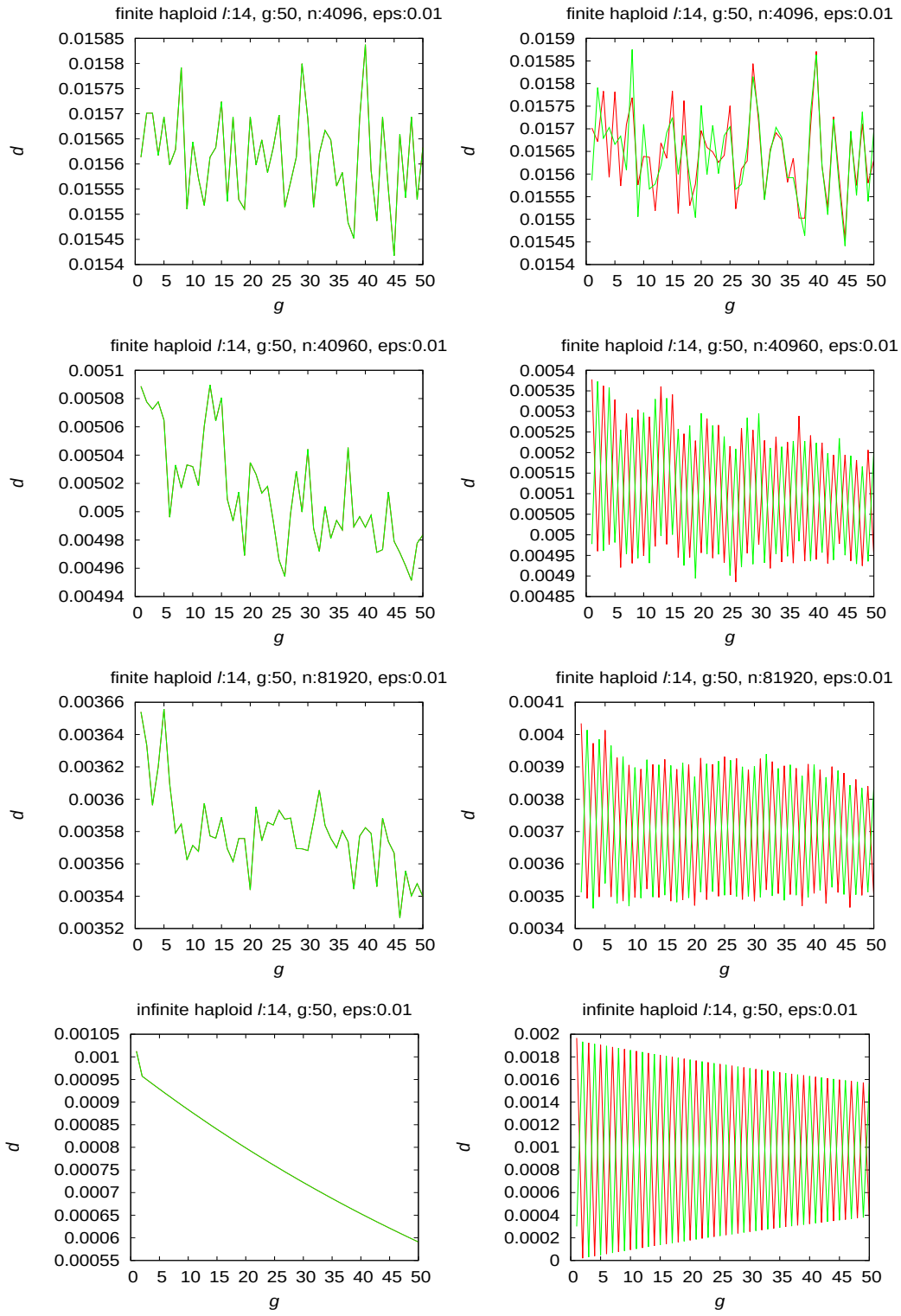


Figure 3.52: Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 14$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

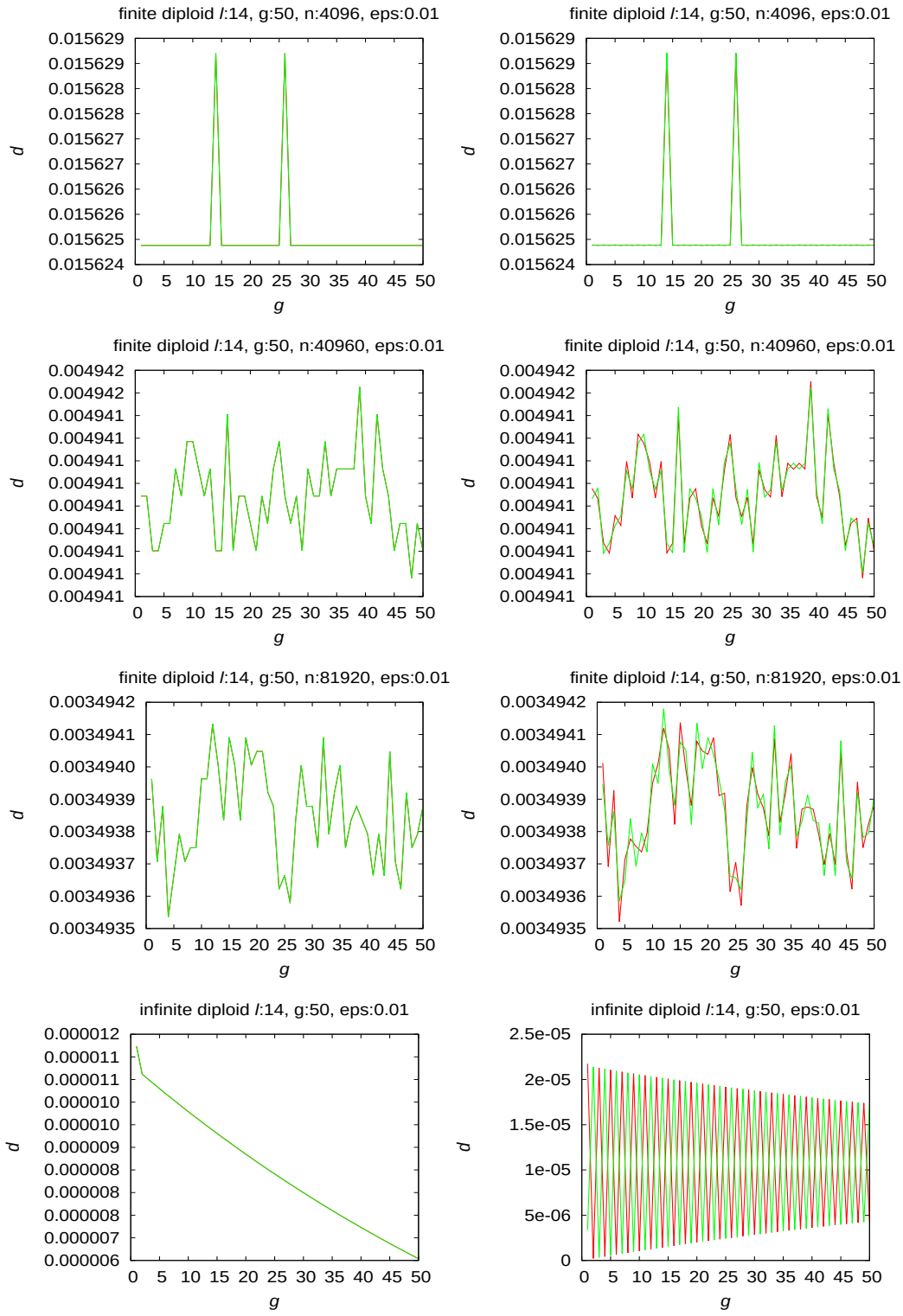


Figure 3.53: Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 14$ and $\epsilon = 0.01$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

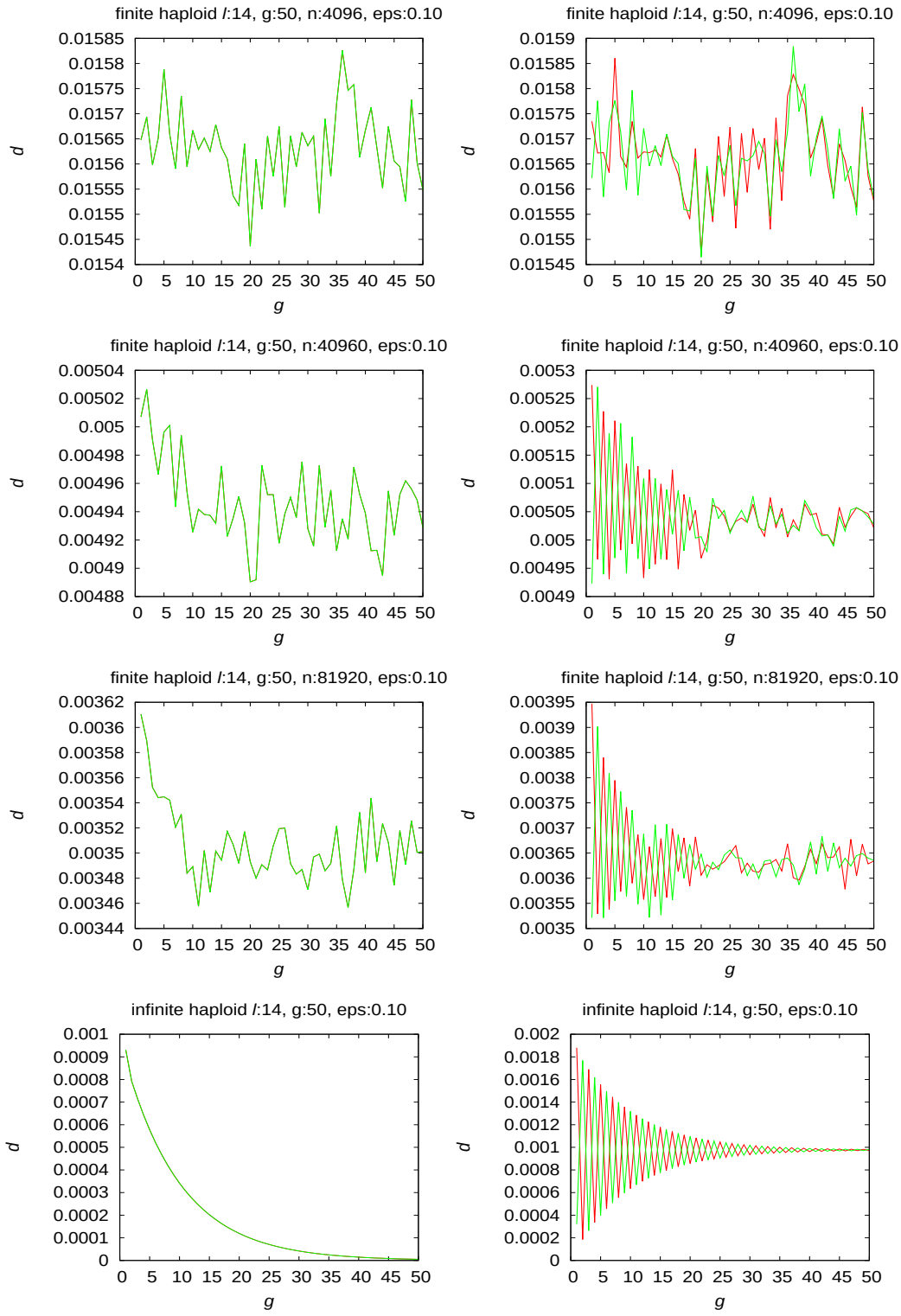


Figure 3.54: Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 14$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

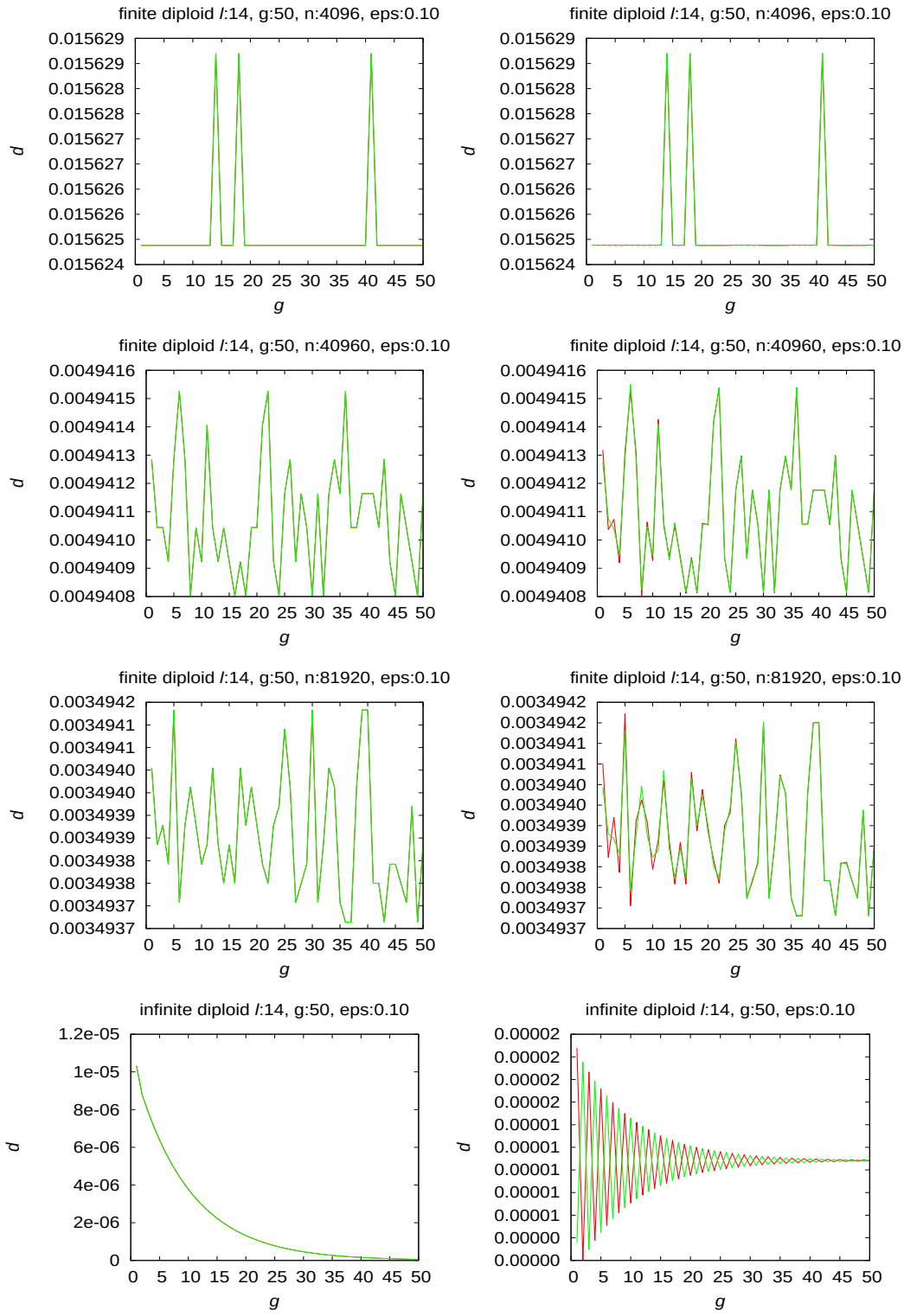


Figure 3.55: Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 14$ and $\epsilon = 0.1$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite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

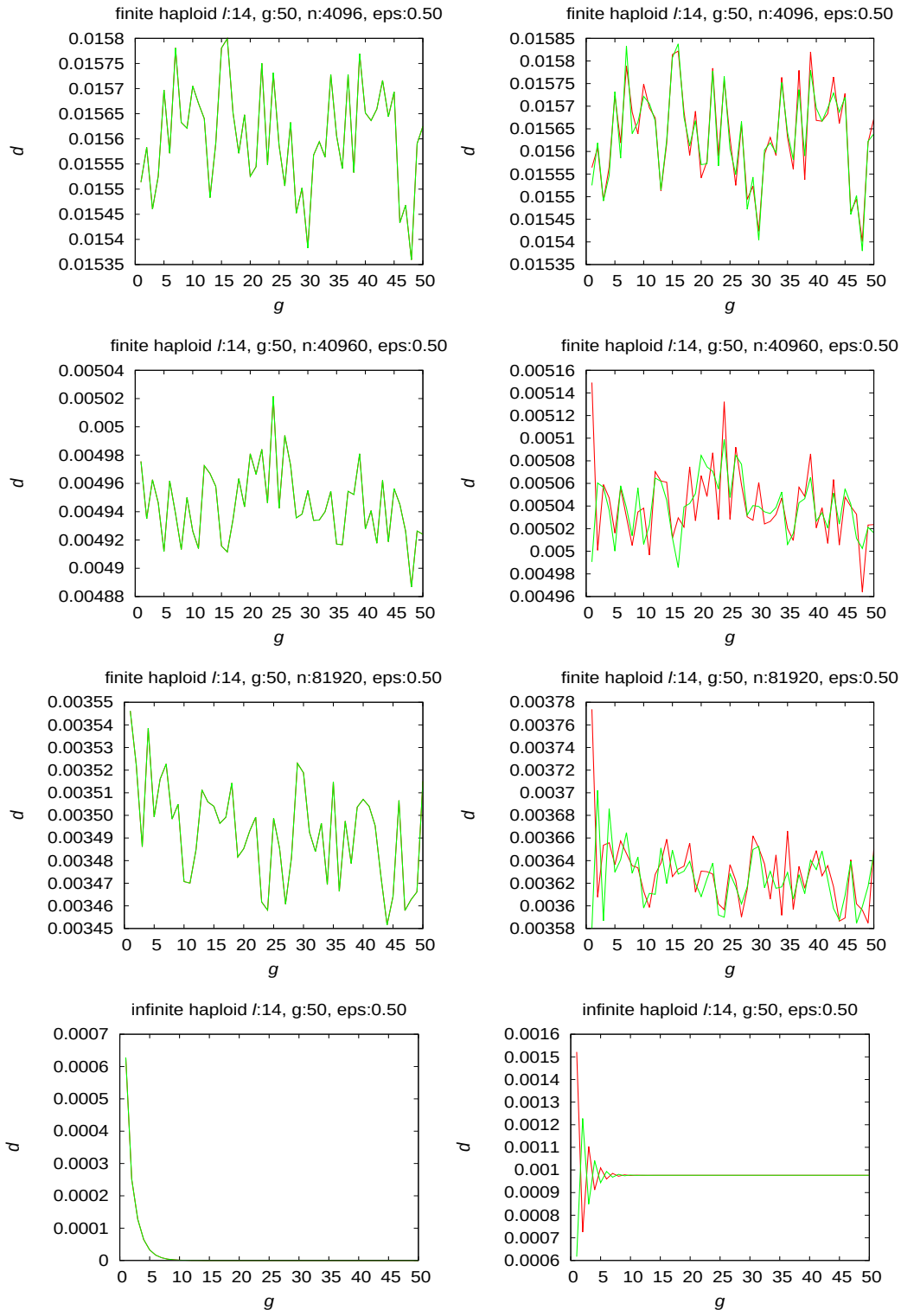


Figure 3.56: Infinite and finite haploid population oscillation behavior in case of violation in χ for genome length $\ell = 14$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite population of size N or infinite population to limits without violation.

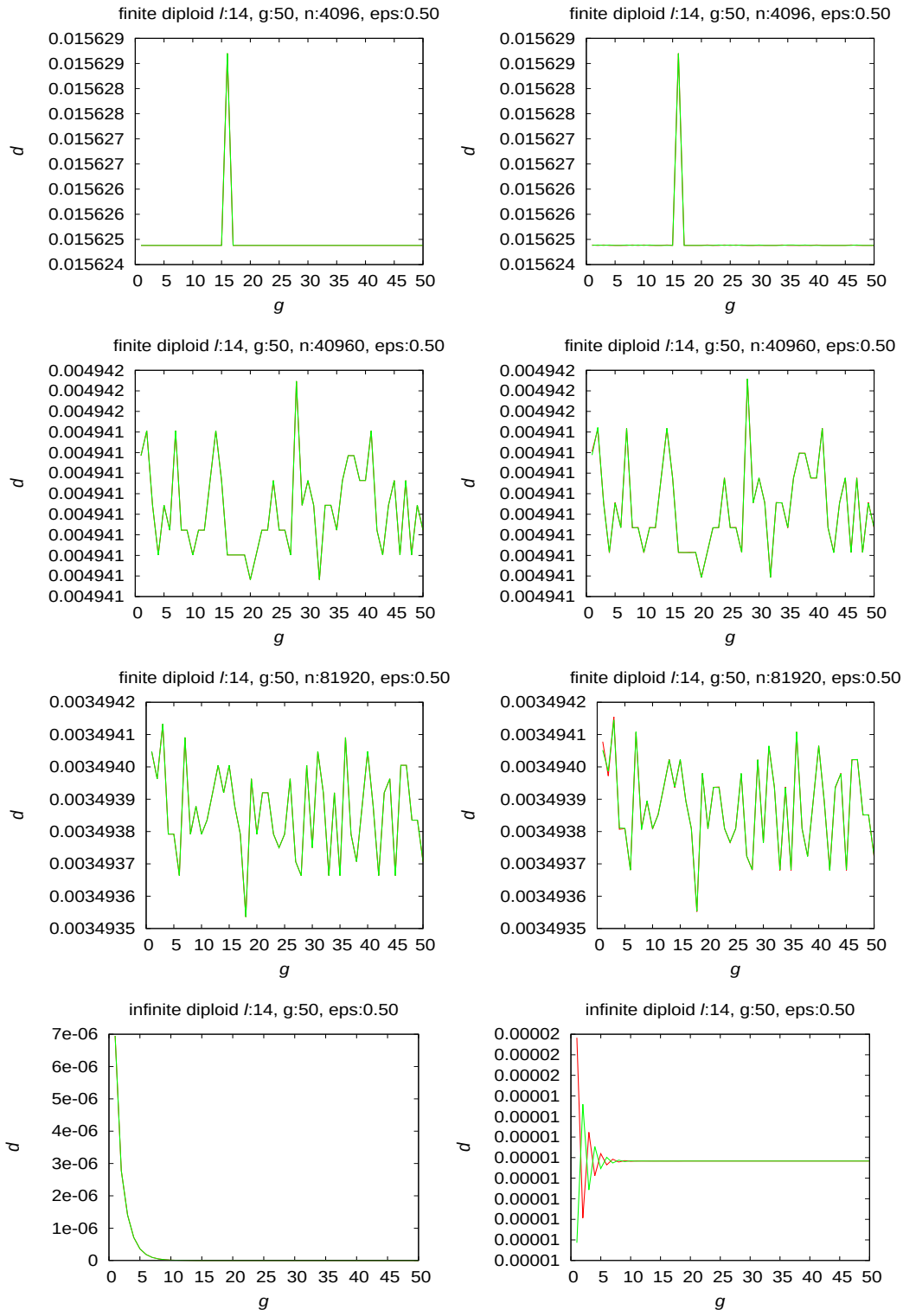


Figure 3.57: Infinite and finite diploid population oscillation behavior in case of violation in χ for genome length $\ell = 14$ and $\epsilon = 0.5$: In left column, d is distance of finite population of size n or infinite population to limit for g generations. In right column, d is distance of finite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

Left column of graphs in above figures for violation in $\boldsymbol{\mu}$ and in $\boldsymbol{\chi}$ shows distance of finite and infinite population to evolutionary limits with violation and right column shows distance finite and infinite population to evolutionary limits without violation.

Graphs in right column give picture of oscillating behavior of population in presence of violation in $\boldsymbol{\mu}$ or $\boldsymbol{\chi}$ distribution where distance of population to limits with no violation. Both finite and infinite population oscillate in presence of violation, however, in case of infinite population, ripples die out quickly as generation increases and ceases to oscillate, giving graph a tapering shape and in case of finite population, even though amplitudes of ripples decreased, ripples didn't die out completely.

Change in oscillating behavior of population with change in $\boldsymbol{\epsilon}$ values $\{0.01, 0.1, 0.5\}$ were also studied. Results show ripples damp out faster with increase $\boldsymbol{\epsilon}$. With smaller values of $\boldsymbol{\epsilon}$, oscillations were sharper and as value of $\boldsymbol{\epsilon}$ increased, rate of damping of ripples was quicker. Error $\boldsymbol{\epsilon}$ introduced to $\boldsymbol{\mu}$ or $\boldsymbol{\chi}$ distribution creates new masks (different than in case of without violation) to be used in mutation or crossover during transmission. With small $\boldsymbol{\epsilon}$, the probablility of using the new masks available due to violation is very small and those masks might not be used at all during crossover or mutation in finite population and with higher values of $\boldsymbol{\epsilon}$, those new masks have higher chance of usage during mutation or crossover which cause oscillation to damp out quickly or cause no oscillation at all. In figures for violation in $\boldsymbol{\mu}$ distribution, with $\boldsymbol{\epsilon} = 0.01$, oscillation is clearly visible and ripples are sharper; with $\boldsymbol{\epsilon} = 0.1$ oscillation was visible but ripples were damping out quickly; with $\boldsymbol{\epsilon} = 0.5$, oscillation was very minimal or not visible at all. As population size increases, chance of new masks created due to violation in $\boldsymbol{\mu}$ or $\boldsymbol{\chi}$ distribution also increases, thus, ripples damp out more quickly.

Graphs in left column show distance between finite population and limit $\boldsymbol{z}^*$ with violation decreases as finite population size increases and shows behavior similar to infinite population behavior as finite population reach large number. Simulation results show infinite population converges to limit $\boldsymbol{z}^*$ quicker with increase in $\boldsymbol{\epsilon}$. The

distance data in case of both $\boldsymbol{\mu}$ and $\boldsymbol{\chi}$ distribution violation with different values of $\boldsymbol{\epsilon}$ for different finite population size N are tabulated below.

Table 3.3: Experimental distance measured for violation in $\boldsymbol{\mu}$: ℓ is genome length, ϵ is error introduced to $\boldsymbol{\mu}$ for violation, $\{d', d'', d'''\}$ are distance measured for population size $\{4096, 40960, 81920\}$ respectively

ϵ	case	ℓ	d'	d''	d'''
0.01	haploid	8	0.017614	0.009411	0.009261
		10	0.016812	0.008761	0.007667
		12	0.016099	0.006417	0.005315
		14	0.015690	0.005141	0.003821
	diploid	8	0.015635	0.004994	0.003582
		10	0.015631	0.004952	0.003508
		12	0.015625	0.004942	0.003495
		14	0.015625	0.004941	0.003494
0.1	haploid	8	0.015805	0.005400	0.004080
		10	0.015825	0.005281	0.003938
		12	0.015672	0.005064	0.003649
		14	0.015645	0.004969	0.003541
	diploid	8	0.015631	0.004946	0.003499
		10	0.015623	0.004942	0.003508
		12	0.015625	0.004941	0.003495
		14	0.015625	0.004941	0.003494
0.5	haploid	8	0.016364	0.005641	0.004239
		10	0.016130	0.005491	0.004048
		12	0.015743	0.005079	0.003642
		14	0.015736	0.005109	0.003684
	diploid	8	0.015630	0.004952	0.003513
		10	0.015626	0.004944	0.003497
		12	0.015626	0.004941	0.003494
		14	0.015625	0.004941	0.003494

Table 3.4: Experimental distance measured for violation in $\boldsymbol{\chi}$: ℓ is genome length, ϵ is error introduced to $\boldsymbol{\chi}$ for violation, $\{d', d'', d'''\}$ are distance measured for population size $\{4096, 40960, 81920\}$ respectively

ϵ	case	ℓ	d'	d''	d'''
0.01	haploid	8	0.018631	0.015022	0.011465
		10	0.015839	0.006201	0.005093
		12	0.015834	0.005639	0.004518
		14	0.015614	0.005012	0.003579
	diploid	8	0.015655	0.005064	0.003658
		10	0.015625	0.004945	0.003498
		12	0.015625	0.004942	0.003494
		14	0.015625	0.004941	0.003494
0.1	haploid	8	0.016264	0.006084	0.005129
		10	0.015677	0.005131	0.003709
		12	0.015671	0.005087	0.003667
		14	0.015632	0.004946	0.003506
	diploid	8	0.015617	0.004958	0.003518
		10	0.015625	0.004941	0.003494
		12	0.015625	0.004941	0.003494
		14	0.015625	0.004941	0.003494
0.5	haploid	8	0.015636	0.005062	0.003629
		10	0.015528	0.004944	0.003529
		12	0.015673	0.004968	0.003515
		14	0.015604	0.004945	0.003493
	diploid	8	0.015631	0.004942	0.003500
		10	0.015626	0.004941	0.003494
		12	0.015625	0.004941	0.003494
		14	0.015625	0.004941	0.003494

From table 3.3, average distance calculated for finite population size 4096 is 0.015861, for size 40960 is 0.005464 and for size 81920 is 0.004123. From table 3.4, average distance calculated for finite population size 4096 is 0.015797, for size 40960 is 0.005520 and for size 81920 is 0.004040. These results show experimental distance between finite population and the limit with violation closely follows expected single step distance between finite and infinite population given by 3.1 and the distance decreased as $1/\sqrt{N}$.

3.6 Summary

In this chapter, we described limits predicted by Vose for infinite population, and necessary and sufficient condition for population to converge in to periodic orbits. Through experiment, we showed finite population also oscillate under condition stated for infinite population to converge in to periodic orbits and converge to infinite population evolutionary limits as population size increases. Then we studied effect of violation in condition for population to converge in to periodic orbit on behavior of infinite and finite population through simulation.

Chapter 4

Conclusion

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

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

Evolutionary limits predicted by Vose for infinite population were explored and analysed. Simulations showed when necessary condition in $\boldsymbol{\mu}$ and $\boldsymbol{\chi}$ distribution is met, finite population also showed oscillating behavior and converge to evolutionary limits for infinite population. In case of violation in the condition, infinite population

ceased to oscillate but finite population when $\boldsymbol{\epsilon}$ introduced to $\boldsymbol{\mu}$ or $\boldsymbol{\chi}$ was not large, continued to oscillate.

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

Bibliography

Bibliography

Beauchamp, K. (1975). *Walsh functions and their applications*. Academic Press. 17

Bethke, A. D. (1980). *Genetic Algorithms As Function Optimizers*. PhD thesis, The University of Michigan. 2, 4

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. 18

Crow, J. and Kimura, M. (1970). *An introduction to population genetics theory*. New York, Evanston and London: Harper & Row. 26

Geiringer, H. (1944). On the probability theory of linkage in mendelian heredity. *Ann. Math. Stat.*, 15(1):25–27. 10, 14

Goldberg, D. E. (1987). Simple genetic algorithms and the minimal, deceptive problem. *Genetic algorithms and simulated annealing*, 74:74–88.

Goldberg, D. E. (1989a). Genetic algorithms and walsh functions: Part i, a gentle introduction. *Complex systems*, 3(2):129–152. 4

Goldberg, D. E. (1989b). Genetic algorithms and walsh functions-partii: Deception and its analysis. *Complex systems*, 3(153–171). 4

Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science*, 28(706):49–50. 10, 11

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

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

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

Mendel, G. (1865). Versuche über plflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn*, IV:3–47. 11

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

Nix, A. E. and Vose, M. D. (1992). Modeling genetic algorithms with markov chains. *Annals of Mathematics and Artificial Intelligence*, 5(1):79–88. 3

Shanks, J. L. (1969). Computation of the fast walsh-fourier transform. *IEEE Trans. Comput.*, 18(5):457–459. 5, 18, 22

Vose, M. and Liepins, G. E. (1991). Punctuated equilibria in genetic search. *Complex systems*, 5(1):31–44. 2

Vose, M. D. (1999). *The simple genetic algorithm: foundations and theory*, volume 12. MIT press. 4, 5, 6, 7, 13, 25, 26

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. 4, 14, 15, 16, 19

Wikipedia, C. (2016). Jensen's inequality. 26

Appendix

Vita

Vita goes here...