Efficient Simulation Of A Simple Evolutionary System

Mahendra Duwal Shrestha

The University Of Tennessee

April 6, 2017

Outline

Part-I: Efficient Computations

Population

Population P: a collection of length ℓ binary strings

Population vector: $\mathbf{p}_j$ proportion of string j in the population

If $P = \langle 00, 01, 01, 10, 11, 11 \rangle$, then $\mathbf{p}_3 = 2/6 = 1/3$

$\mathbf{1}$ & $\mathcal{R}$

$\mathbf{1}$ is column vector of all 1s of ℓ components, $\langle 1, 1, \cdots, 1 \rangle$

$\mathcal{R}$ denotes the set of binary strings of length ℓ

$|\mathcal{R}| = Z = 2^{\ell}$

Addition and multiplication of elements in $\mathcal{R}$ are bitwise operations modulo 2

$x = 1101$, $y = 1010$

$x + y = 1101 + 1010 = 0111$

$xy = 1101 \cdot 1010 = 1000$

$\bar{x} = x + \mathbf{1} = 0010$

Crossover

Crossover : Choose parents u and v, exchange bits using crossover mask m:

$u' = um + v\bar{m}, v' = u\bar{m} + vm$

$u = \mathbf{1100}, v = 1101, m = 1100$

$\{\mathbf{1100}, 1101\} \rightarrow \{\mathbf{11}00 + 0001,\ 00\mathbf{00} + 1100\} \rightarrow \{\mathbf{11}01, 11\mathbf{00}\}$

$\chi_m =$ probability of using crossover mask m

Mutation

Mutation: Flip bits using mutation mask m:

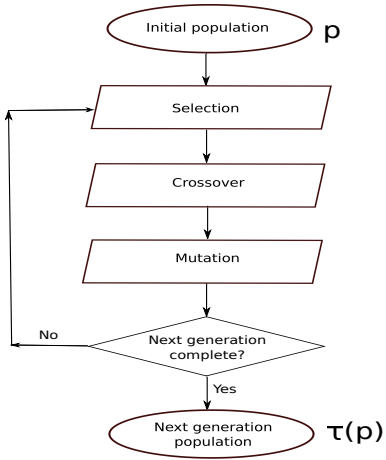
$x \to x + m$

$x = 1100, m = 0001$

$1100 \to 1100 + 0001 \to 1101$

$\boldsymbol{\mu}_m =$ probability of using mutation mask m

Finite Population GA (Haploid)



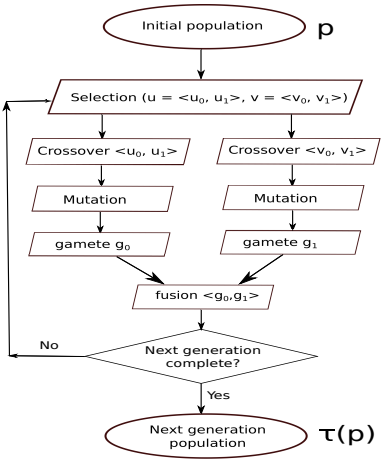
Randomly select parents u and v

Crossover u and v to produce u' and v'

Keep one of u', v' , and mutate to produce gamete g

Repeat above to form next generation

Finite Population GA (Diploid)



Random Heuristic Search

τ is a stochastic transition rule that maps $\mathbf{p}$ to

$\mathbf{p'} \in \Lambda_N = \{\langle \frac{X_1}{N}, \cdots, \frac{X_Z}{N} \rangle \mid X_i \in \mathbb{Z}^{\geq 0}, \sum X_i = N\}$

$\tau(\mathbf{p})$ cannot be predicted with certainty

$\mathbf{p}, \tau(\mathbf{p}), \tau^2(\mathbf{p}), \cdots$ forms Markov chain

Infinite Population Model

Population modeled as vector $\mathbf{p} \in \Lambda = \{\langle \mathbf{p}_1, \cdots, \mathbf{p}_Z \rangle \mid \mathbf{p}_i \geq 0,\ \sum \mathbf{p}_i = 1\}$

$\mathcal{G}$ maps $\mathbf{p}$ to the next generation $\mathbf{p}'$
$\mathcal{G}(\mathbf{p})_j$ = proportion of string j in the next generation
The infinite population model

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

$\mathcal{G}(\mathbf{p}) = \mathcal{E}(\tau(\mathbf{p}))$

The variance is

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

Diploid Population Model

Diploid genome: $\alpha = \langle \alpha_0, \alpha_1 \rangle$

Population vector $\mathbf{q}_\alpha$: prevalence of diploid α

$t_\alpha(g)$: probability that gamete g is produced from parent α

$$\mathbf{q}'_\gamma = \sum_\alpha \mathbf{q}_\alpha \, t_\alpha(\gamma_0) \sum_\beta \mathbf{q}_\beta \, t_\beta(\gamma_1)$$

Diploid Model Reduction to Haploid Model

Diploids in terms of haploids:

$$\mathbf{q}_{\langle \gamma_0, \gamma_1 \rangle} \;=\; \mathbf{p}_{\gamma_0}\, \mathbf{p}_{\gamma_1}$$

Haploids in terms of diploids:

$$\mathbf{p}_g \;=\; \frac{1}{2} \sum_{\alpha_0,\, \alpha_1} \mathbf{q}_{\langle \alpha_0, \alpha_1 \rangle} ([g = \alpha_0] + [g = \alpha_1])$$

Evolution equation in terms of haploid distribution $\mathbf{p}$:

$$\mathbf{p}'_{\gamma_0} \;=\; \sum_{\alpha_0,\, \alpha_1} \mathbf{p}_{\alpha_0}\, \mathbf{p}_{\alpha_1}\, t_{\langle \alpha_0,\, \alpha_1 \rangle}(\gamma_0)$$

Matrix form:

$$\mathbf{p}'_g \;=\; \mathbf{p}^T M_g\, \mathbf{p} \quad \text{where } \Big(M_g\Big)_{u,v} \;=\; t_{\langle u,v \rangle}(g)$$

Specialization to Vose's Haploid Model

Mutation distribution:

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

Crossover distribution:

$$\boldsymbol{\chi}_i = \begin{cases} \chi c_i & \text{if } i > 0 \\ 1 - \chi + \chi c_0 & \text{if } i = 0 \end{cases} \qquad c_i = 2^{-\ell}$$

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

Walsh Basis

$$W_{n,t} = Z^{-1/2}(-1)^{n^T t} \text{ where } Z = 2^\ell$$

$$\widehat{w} = Ww \qquad O(Z \log Z)$$

$$\widehat{A} = WAW \qquad O(Z^2 \log Z)$$

```
procedure FWT
    n = 2^d ← size of array X where d is positive integer
    for i = 0 to d − 1 do
        m = n/2^i
        z = m/2
        for j = 0 to 2^i − 1 do
            for k = 0 to z − 1 do
                t1 = m × j + k
                t2 = m × j + z + k
                a = X[t1]
                b = X[t2]
                X[t1] = a + b
                X[t2] = a − b
            end for
        end for
    end for
    return X
end procedure
```

Computations in Walsh basis (Vose's Haploid Model)

Mixing matrix M in Walsh basis

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

Evolution eqn in Walsh basis

$$\widehat{\mathbf{p}}_g' \;=\; 2^{\ell/2} \sum_{i \in g\mathcal{R}} \widehat{\mathbf{p}}_i\,\widehat{\mathbf{p}}_{i+g}\,\widehat{M}_{i,i+g} \qquad \text{where} \quad g\mathcal{R} = \{g\,i \mid i \in \mathcal{R}\}$$

Computational Comparison

In Walsh basis:

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

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

Before Walsh basis:

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

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

Computational Significance

Reduction to haploid model and Walsh basis simplifiy computation, which otherwise for diploid case would have been impractical

Only one mixing matrix as opposed to 2^{ℓ} is needed to compute next generation

For $\ell = 14$, using 2^{14} matrices with each having $2^{14} \cdot 2^{14}$ entries would require 32 TB of memory, whereas one mixing matrix requires only 2 GB

Distance

Naive implementation:

$$\|\mathbf{f}-\mathbf{q}\|^2 \;=\; \sum_{\alpha}(\mathbf{f}_\alpha-\mathbf{q}_\alpha)^2 \longrightarrow 2^\ell\cdot 2^\ell \text{ terms}$$

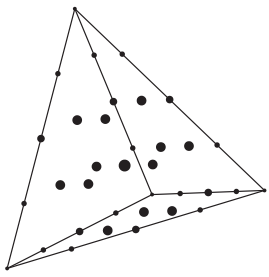
Our implementation:

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

$$\|\mathbf{f}-\mathbf{q}\|^2 \;=\; \sum_{g}^{2}(\mathbf{p}_g)^2 + \sum_{\alpha\in S_\mathbf{f}} \mathbf{f}_\alpha(\mathbf{f}_\alpha-2\mathbf{q}_\alpha) \to 2^\ell + |S_\mathbf{f}| \text{ terms}$$

Part-II: Applications

Population Points



Finite population points are represented by dots

Infinite population can be any where in the space

$$\sup_{\boldsymbol{\xi}\in\Lambda} \inf_{\mathbf{p}\in\Lambda_N} \|\boldsymbol{\xi} - \mathbf{p}\| \;=\; O(1/\sqrt{N})$$

Question 1: Distance Between Finite and Infinite Population

Chebyshev's inequality: suggests that perhaps

$$\|\tau(\mathbf{p}) - \mathcal{G}(\mathbf{p})\| \leq \frac{k}{\sqrt{N}} \quad \text{with probability approaching 1}$$

Jensen's inequality:

$$\mathcal{E}(\|\tau(\mathbf{p}) - \mathcal{G}(\mathbf{p})\| \leq \frac{\sqrt{1 - \|\mathcal{G}(\mathbf{p})\|^2}}{\sqrt{N}}$$

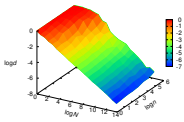
Geometric point of view:

$$\sup_{\boldsymbol{\xi} \in \Lambda} \inf_{\mathbf{p} \in \Lambda_N} \|\boldsymbol{\xi} - \mathbf{p}\| = O(1/\sqrt{N})$$

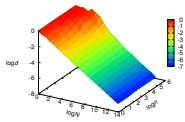
Can the distance decrease in practice like $1/\sqrt{N}$?

Convergence: Results

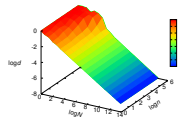
$$\chi = 0.1,\ \mu = 0.001$$



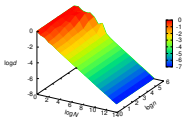
(a) $\ell = 4$



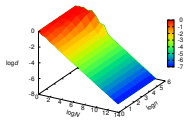
(b) $\ell = 6$



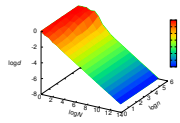
(c) $\ell = 8$



(d) $\ell = 10$



(e) $\ell = 12$

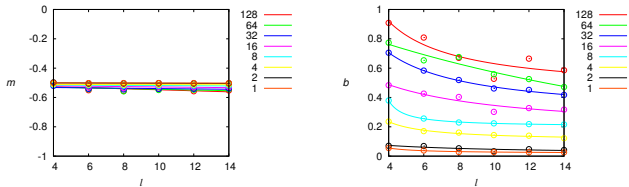


(f) $\ell = 14$

Figure : Convergence of finite population behavior

Regression

$$\log d = m \log N + b$$



(a) Slope m **(b)** Intercept b

Figure : Regression parameter for generation $n \in \{1, 2, 4, 8, 16, 32, 64, 128\}$

$d \approx N^m e^b$

From figure (a) above, $m \approx -(\frac{1}{2})$

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

Convergence: Conclusion

The distance between finite and infinite population can decrease like $1/\sqrt{N}$

Question 2

Finite Population Oscillation

Limits

Infinite population evolution

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

may converge to a fixed point

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

But under some circumstances, evolution converges to a periodic orbit that oscillates between two fixed points, $\mathbf{p}^*$ and $\mathbf{q}^*$

$$\mathbf{p}^* = \lim_{n\to\infty} \mathcal{G}^{2n}(\mathbf{p}), \quad \mathbf{q}^* = \lim_{n\to\infty} \mathcal{G}^{2n+1}(\mathbf{q})$$

Periodic Orbit: Necessary and Sufficient Conditions

For some $g \in \mathcal{R}, g \neq 0$

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

Infinite populations converge to a periodic orbit

Can finite populations also exhibit oscillation from random initial populations?

Previous Works on Oscillation

Akin (1982) proved existence of cycling for continuous-time 2-bit diploid model

Hasting (1981) studied cycling in populations with infinite 2-bit diploid population model

Wright and Bidwell (1997) provided examples of cycling in an infinite haploid model with crossover and mutation for 3 bit and 4 bit populations

Wright and Agapie (2001) described cycling in infinite population for up to 4 bits, and also presented data for cycling in finite population

Difference From Previous Works

Akin considers continuous time, we consider discrete time

Hastings' study is limited to two bits, and only crossover, no mutation

Wright and Bidwell consider specific parameter values

Wright and Agapie use dynamic mutation

Simulation

Simulations were run for both haploid and diploid populations

Random initial population

$\ell \in \{8, 10, 12, 14\}$

$N \in \{4096, 40960, 81920\}$

To visualize, distance of population to fixed points $\mathbf{p}^*$ and $\mathbf{q}^*$ is plotted

Oscillation: Results

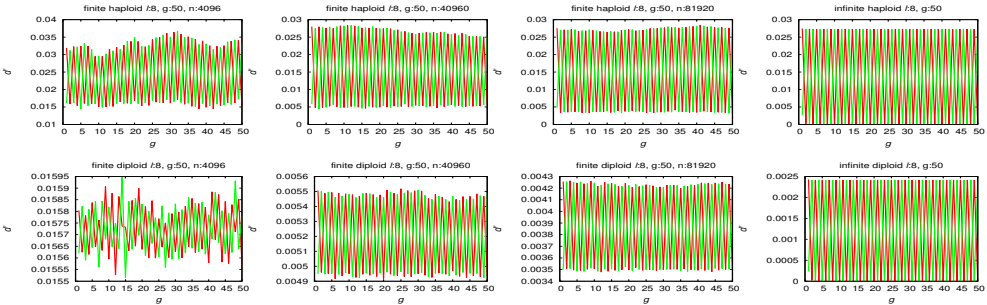


Figure : Infinite and finite population behavior for genome length $\ell = 8$

Oscillation: Conclusion

Finite populations can exhibit approximate oscillations

Question 3

Oscillation Under Mutation-Violation

For all g,

$$-1 \neq \sum_j (-1)^{g^T j} \boldsymbol{\mu}_j$$

No periodic orbits for infinite population

Mutation-Violation

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

$\boldsymbol{\mu}_i := (1 - \epsilon)\boldsymbol{\mu}_i$

This modification makes the Markov chain regular

No periodic orbits for finite population

Can finite population exhibit approximate oscillations?

Simulation

$\epsilon \in \{0.01, 0.1, 0.5\}$

$\ell \in \{8, 10, 12, 14\}$

$N \in \{4096, 40960, 81920\}$

Distance to limits p^* and q^* without violation ($\epsilon = 0$) are plotted

Mutation-Violation: Results

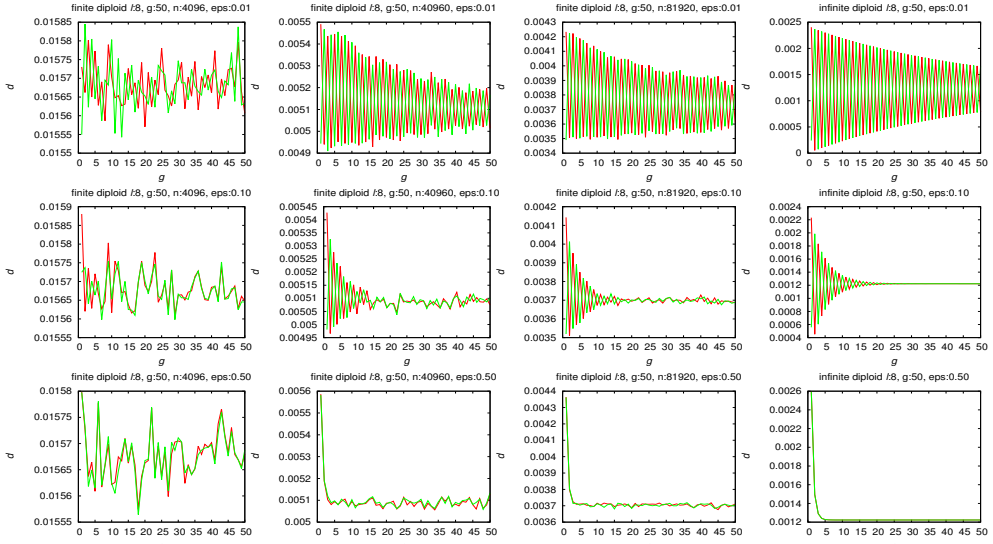


Figure : Infinite and finite diploid population behavior for μ violation and $\ell = 8$

Mutation-Violation: Conclusion

Finite populations can exhibit approximate oscillation when mutation-violation is small

If violation is large, then oscillation can decrease

Question 4

Oscillation under Crossover-Violation

For all g,

$$1 \neq \sum_{k \in \bar{g}\mathcal{R}} \boldsymbol{\chi}_{k+g} + \boldsymbol{\chi}_k$$

No periodic orbit exists for infinite population

Crossover-Violation

$$\chi_i := (1-\epsilon)\chi_i$$

$$\chi_j := \epsilon \qquad j \text{ is chosen such that } j \notin \bar{g}\mathcal{R}$$

Can finite populations exhibit approximate oscillation?

Simulation

$\epsilon = \{0.01, 0.1, 0.5\}$

$\ell = \{8, 10, 12, 14\}$

$N = \{4096, 40960, 81920\}$

Distance to limits $\mathbf{p}^*$ and $\mathbf{q}^*$ without violation ($\epsilon = 0$) are plotted

Crossover-Violation: Results

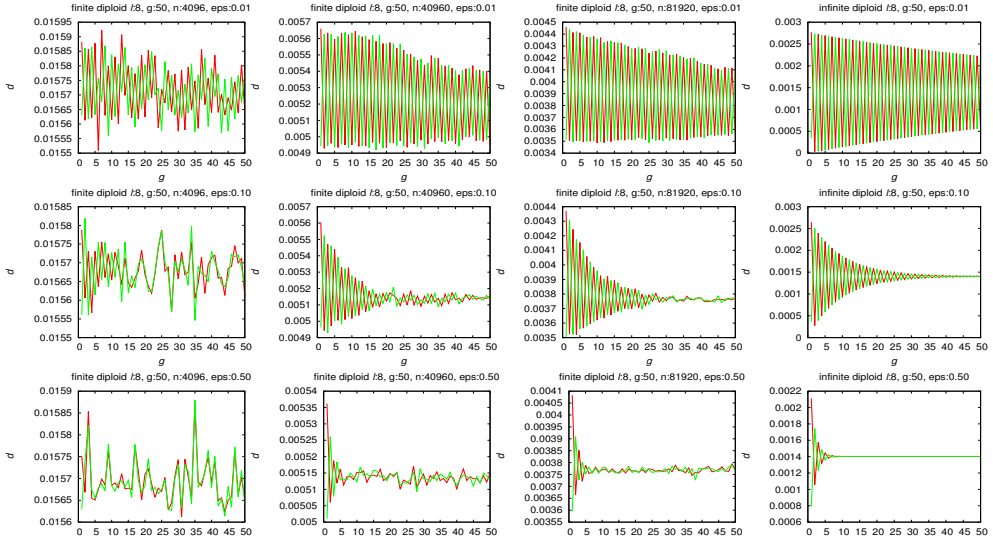


Figure : Infinite and finite diploid population behavior for χ violation and $\ell = 8$

Crossover-Violation: Conclusion

Finite populations can exhibit approximate oscillation when crossover-violation is small

If violation is large, then oscillation can decrease

Conclusion

By reducing to haploid case, Vose's haploid model makes computation efficient in diploid case

Distance between finite population and infinite population can decrease like $1/\sqrt{N}$

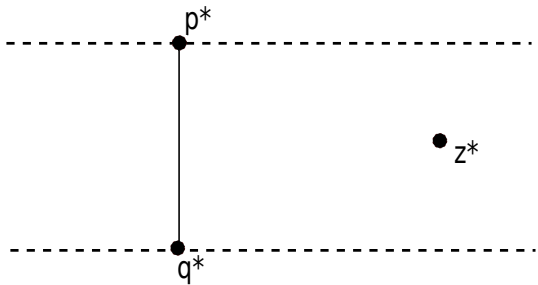
When infinite populations oscillate, finite populations can exhibit approximate oscillation

Finite populations can exhibit approximate oscillation for small mutation-violation

Finite populations can exhibit approximate oscillation for small crossover-violation

Part-III: Future Work

Violation-limit ($\mathbf{z}^*$) between non-violation-limits ($\mathbf{p}^*$ and $\mathbf{q}^*$)



$\mathbf{z}^*$ is between and equidistant from $\mathbf{p}^*$ and $\mathbf{q}^*$

Thank You!!

Chebyshev's Inequality

Let $\epsilon = f(r)/\sqrt{r}$, where $f(r)$ grows arbitrarily slowly such that

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

and

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

From Chebyshev's inequality,

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

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

Jensen's Inequality

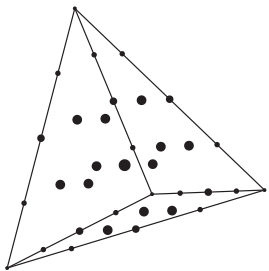
Let η be the random variable $\|\tau(\mathbf{p}) - \mathcal{G}(\mathbf{p})\|$, and convex function be $\phi(x) = x^2$

Then from Jensen's Inequality,

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

This also suggests the distance might decrease as $1/\sqrt{r}$

Population Points



Finite populations are represented by dots

Infinite population can be any where in the space

Distance between a finite population and an infinite population is $O(1/\sqrt{r})$

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

Computation of $\mathbf{p}^*$ and $\mathbf{q}^*$

$\mathbf{p}^* = \lim_{n\to\infty} \mathcal{M}^{2n}(\mathbf{p})$ $\qquad$ $\mathbf{q}^* = \lim_{n\to\infty} \mathcal{M}^{2n+1}(\mathbf{q})$

Let $S_g = g\mathcal{R} \setminus \{\mathbf{0}, g\}$, $|g|$ be the number of non zero bits in g, and

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

Moreover,

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

Computation of $\mathbf{p}^*$ and $\mathbf{q}^*$ continued...

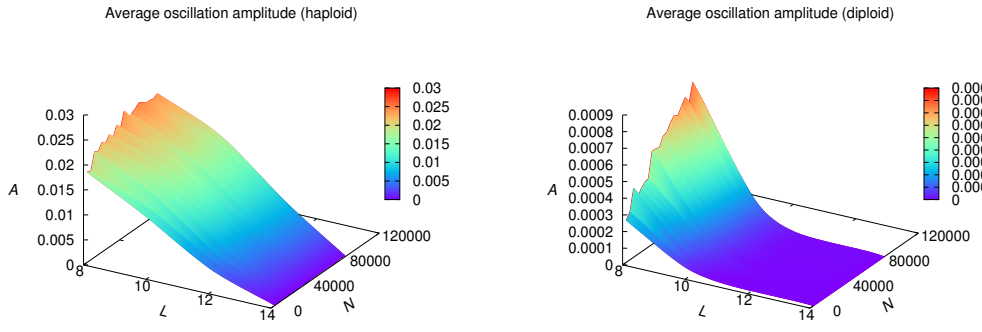
The limits in the Walsh basis in form of recursive equations

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

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

Limits $\widehat{\mathbf{p}^*}_g$ and $\widehat{\mathbf{q}^*_g}$ can be computed considering gth components in order of increasing $|g|$.

Oscillation Amplitude



(a) **(b)**

Figure : Average oscillation amplitude