Linear Algebra of Eigen's Quasispecies Model

Artem Novozhilov

Department of Mathematics
North Dakota State University

Midwest Mathematical Biology Conference,
University of Wisconsin — La Crosse

May 17, 2014

Collaborators:

Yuri S. Semenov
Moscow State University of Railway Engineering, Moscow, Russia

Alexander S. Bratus
Lomonosov Moscow State University
Moscow, Russia

Outline:

- Motivation and historic outline
- Mathematical problem
- Known results
- New results

M. Eigen, Naturwisenschaften, 58(10), 1971: 465–523

DIE NATURWISSENSCHAFTEN

58. Jahrgang, 1971 Heft 10 Oktober

Selforganization of Matter and the Evolution of Biological Macromolecules

MANFRED EIGEN*

Max-Planck-Institut für Biophysikalische Chemie,
Karl-Friedrich-Bonhoeffer-Institut, Göttingen-Nikolausberg

Quasispecies theory:

Manfred Eigen, born 1927

- M. Eigen, Naturwisenschaften, 58(10), 1971:465–523
- M. Eigen, P. Schuster, The Hypercycle, Springer, 1979
- M. Eigen, J. McCaskill, P. Schuster, J Phys Chem, 92(24), 1988:6881-6891

Mathematical side of the story:

The variables of the dynamical system are the concentrations of individual polynucleotide sequences: $[I_i] = c_i(t)$. We are interested, essentially, in the relative concentrations of the different species

$$x_i(t) = c_i(t) / \sum_{i=1}^{n} c_i(t); \quad i = 1, 2, \ldots, n \tag{12}$$

The resulting kinetic equations, around which quasi-species theory centers, are then

$$\mathrm{d}x_i(t)/\mathrm{d}t \equiv \dot{x}_i(t) = (W_{ii} - \bar{E}(t))x_i(t) + \sum_{k \neq i} W_{ik}x_k(t); \quad i, k = 1, 2, \ldots, n \tag{13}$$

The mean excess production

$$\bar{E}(t) = \sum_{i=1}^{n} x_i(t)E_i \tag{14}$$

of the population may be physically compensated by a dilution

Ref: M. Eigen, J. McCaskill, P. Schuster, J Phys Chem, 92(24), 1988: 6881-6891

Model statement:

Consider a population of sequences of fixed length N composed of two-letter alphabet, say, $\{0, 1\}$, therefore 2^N different sequences.

The population is subject to two evolutionary forces. First evolutionary force is *selection*, which is included in the system through the *Malthusian fitness*, defined here for simplicity as

$$m(\text{particular sequence } \sigma) = m(H_\sigma),$$

where H_σ is the Hamming norm of this sequence, i.e., number of 1s in sequence σ. In this way we do not distinguish between sequences with the same number of 1s and hence reduce the dimensionality of the problem from $2^N \times 2^N$ to $(N+1) \times (N+1)$. Hence, we consider at this point only *permutation invariant* fitness landscapes

$$\boldsymbol{M} = \mathrm{diag}(m_0, \ldots, m_N) \quad \text{or} \quad \boldsymbol{m} = (m_0, \ldots, m_N)^\top.$$

Model statement:

The second evolutionary force is *mutation*.

In particular, assuming $N+1$ classes of sequences, we have that the mutations μ_{ij} (i.e., the mutation rate from class j to class i) can be described by the matrix

$$\boldsymbol{\mathcal{M}} = (\mu_{ij}) = \mu \boldsymbol{Q} = \begin{bmatrix} -N & 1 & 0 & 0 & \dots & \dots & 0 \\ N & -N & 2 & 0 & \dots & \dots & 0 \\ 0 & N-1 & -N & 3 & \dots & \dots & 0 \\ 0 & 0 & N-2 & -N & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 0 & \dots & \dots & 2 & -N & N \\ 0 & 0 & \dots & \dots & 0 & 1 & -N \end{bmatrix},$$

where μ is the mutation rate per site per sequence per replication event.

Model statement:

Let $\boldsymbol{p}(t)$ denote the vector of frequencies of different classes of sequences, then, assuming uncoupled reproduction and mutation events, we arrive at

$$\dot{\boldsymbol{p}}(t) = (\boldsymbol{M} + \mu\boldsymbol{Q})\boldsymbol{p}(t) - \overline{m}(t)\boldsymbol{p}(t),$$

where

$$\overline{m}(t) = \boldsymbol{m} \cdot \boldsymbol{p}(t) = \sum_{i=0}^{N} m_i p_i(t)$$

is the mean population fitness.

This model is often called a paramuse of Crow–Kimura quasispecies model with permutation invariant fitness landscape.

Ref: Baake and Gabriel, Annual Reviews of Computational Physics VII, 1999: 203–264

Ref: Crow and Kimura, An introduction to population genetics theory, 1970

Elementary results:

The asymptotic behavior of the quasispecies model is determined by the equilibrium $\boldsymbol{p} = \lim_{t\to\infty} \boldsymbol{p}(t)$, which solves the eigenvalue problem

$$(\boldsymbol{M} + \mu\boldsymbol{Q})\boldsymbol{p} = \overline{m}\,\boldsymbol{p},$$

where

$$\overline{m} = \boldsymbol{m} \cdot \boldsymbol{p}.$$

By Perron–Frobenius theorem it follows that there is a unique positive solution $\boldsymbol{p} > 0$, which is the right eigenvector of $\boldsymbol{M} + \mu\boldsymbol{Q}$ corresponding to the simple real dominant eigenvalue $\lambda = \overline{m}$.

This vector $\boldsymbol{p}$ was called by Eigen the *quasispecies*. It is globally stable for the quasispecies system. We are mostly interested in properties of $\overline{m}$ and $\boldsymbol{p}$ depending on the fitness landscape $\boldsymbol{M}$ and mutation rate μ, therefore, we use the notation $\overline{m} = \overline{m}(\mu)$ and $\boldsymbol{p} = \boldsymbol{p}(\mu)$ for the mean fitness and equilibrium distribution.

Known results:

- Thompson and McBride, Math Biosciences, 21: 127–142, 1974

  The quasispecies model is essentially linear.
- Rumschitzki, J of Math Biol, 24: 667–680, 1987

  For zero epistasis fitness landscape the spectral properties of the Eigen evolutionary matrices can be inferred with the representation of $\boldsymbol{\mathcal{M}}$ and $\boldsymbol{M}$ using tensor products.
- Swetina and Schuster, Bioph Chem, 16: 329–345, 1982

  Numerical analysis of single peaked fitness landscape yields the *error threshold*.

Known results: The error threshold

Ref: Swetina and Schuster, Bioph Chem, 16: 329–345, 1982

Consider the single peaked fitness landscape

$$\boldsymbol{M} = \mathrm{diag}(m_0, 0, \ldots, 0), \quad m_0 > 0.$$

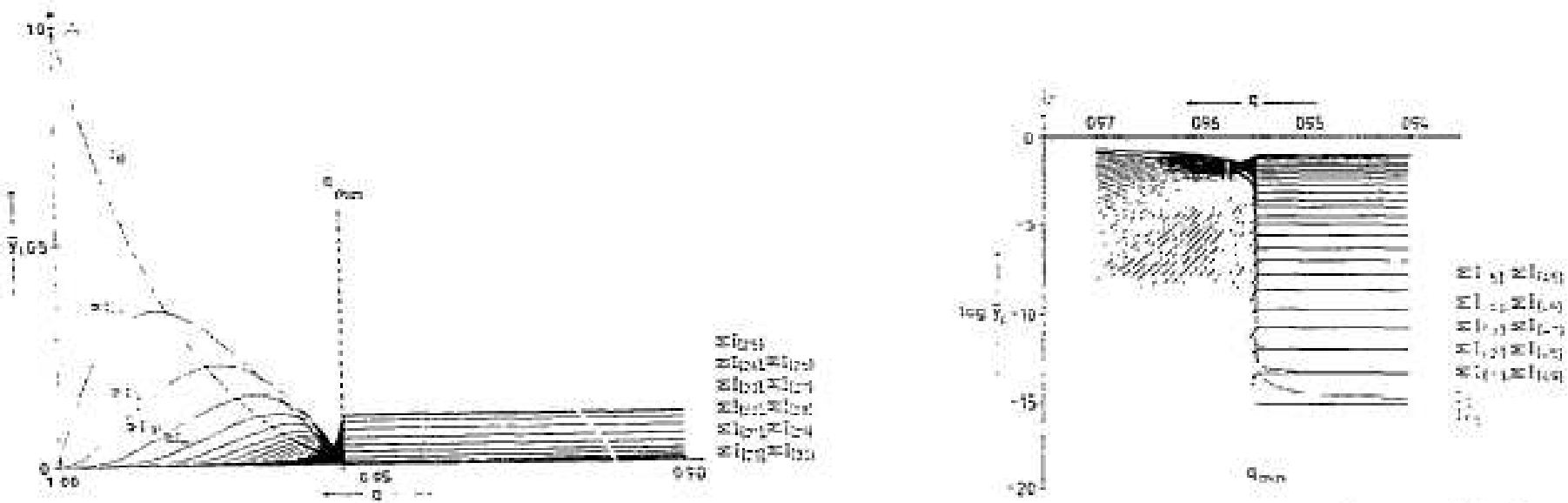
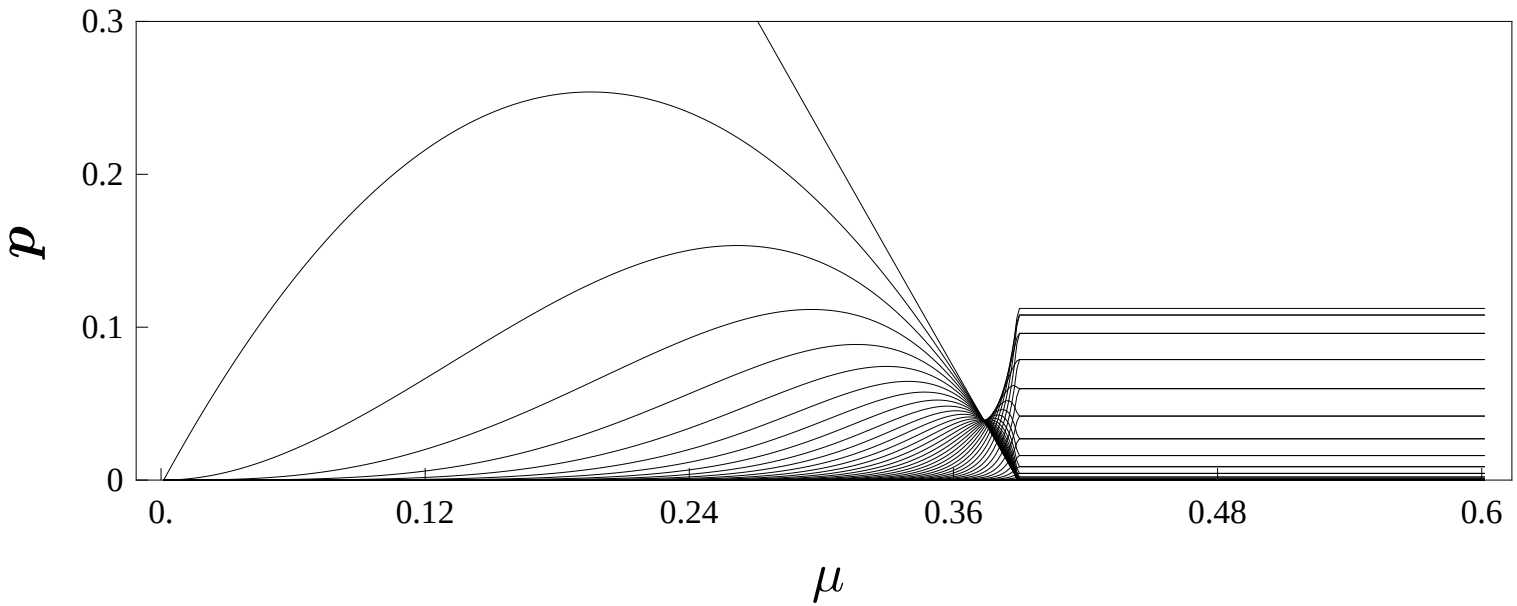


Fig. 10. Distribution of mutant classes as a function of the single-digit accuracy q for $\nu = 50$. Note the sharpness of the transition from direct to stochastic replication around q_{min}. This is seen best on the logarithmic plot. In the domain of stochastic replication individual concentrations become exceedingly small: $\bar{x}_i = 8.9 \times 10^{-16}$, $i = 0, \ldots, 2^{50} - 1$. For basic definitions and numerical values see fig. 7.

Known results: The error threshold

Consider the single peaked fitness landscape

$$\boldsymbol{M} = \mathrm{diag}(m_0, 0, \ldots, 0), \quad m_0 > 0.$$



Known results: Statistical Physics

- Leuthäusser, I. (1986). An exact correspondence between Eigen's evolution model and a two-dimensional Ising system. The Journal of Chemical Physics, 84(3), 1884-1885.
- Tarazona, P. (1992). Error thresholds for molecular quasispecies as phase transitions: From simple landscapes to spin-glass models. Physical Review A, 45(8), 6038.
- Baake, E., Baake, M., & Wagner, H. (1997). Ising quantum chain is equivalent to a model of biological evolution. Physical Review Letters, 78(3), 559-562.
- Galluccio, S. (1997). Exact solution of the quasispecies model in a sharply peaked fitness landscape. Physical Review E, 56(4), 4526.
- Baake, E., & Wagner, H. (2001). Mutation-selection models solved exactly with methods of statistical mechanics. Genetical Research, 78(01), 93-117.
- Saakian, D. B., & Hu, C. K. (2006). Exact solution of the Eigen model with general fitness functions and degradation rates. Proceedings of the National Academy of Sciences of the USA, 103(13), 4935-4939.

Known results: Maximum principle

- Hermisson, J., Redner, O., Wagner, H., & Baake, E. (2002). Mutation-selection balance: Ancestry, load, and maximum principle. Theoretical Population Biology, 62(1), 9-46.
- Baake, E., & Georgii, H. O. (2007). Mutation, selection, and ancestry in branching models: a variational approach. Journal of Mathematical Biology, 54(2), 257-303.

Assume that $m_i = Nr_i = Nr(x_i)$, $x_i = \frac{i}{N} \in [0,1]$ and define $g(x) = \mu\big(1 - 2\sqrt{x(1-x)}\big)$. Then the mean fitness $\overline{m}(\mu) = N\overline{r}$ is given by

$$\overline{r} \approx \overline{r}_\infty = \sup_{x\in[0,1]} \big(r(x) - g(x)\big).$$

$r(x)$ may have only finite number of discontinuities and be either left or right continuous at every point.

Main idea:

We consider the eigenvalue problem

$$(\boldsymbol{M} + \mu \boldsymbol{Q})\boldsymbol{p} = \overline{m}\,\boldsymbol{p}, \quad \overline{m} = \boldsymbol{m} \cdot \boldsymbol{p},$$

where $\boldsymbol{p} = \boldsymbol{p}(\mu)$, $\overline{m} = \overline{m}(\mu)$ with a fixed fitness landscape $\boldsymbol{m}$.

We claim that this problem simplifies in the coordinates of the basis composed of the eigenvectors of the matrix $\boldsymbol{Q} = \boldsymbol{Q}_N$.

Proposition: For the matrix $Q = Q_N$:

1. The eigenvalues of $\boldsymbol{Q}_N$ are simple (all have algebraic multiplicities one) and given by
$$q_k = -2k, \quad k = 0, \ldots, N.$$

2. Let $\boldsymbol{v}_k^\top = (c_{0k}, \ldots, c_{Nk})$ be the right eigenvector of $\boldsymbol{Q}_N$ corresponding to q_k and normalized such that $c_{0k} = 1$, $\boldsymbol{C} = \boldsymbol{C}_N = (c_{ik})_{(N+1)\times(N+1)}$ be the matrix composed of $\boldsymbol{v}_k$ ($\boldsymbol{v}_k$ is the k-th column of $\boldsymbol{C}_N$). Then the generating function for the elements of the k-th column has the form
$$P_k(t) = \sum_{i=0}^{N} c_{ik}\, t^i = (1-t)^k (1+t)^{N-k}, \quad k = 0, \ldots, N.$$

3. $\boldsymbol{C}^2 = 2^N \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix, or, equivalently,
$$\boldsymbol{C}^{-1} = 2^{-N} \boldsymbol{C}.$$

4. 1-norm of $\boldsymbol{C}$ is
$$\|\boldsymbol{C}\|_1 = \max_{0 \le k \le N} \sum_{i=0}^{N} |c_{ik}| = 2^N.$$

Behavior for $\mu \to \infty$:

Let $\boldsymbol{x} = \boldsymbol{C}^{-1}\boldsymbol{p}$ and $\hat{\boldsymbol{x}} = 2^{-N}(1, 0, \ldots, 0)$. Then

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_1 \leq \frac{1}{\mu}\|\boldsymbol{M}\|_1,$$

and

$$\|\boldsymbol{x}'(\mu)\|_1 \leq \frac{2N}{2\mu - (2^{N+1}+1)\|\boldsymbol{M}\|_1}\,.$$

Therefore,

$$\boldsymbol{p}(\mu) \to \hat{\boldsymbol{p}} = 2^{-N}\left(\binom{N}{0}, \ldots, \binom{N}{N}\right).$$

Parametric solution:

Consider fitness landscape such that $m_j > 0$ for some j, and all the rest $m_i = 0$ for $i \neq j$. Then

$$x_k(s) = 2^{-N} \frac{c_{kj}}{1+ks},$$

$$p_i(s) = 2^{-N} \sum_{k=0}^{N} \frac{c_{ik}c_{kj}}{1+ks},$$

$$\overline{m}(s) = m_j p_j(s),$$

$$\mu = \frac{s}{2}\overline{m}(s).$$

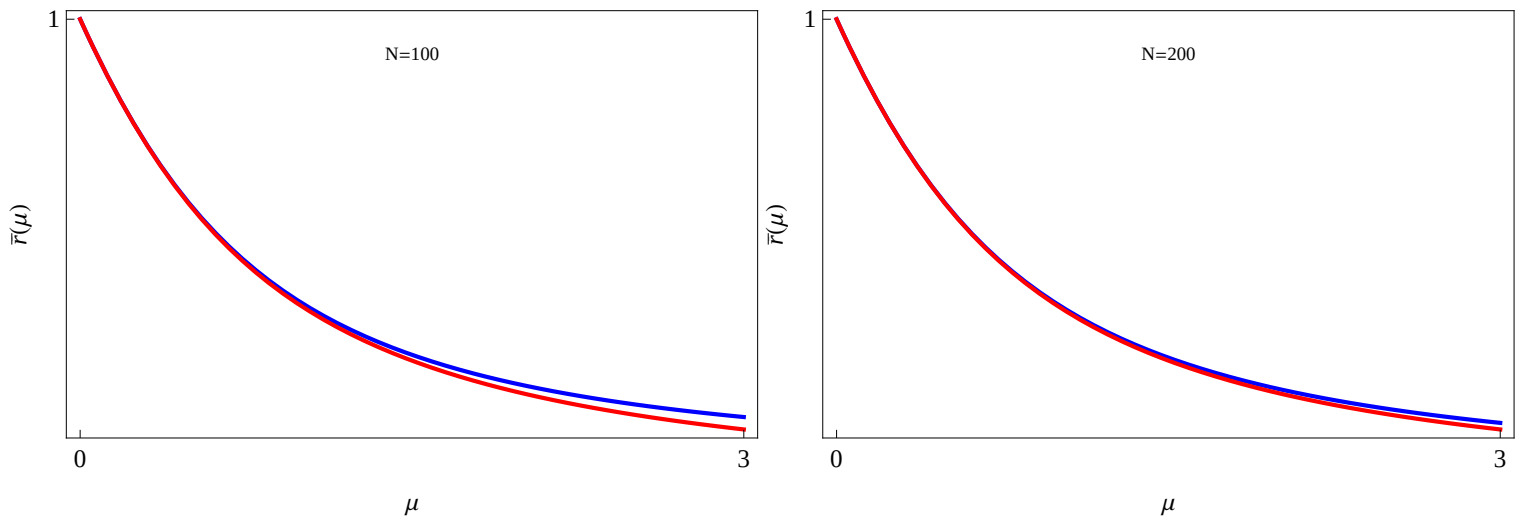
Example 1:

Let $N = 2A$ and

$$\boldsymbol{m} = (0, \ldots, 0, N, 0, \ldots, 0),$$

where N is exactly on A-th position. Consider a scaled fitness landscape $N\boldsymbol{r} = \boldsymbol{m}$ and $\overline{r}(\mu) = \overline{m}(\mu)/N$. Then, using the properties of the mutation matrix $\boldsymbol{Q}$ and the parametric formulas from the previous slide, it can be proved that

$$\overline{r} \approx \overline{r}_\infty = \lim_{N\to\infty} \overline{r}(\mu) = \sqrt{\mu^2 + 1} - \mu.$$



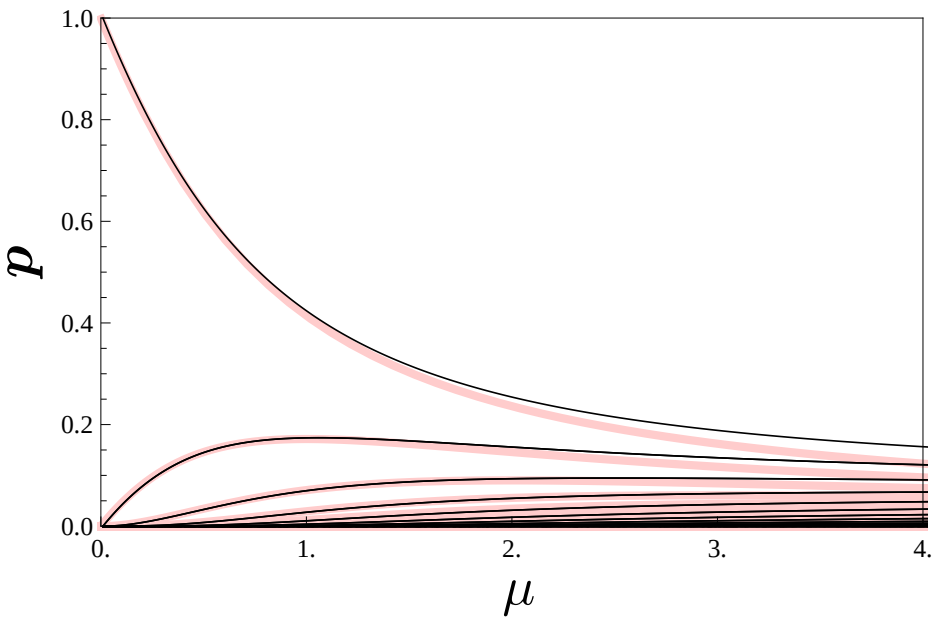
Example 1:

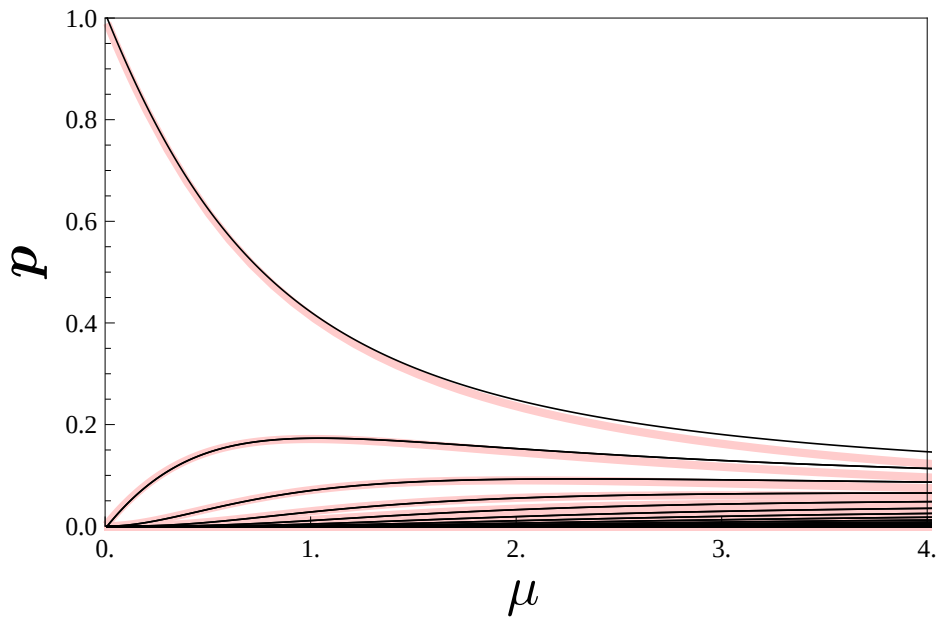
Let $N = 2A$ and

$$\boldsymbol{m} = (0, \ldots, 0, N, 0, \ldots, 0),$$

where N is exactly on A-th position. Consider a scaled fitness landscape $N\boldsymbol{r} = \boldsymbol{m}$ and $\overline{r}(\mu) = \overline{m}(\mu)/N$. Denoting $\overline{r}_\infty = \sqrt{\mu^2 + 1} - \mu$ we can prove that

$$\lim_{N\to\infty} p_{A\pm k}(\mu) = \overline{r}_\infty \left(\frac{1 - \overline{r}_\infty}{1 + \overline{r}_\infty}\right)^k, \quad k = 0, \ldots, A.$$





Example 2: Single peaked landscape

Let

$$\boldsymbol{m} = (N, 0, \ldots, 0).$$

Consider a scaled fitness landscape $N\boldsymbol{r} = \boldsymbol{m}$ and $\overline{r}(\mu) = \overline{m}(\mu)/N$. Then, using the properties of the mutation matrix $\boldsymbol{Q}$ and the parametric formulas we can prove that

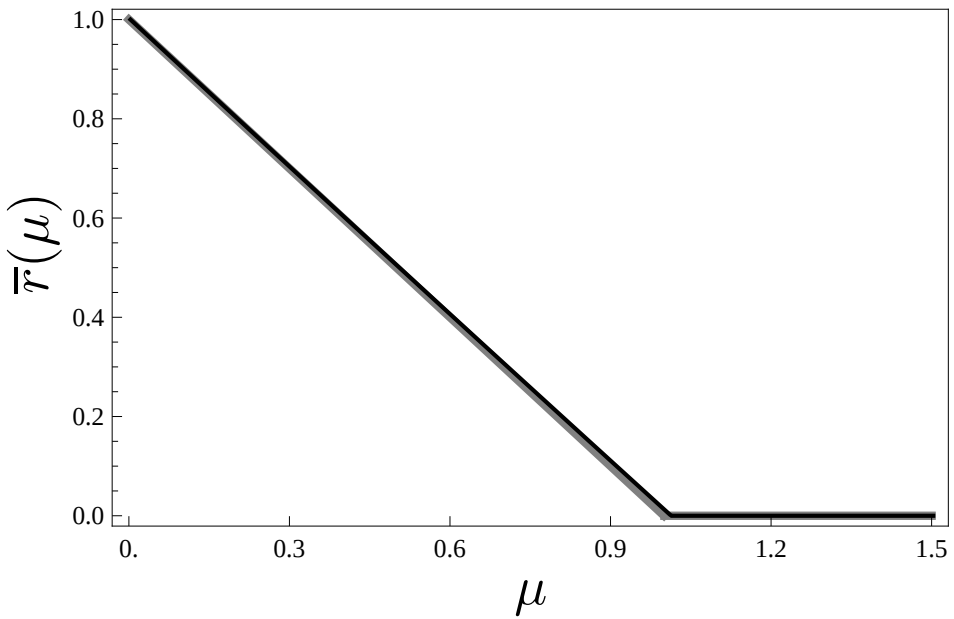
$$\overline{r}_\infty = \lim_{N\to\infty} p_0 = 1 - \mu,$$
$$\lim_{N\to\infty} p_i = (1-\mu)\mu^i, \quad i \geq 1.$$

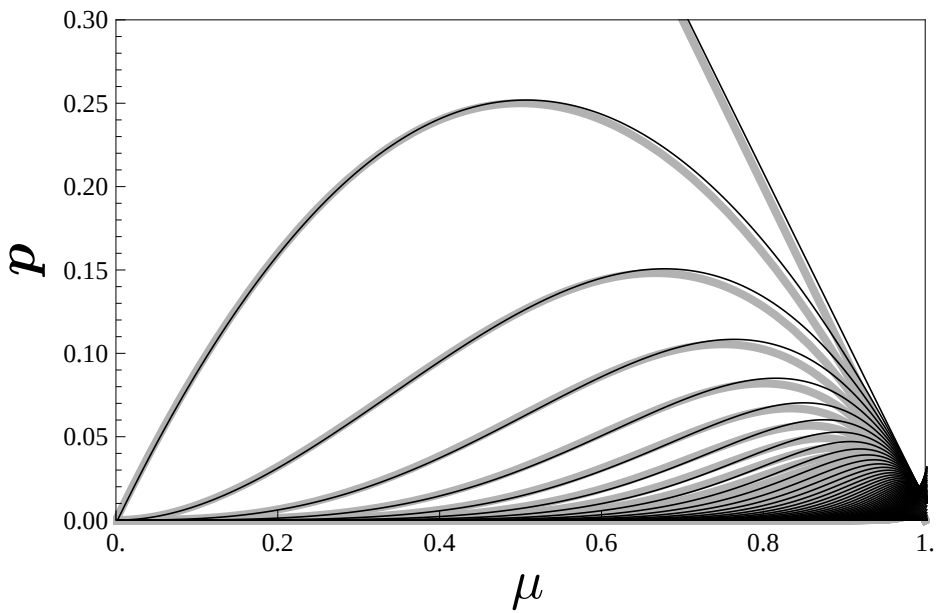
if $\mu < 1$ and $p_j = 0$ for all j if $\mu \geq 1$.

Example 2: Single peaked landscape

Comparison with numerical computations for $N = 100$.

$$\overline{r}_\infty = \lim_{N\to\infty} p_0 = 1 - \mu, \qquad \lim_{N\to\infty} p_i = (1-\mu)\mu^i, \quad i \geq 1$$





Epsilon stabilization:

Definition: Dominant eigenvalue $\overline{m}(\mu)$ admits epsilon stabilization if for a small enough $\varepsilon > 0$ there exists constant $\overline{m}^*_\varepsilon$ and μ^*_ε such that for all $\mu > \mu^*_\varepsilon$ we have

$$|\overline{m}(\mu) - \overline{m}^*_\epsilon| < \varepsilon, \quad |\overline{m}'(\mu)| < \varepsilon.$$

We know that epsilon stabilization is a sure event, due to the previous. The question is how to determine μ^*_ε given $\varepsilon > 0$.

Epsilon stabilization:

Let $\boldsymbol{m} = (m_0, m_1, \ldots, m_N)$ such that $m_0 > m_1 \geq m_2 \geq \ldots \geq m_N$. Then we have

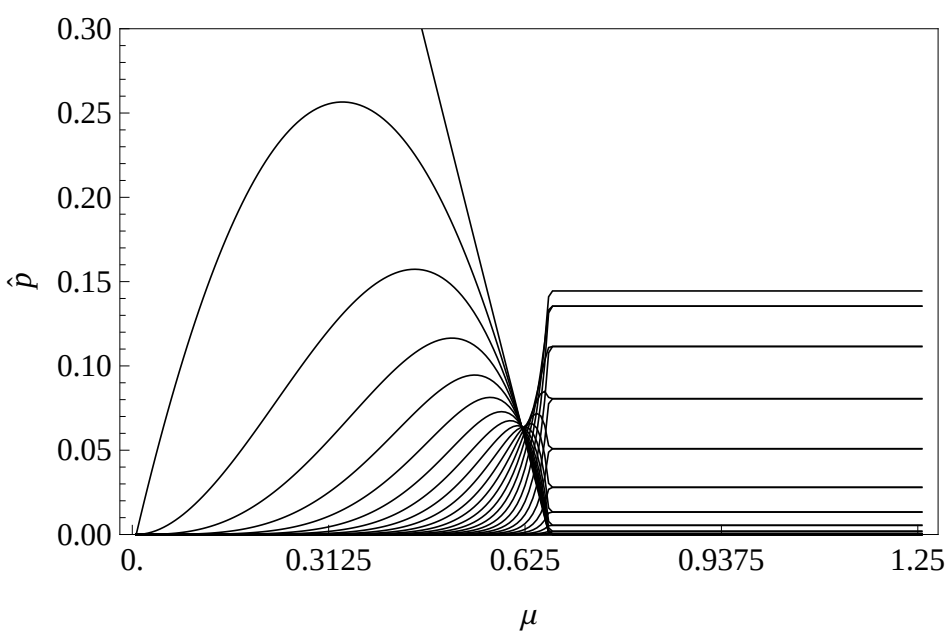
$$\mu^* = \frac{m_0 - m_1}{N}\,, \qquad \text{(Classical Result)}$$

$$\mu^*_\varepsilon = (m_0 - m_1)\left(1 - \sqrt{1 - \frac{2(m_0 - 1)}{(m_0 - m_1)N}}\right), \qquad \text{(Our result 1)}$$

$$\text{if } \delta = 2^{-N}\sum_{k=0}^{N}(m_k - 1)\binom{N}{k} < \varepsilon,$$

$$\overline{m}(\mu^*) = m_{\min} + 2\mu^*. \qquad \text{(Our result 2)}$$

Example:



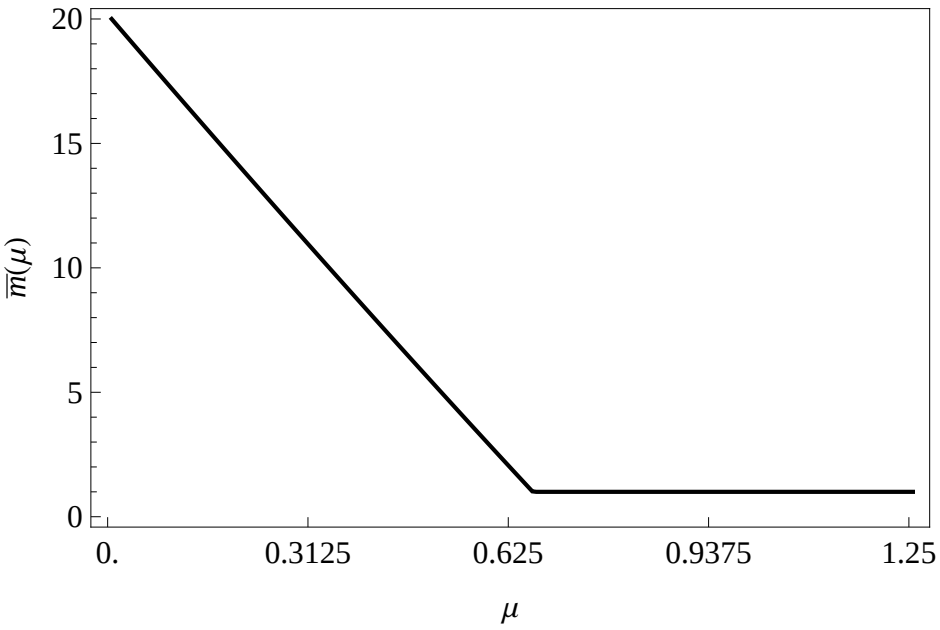


Figure : Error threshold in the quasispecies model with the single picked fitness landscape

Estimates from the previous slide:

$$\mu_1^* = 0.633,$$
$$\mu_2^* = 0.644,$$
$$\mu_3^* = 0.613$$

Example:

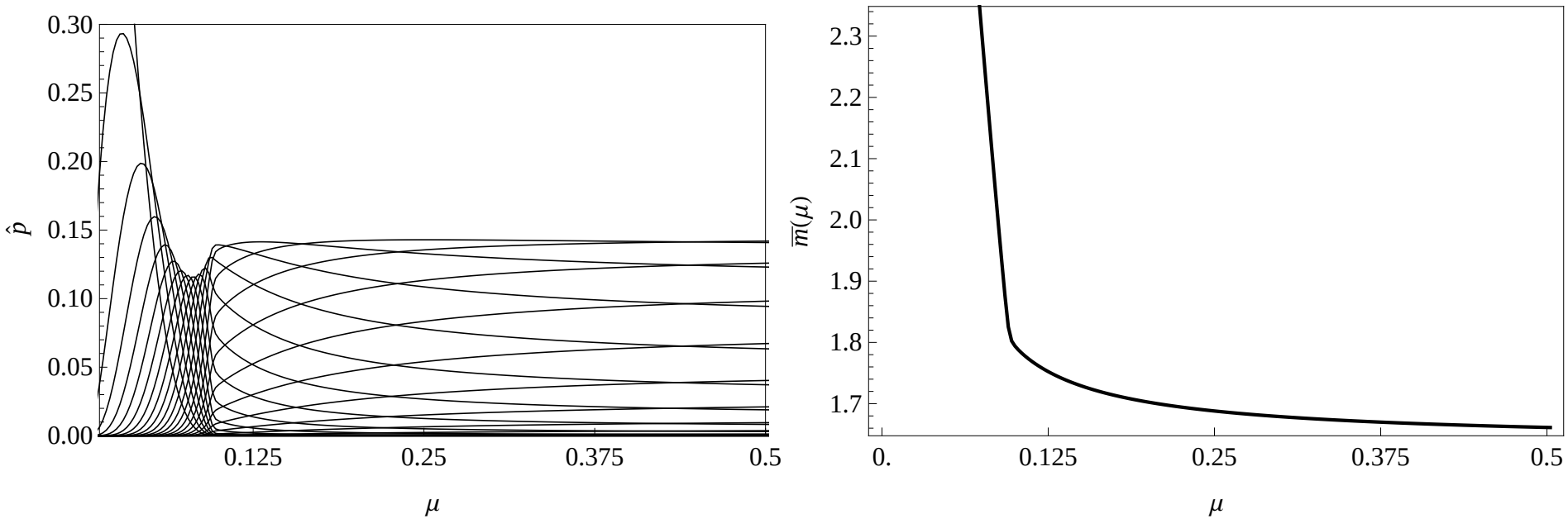


Figure : Error threshold in the quasispecies model with $m_k = k^\alpha \log(1-s)$

Estimates from the previous slide:

$$\mu_1^* = 0.09,$$
$$\mu_2^* = 0.19,$$
$$\mu_3^* = 0.33$$

Acknowledgements:

Startup grant from Department of Mathematics, NDSU

ND EPSCoR and NSF grant # EPS-0814442

Thank you for your attention!

e-mail: artem.novozhilov@ndsu.edu

site: https://www.ndsu.edu/pubweb/ novozhil/

References:

- Bratus, A. S., Novozhilov, A. S., & Semenov, Y. S. (2013). Linear algebra of the permutation invariant Crow–Kimura model of prebiotic evolution. arXiv:1306.0111.
- Semenov, Y. S., Bratus, A. S., & Novozhilov, A. S. (2014). On the behavior of the leading eigenvalue of the Eigen evolutionary matrices, in preparation.