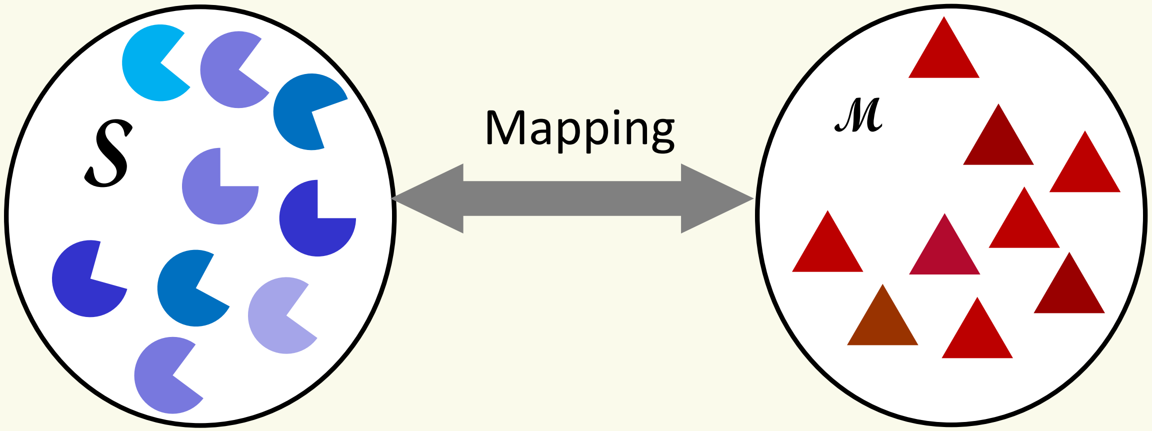
Lecture IV – A

Shannon’s theory of noisy channels and molecular codes

Noisy molecular codes: Rate-Distortion theory

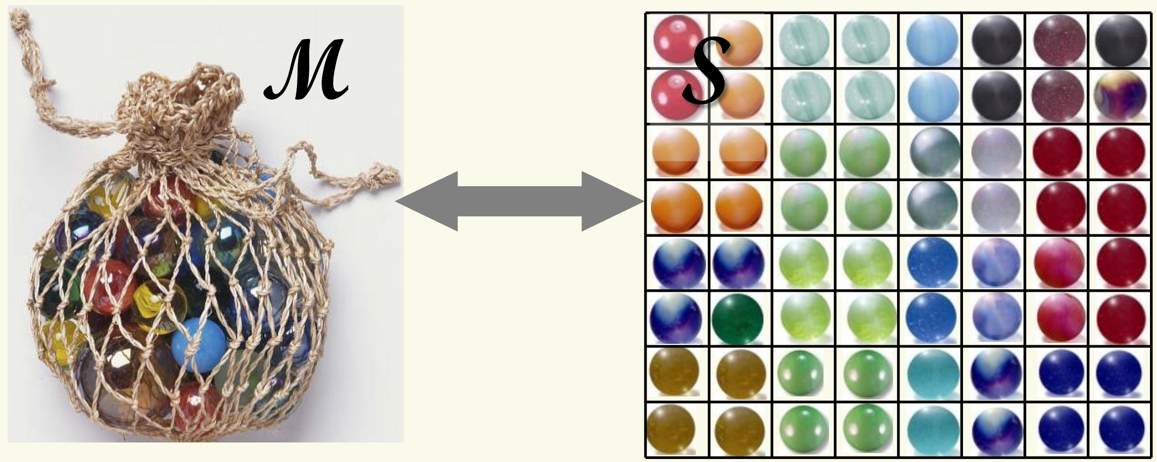


- Channel/Code = **mapping** between two molecular spaces.
- Two functionals determine the "*fitness*" of the code:

  Fitness = **Rate**(map) + **Distortion**(map).

- Mapping becomes non-random at a *coding transition*.
- *Topological* aspects.

Geometry of molecular information channels

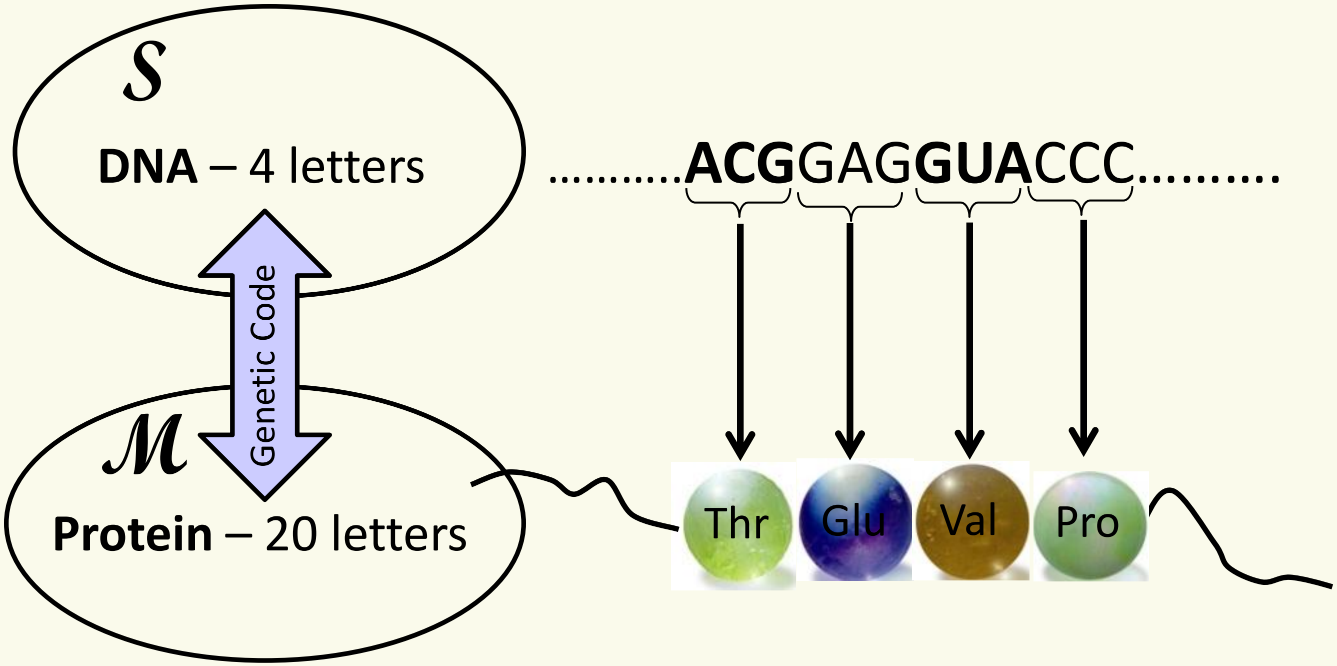


<u>“Marble packing”</u>

(A) Max colors.

(B) Same\similar color of neighbors.

The genetic code is main info channel of life



- **Genetic code**: translates 3-letter words in 4-letter DNA language ($4^3 = 64$ codons) to protein language of 20 amino acids.

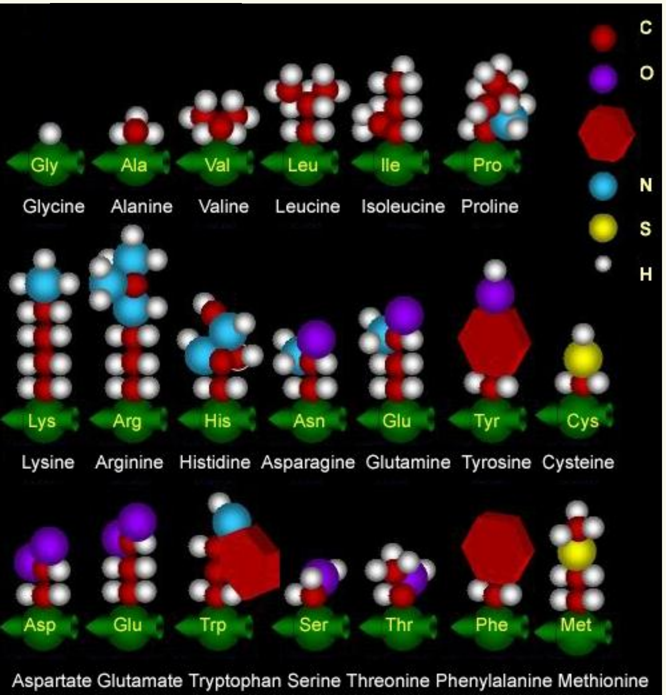
- Proteins are amino acid polymers.

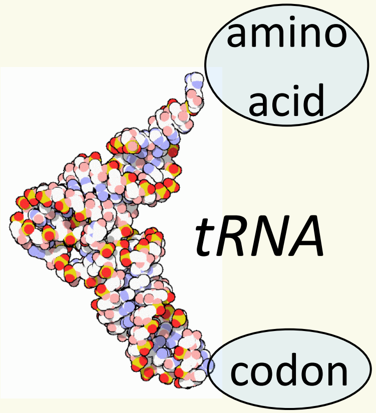
- **Diversity** of amino-acids is essential to protein functionality.

The genetic code maps codons to amino-acids

- Molecular code = map relating two sets of molecules
  (spaces, “languages”) **via molecular recognition**.
- Spaces defined by similarity of molecules (size, polarity etc.)

20 amino-acids





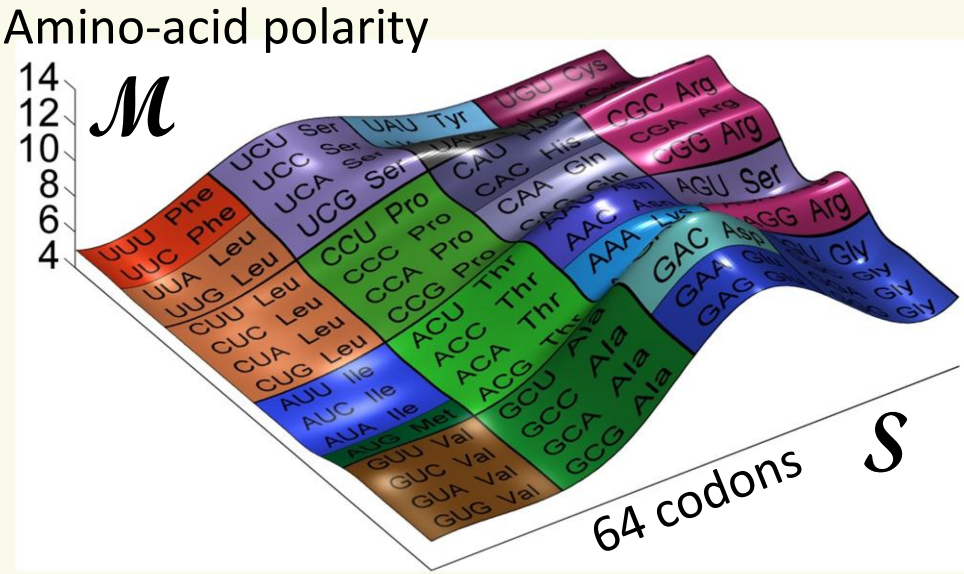
Genetic Code

64 codons

UUU	UUA	UCU	UCA	UAU	UAA	UGU	UGA
UUC	UUG	UCC	UCG	UAC	UAG	UGC	UGG
CUU	CUA	CCU	CCA	CAU	CAA	CGU	CGA
CUC	CUG	CCC	CCG	CAC	CAG	CGC	CGG
AUU	AUA	ACU	ACA	AAU	AAA	AGU	AGA
AUC	AUG	ACC	ACG	AAC	AAG	AGC	AGG
GUU	GUA	GCU	GCA	GAU	GAA	GGU	GGA
GUC	GUG	GCC	GCG	GAC	GAG	GGC	GGG

The genetic code is a smooth mapping

UUU Phe	UCU Ser	UAU Tyr	UGU Cys
UUC Phe	UCC Ser	UAC Tyr	UGC Cys
UUA Leu	UCA Ser	UAA TER	UGA TER
UUG Leu	UCG Ser	UAG TER	UGG Trp
CUU Leu	CCU Pro	CAU His	CGU Arg
CUC Leu	CCC Pro	CAC His	CGC Arg
CUA Leu	CCA Pro	CAA Gln	CGA Arg
CUG Leu	CCG Pro	CAG Gln	CGG Arg
AUU Ile	ACU Thr	AAU Asn	AGU Ser
AUC Ile	ACC Thr	AAC Asn	AGC Ser
AUA Ile	ACA Thr	AAA Lys	AGA Arg
AUG Met	ACG Thr	AAG Lys	AGG Arg
GUU Val	GCU Ala	GAU Asp	GGU Gly
GUC Val	GCC Ala	GAC Asp	GGC Gly
GUA Val	GCA Ala	GAA Glu	GGA Gly
GUG Val	GCG Ala	GAG Glu	GGG Gly



- Degenerate (20 out of 64).
- Compactness of amino-acid regions.
- **Smooth** (similar “color” of neighbors).

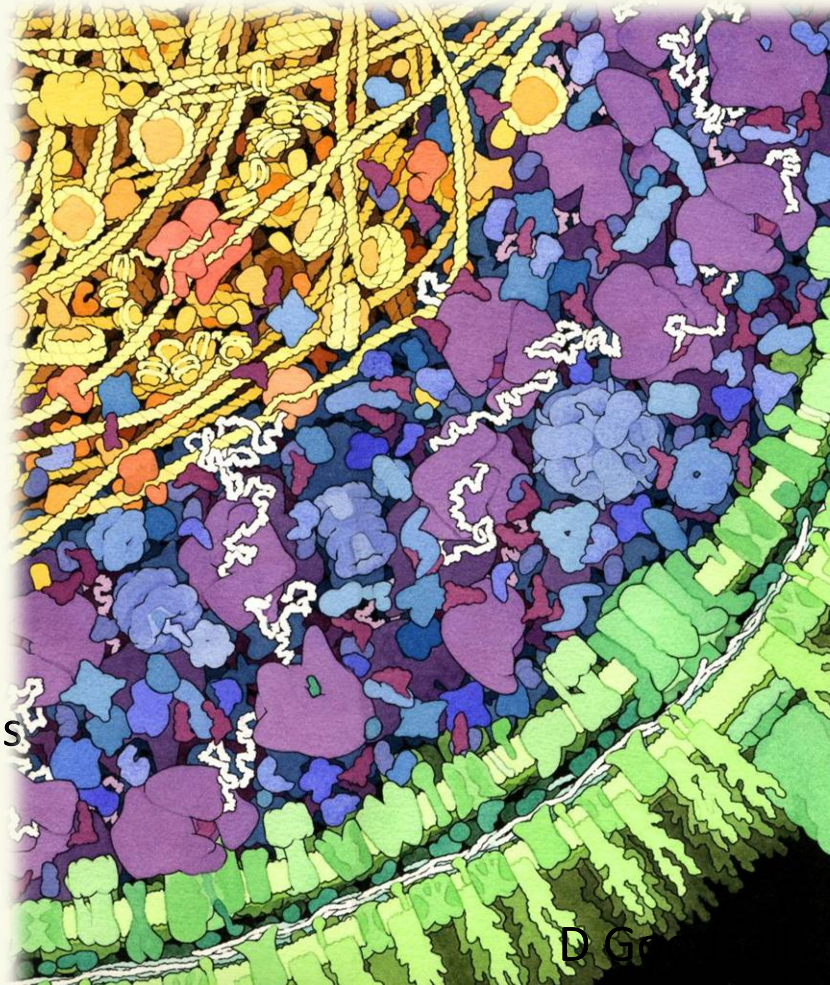
Generic properties of molecular codes?

Challenges of molecular codes: rate and distortion

Distortion

- Noise, crowded milieu.
- Competing lookalikes.
- Weak recognition interactions ~ $k_B T$.
- Need diverse meanings.

"Synthesis of reliable organisms from unreliable components"
(von Neumann, *Automata Studies* 1956)

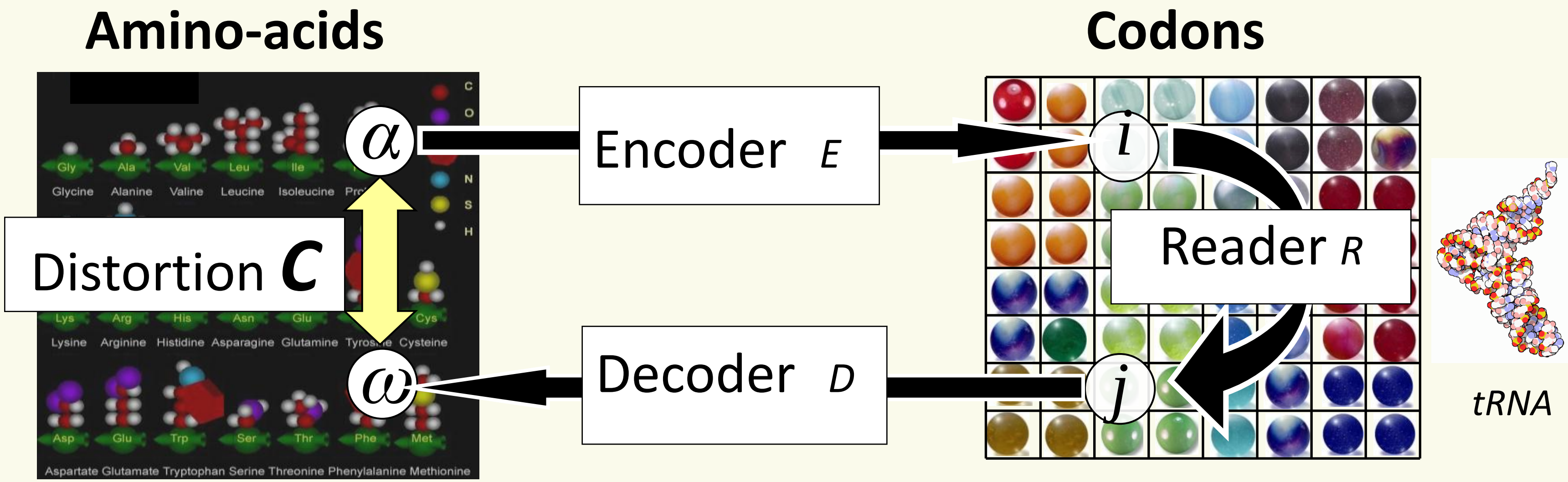


Rate

- How to construct the low-rate molecular codes at minimal cost of resources?

Rate-distortion theory (Shannon 1956)

Fitter codes have minimal distortion

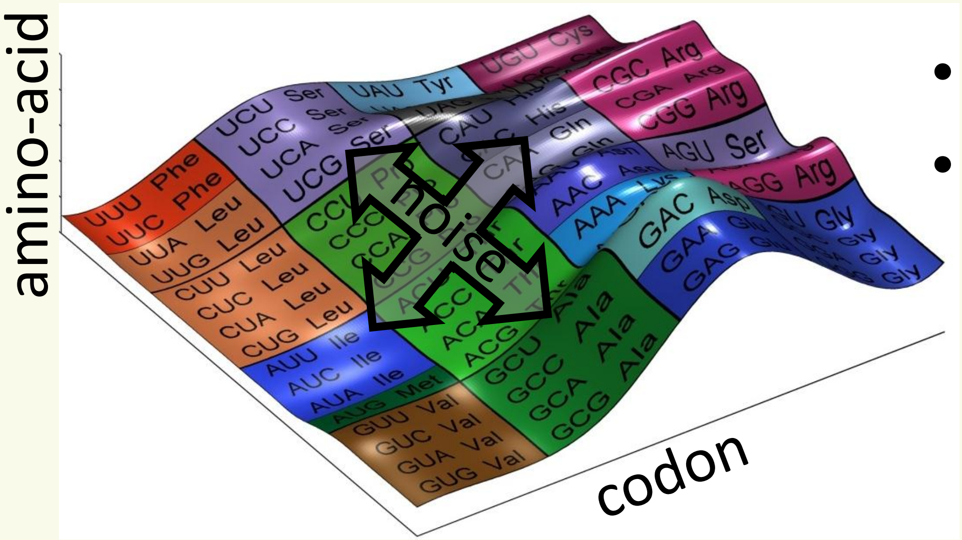


$$Q = \langle C_{\alpha\omega} \rangle = \sum_{paths} P_{path} C_{\alpha\omega} = \mathrm{Tr}\left(E \cdot R \cdot D \cdot C\right)$$

- Distortion of noisy channel, Q = average distortion of AA.
- R defines topology of codon space.
- C defines topology of amino-acid space.

(TT, J Theo Bio 2007, PRL 2008, PNAS 2008)

Smooth codes minimize distortion

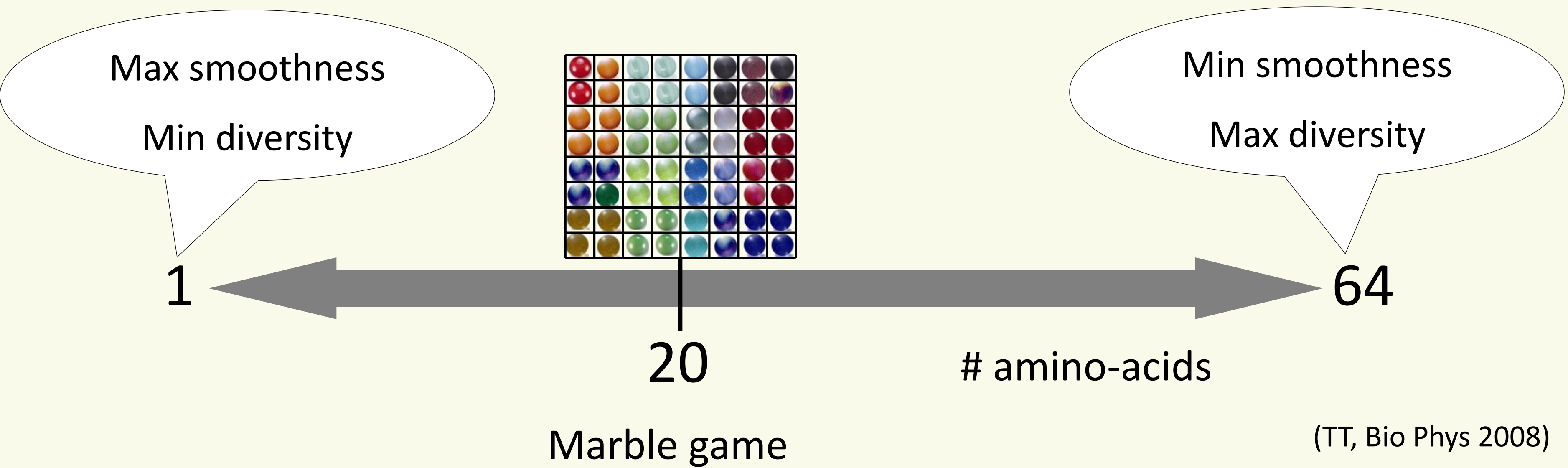


- Noise confuses close codons.
- Smooth code:

  close codons = close amino-acids.

  → minimal distortion.

- Optimal code must balance contradicting needs for **smoothness** and **diversity**.

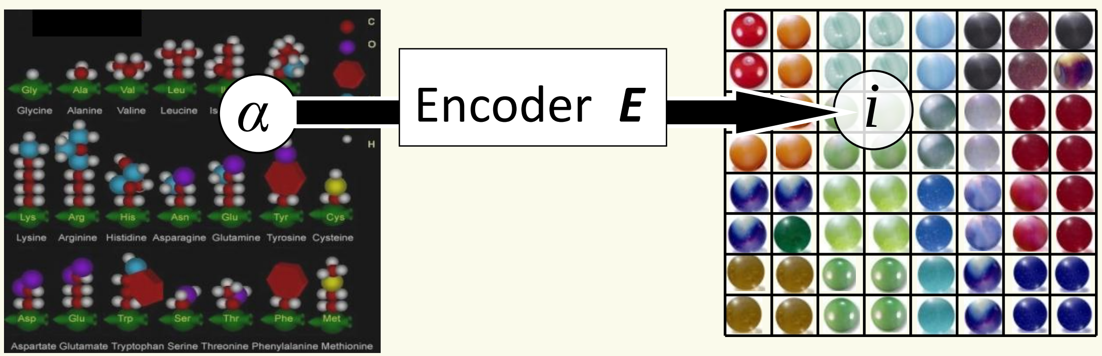


(TT, Bio Phys 2008)

Channel rate is code's cost

- Diverse codes require high specificity = high binding energies ε.
- Cost ~ average binding energy $< \varepsilon >$.
- Binding prob. ~ Boltzmann: $E \sim e^{\varepsilon/T}$.

$$I \sim \sum_{\alpha,i} E_{\alpha i} \ln E_{\alpha i} \sim \langle \varepsilon_{\alpha i} \rangle_E$$



- Cost I = Channel Rate (bits/message)

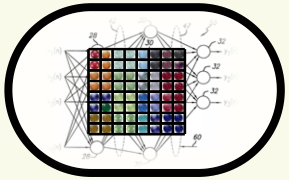
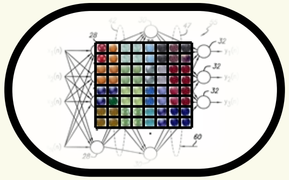
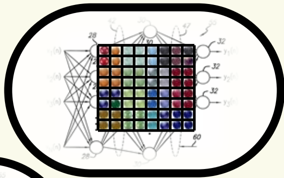
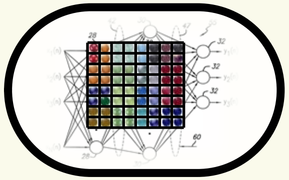
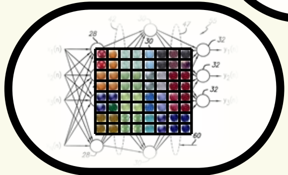
Code's fitness combines rate and distortion of map

Fitness = Gain x *Distortion + Rate*

$$H = \beta Q + I$$

- ***Gain*** β increases with organism complexity and environment richness.
- Fitness ***H*** is "free energy" with inverse "temperature" β.
- Evolution varies the gain β.

- Population of self-replicators evolving according to code fitness *H*: mutation, selection, random drift.

Code emerges at a critical coding transition

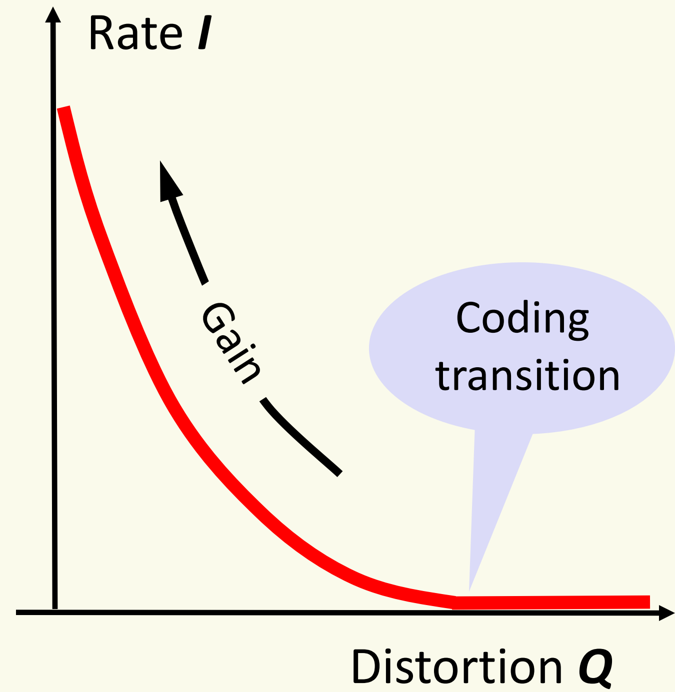
- Low gain β : Cost too high

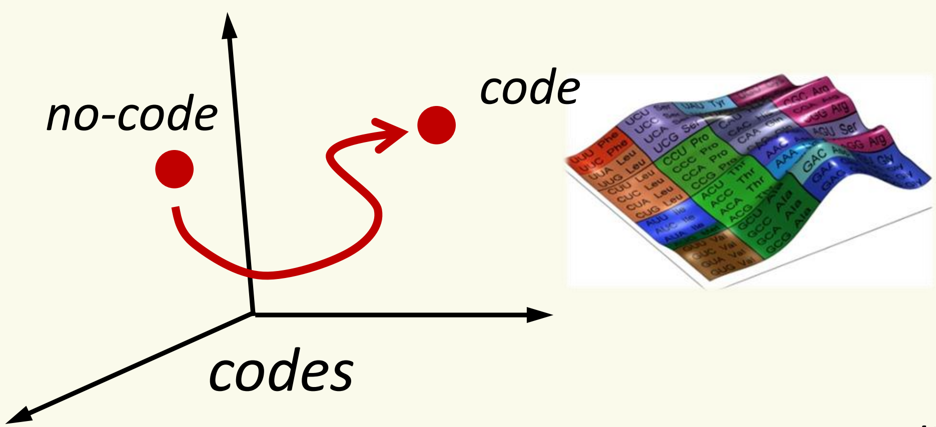
  → no specificity → **no code**.

- **Code emerges** when β increases:

  channel starts to convey information ($I \neq 0$).

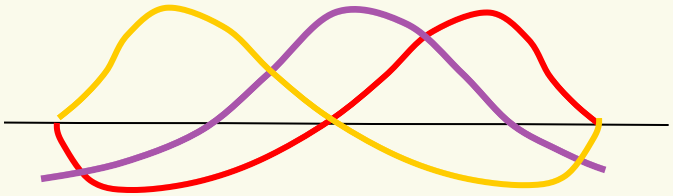
- Continuous phase transition.
- Emergent code is smooth, low mode of ***R***.





Rate-distortion theory (Shannon 1956)

Emergent code is a smooth mode of error-Laplacian

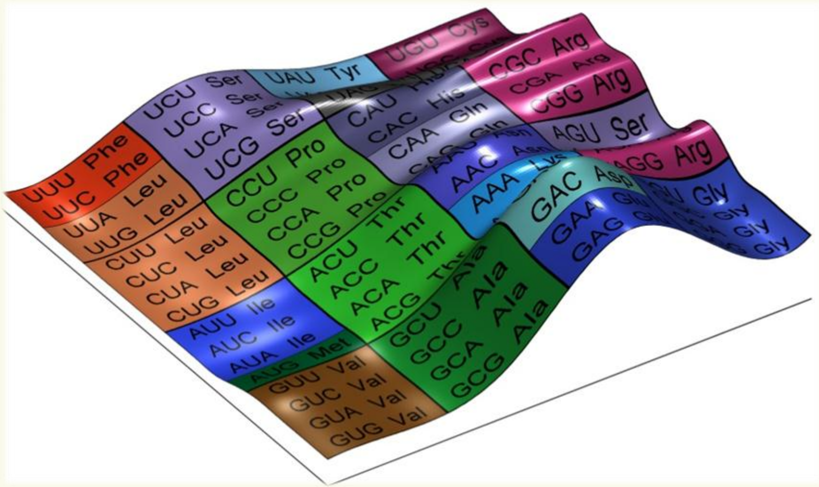


- Lowest excited modes of graph-Laplacian *R* .
- Single maximum for lowest excited modes (Courant).
- Every mode corresponds to amino-acid :

 # low modes = # amino-acids.

→ single contiguous domain for each amino-acid.

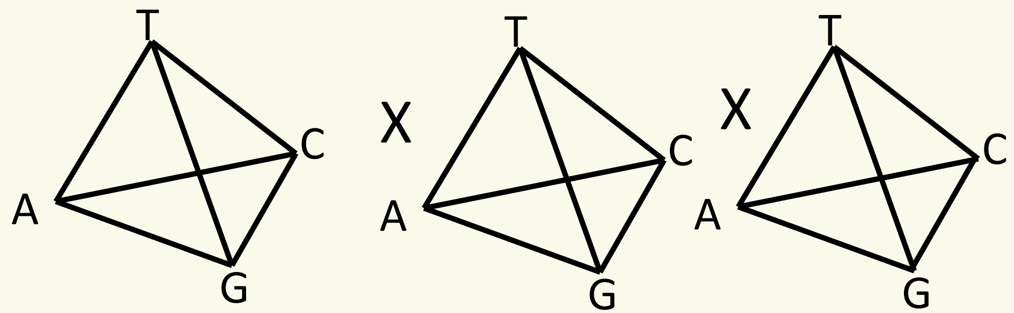
→ **Smoothness**.

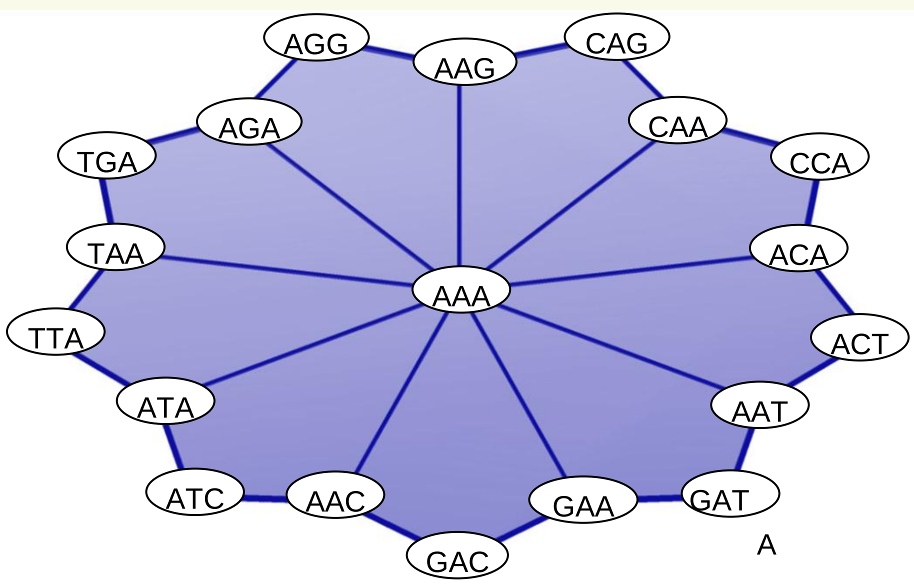


Probable errors define the graph and the topology of the genetic code

- Codon graph = codon vertices + 1-letter difference edges (mutations).

$K_4 \times K_4 \times K_4$





- Non-planar graph (many crossings).
- Genus γ = # holes of embedding manifold.
- Graph is holey : embedded in $\gamma = 41$

(lower limit is $\gamma = 25$)

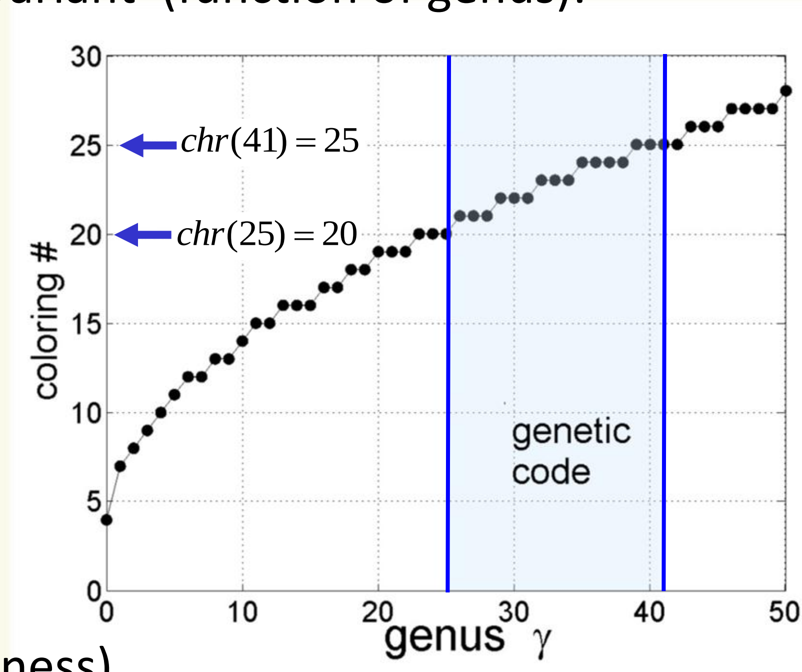
Coloring number limits number of amino-acids

- Q: Minimal # colors suffices to color a map where neighboring countries have different colors?
- A: Coloring number, a topological invariant (function of genus):

$$chr(\gamma) = \left\lfloor \frac{1}{2}\left(7 + \sqrt{1 + 48\gamma}\right) \right\rfloor.$$

(Ringel & Youngs 1968)

$$\max(\#\ \text{amino-acids}) = chr(\gamma)$$

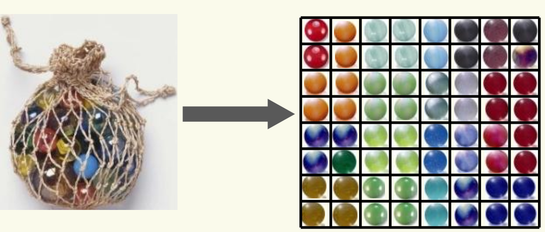


- From Courant 's theorem + "convexity" (tightness).
- Genetic code: γ = 25-41 → coloring number = 20-25 amino-acids

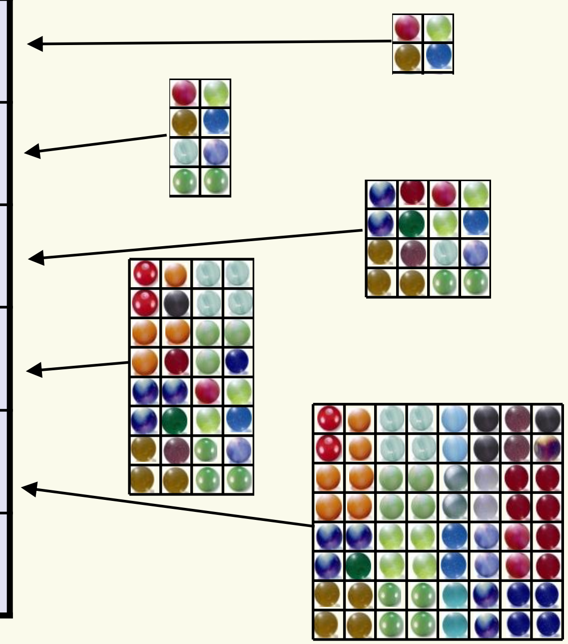
(TT *J Lin Alg* 2008)

The genetic code coevolves with accuracy

- A path for evolution of codes: from early codes with higher codon degeneracy and fewer amino acids to lower degeneracy codes with more amino acids.



1^{st}	2^{nd}	3^{rd}	γ	*chr #*
1	4	1	0	4
2	4	1	1	7
4	4	1	5	11
4	4	2	13	16
4	4	3	25	20
4	4	4	41	25



Lecture IV – B

Growth rate as entropy rate: Kelly's horse race

A New Interpretation of Information Rate

reproduced with permission of AT&T

By J. L. KELLY, JR.

(Manuscript received March 21, 1956)

Horse race basics

Lucky Star:
Odds: 2:1

White light:
Odds: 3:1

Sea biscuit
Odds: 6:1

Horse race basics (cont.)

- Problem: You dedicate 100$ for gambling, you intend to reinvest the money over and over again, what is the optimal strategy?

- Kelly's idea: try to optimize asymptotic growth rate.

Asymptotic growth

- Let *W*(*n*) be your wealth after *n* bets.
- Let *W*(*0*) be you initial wealth
- Growth rate

$$\Lambda = \frac{1}{n}\log_2 \frac{W(n)}{W(0)}$$

- Asymptotic growth rate

$$\Lambda_\infty = \lim_{n\to\infty} \frac{1}{n}\log_2 \frac{W(n)}{W(0)}$$

Constant rebalancing

- Each race has a random outcome X drawn from the distribution P(X) assumed constant ($\partial_t P=0$).

- The percentage of money placed on the i-th horse of the n-th round is W(n-1)*b(i).

- The amount gained :

$$W(n)=O(x)\ W(n-1)\ b(x)$$

Constant rebalancing

- After N such trials (with rebalancing) :

$$W(n) = W(0) \times O(X_1)b(X_1) \times \ldots O(X_N)b(X_N)$$

- So

$$\log_2 \frac{W(n)}{W(0)} = \log_2 O(X_1)b(X_1) + \ldots + \log_2 O(X_N)b(X_N)$$

- Since X is memoryless and P(X) is constant we obtain for N>>1

$$\log_2 \left(O(X_1)b(X_1) \right) + \ldots + \log_2 \left(O(X_N)b(X_N) \right) =$$

$$N \left[P(X_1) \log_2 \left(O(X_1)b(X_1) \right) + \ldots + P(X_N) \log_2 \left(O(X_N)b(X_N) \right) \right]$$

Conclusion so far

Asymptotic growth rate

$$\Lambda = \frac{1}{n}\log_2\frac{W(n)}{W(0)} =$$

$$\frac{1}{n}\Big[\log_2\big(O(X_1)b(X_1)\big) + \ldots + \log_2\big(O(X_n)b(X_n)\big)\Big] =$$

$$\underset{n\to\infty}{\longrightarrow} \quad \sum_i p_i \log_2(O_i b_i)$$

Optimal strategy if P(X) is known

- Suppose we know the probability of winning for all the horses p_i.
- What is the optimal bet-hedging strategy?

$$\Lambda = \sum_i p_i \log_2 \left(O_i b_i \right) - \lambda \sum_i b_i$$

$$\frac{\partial \Lambda}{\partial b_i} = \frac{p_i}{b_i} - \lambda = 0 \quad \Rightarrow \quad b_i = p_i$$

- This strategy is termed proportional betting.

Example: Two horses

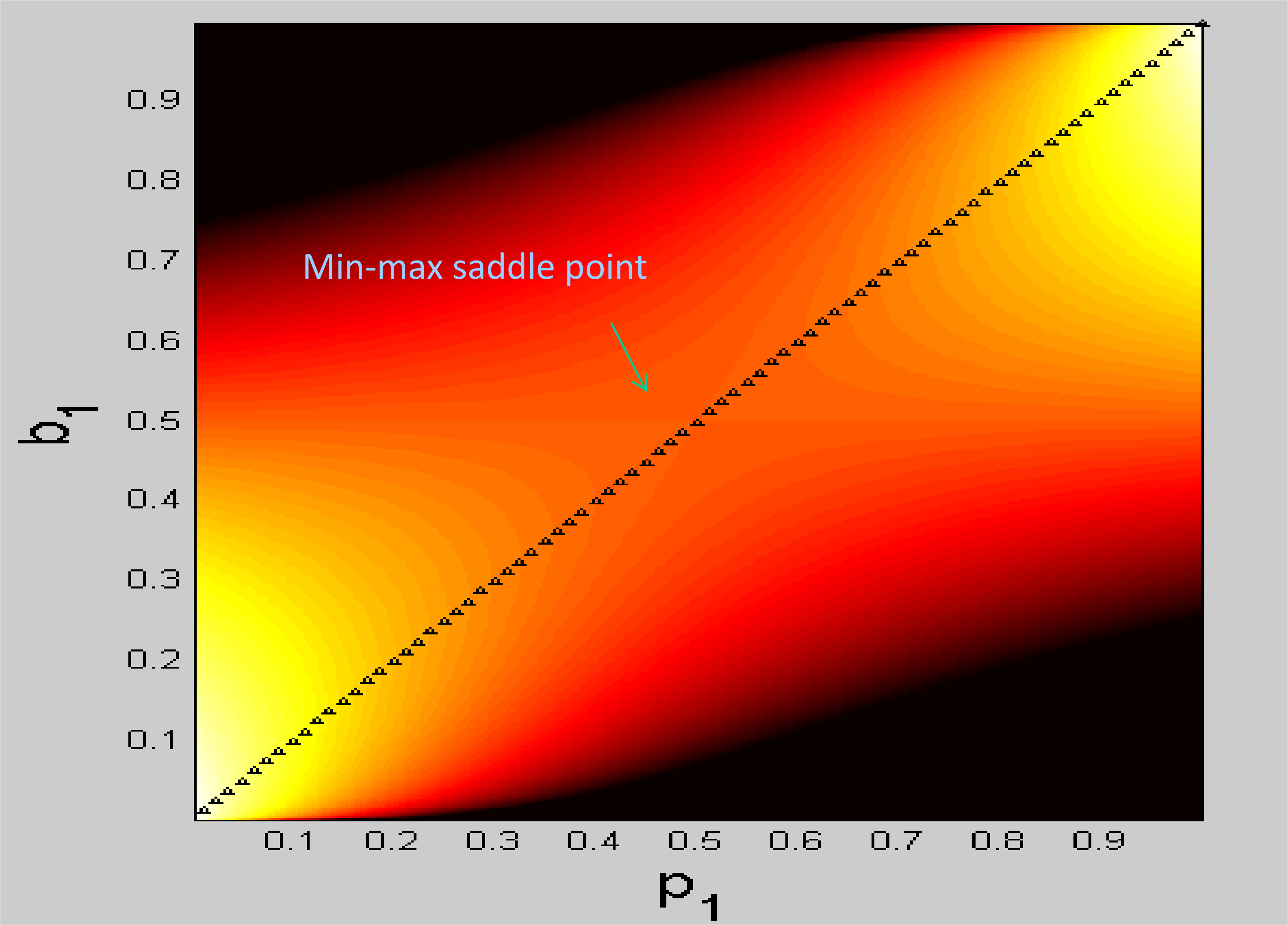
- Odds: 2:1 (double or nothing), i.e. $O_1 = O_2 = 2$.

- Let p_1 be the probability the 1st horse will win and

b_1 the portion of the wealth that placed on this 1st horse.

- Let's plot the asymptotic growth rate:

Example : a race with 2 horses double or nothing



The saddle point is A zero-sum game against “nature”

- The asymptotic growth rate $\Lambda = \Lambda(\mathbf{b}, \mathbf{p})$
- The game is: I choose ***b*** / nature choose ***p***
- What is the minimal growth $\Lambda = \Lambda(\mathbf{b}, \mathbf{p})$ I can assure if nature is “evil”?
- Answer: min-max solution $$\mathbf{b}_{mnmx} = \mathbf{p}_{mnmx} = \frac{O_i^{-1}}{\sum_j O_j^{-1}}$$
- In the example shown $\mathbf{b}_{mnmx} = \mathbf{p}_{mnmx} = \frac{1}{2}$

Growth rate in horse race

$$\Lambda(b,p) = D(p \| p_{\mathrm{mnmx}}) - D(b \| p) + v$$

- D(p||q) – relative (KL) entropy
- 1st term - pessimists surprise (free lunch).
- 2nd term - ''distance'' from optimum (note the sign).
- 3rd term - game value.

Side information

- Race at LA, bookie in NY and I have a friend in the telegraph company…
- Perfect side information = exponential growth.
- What about partial information?

Side information (cont.)

- Informer says horse *j* will win.

- The probability for *i*-th to win given the side information that the j-th horse will win is $p_{i|j}$

- The adjusted portfolio is $b_{i|j}$

Optimal betting with side information

$$\Lambda(b,p)=\sum_{i,j} p_j p_{i|j}\log_2(O_i b_{i|j})=$$

$$D(p\,\|\,p_{\mathrm{mnmx}})+I(X;Y)-\sum_{i,j} p_j D(p_{\cdot|j}\,\|\,b_{\cdot|j})+v$$

$I(X;Y)$ is the mutual information between the informer and us

(X – horse, Y – side information).

Kelly's famous result retrieved at optimality:

Optimal gain of capital = Channel Capacity

So why study horse races? Biology

- Only manifestation of channel capacity without an explicit code.
- Cells ~ money,
- Phenotype ~ betting,
- nature's state ~ winning horse,
- Portfolio = phenotype distribution
- side Info. = sensing

- **BUT**: (i) Suboptimal phenotype ≠ immediate ruin.
- (ii) P=P(t) (non-stationary).

Generalized Kelly (Main result)

$$\Lambda(b,p)=\sum_{i,j} p_j p_{i|j}\log_2\left(\sum_k O_{ik} b_{k|j}\right)=$$

Given that the sum of O^{-1} columns is positive, Λ decomposes to a sum of entropies:

$$D(p\,\|\,p_{\text{mnmx}})+I(X;Y)-\sum_{i,j} p_j D(p_{\cdot|j}\,\|\,S^{-1}b_{\cdot|j})+v$$

new game value

Free lunch term

Side info. Channel rate

Penalty term: ''distance'' between actual p(|j) to the p(|j) you happen to be optimal for.

Generalized Kelly (Main result) cont.

- Iff $\mathbf{b}_{opt}(\mathbf{p}) > 0$ then $b_{opt}(\mathbf{p}) = S\,\mathbf{p}$.

- $$S_{ij}^{-1} = \frac{O_{ij}^{-1}}{\sum_j O_{ij}^{-1}}$$ is a stochastic matrix.

- $\mathbf{S}^{-1}\mathbf{b}_{(\cdot|j)}$ is the conditional environment probability that $\mathbf{b}_{(\cdot|j)}$ is optimal for in the adjusted game.

So the penalty term :

= average loss due to the sub-optimal response to the side information.

$$\sum_{i,j} p_j D(p_{\cdot|j} \parallel S^{-1} b_{\cdot|j})$$

Environment is non-stationary

Slow changes: $\Lambda(p(t),b(t))$ is meaningful

Problem:

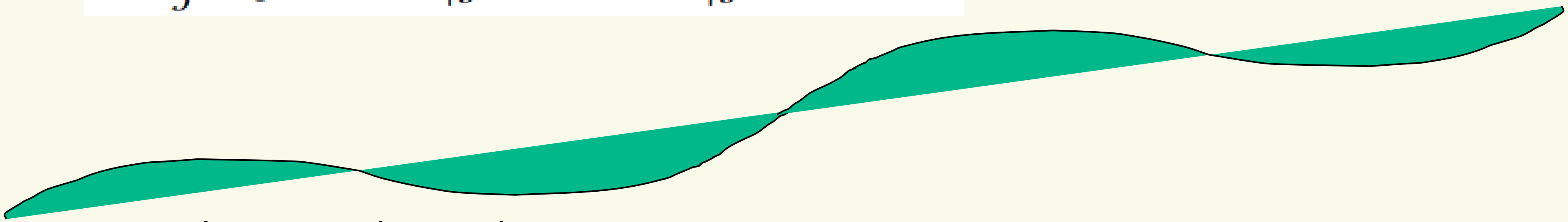
given K phenotypic switchings allowed within [0,T] find optimum switching strategy (when and to what)

When P=P(t)

Our solution:

when ? : equipartition of loss criterion

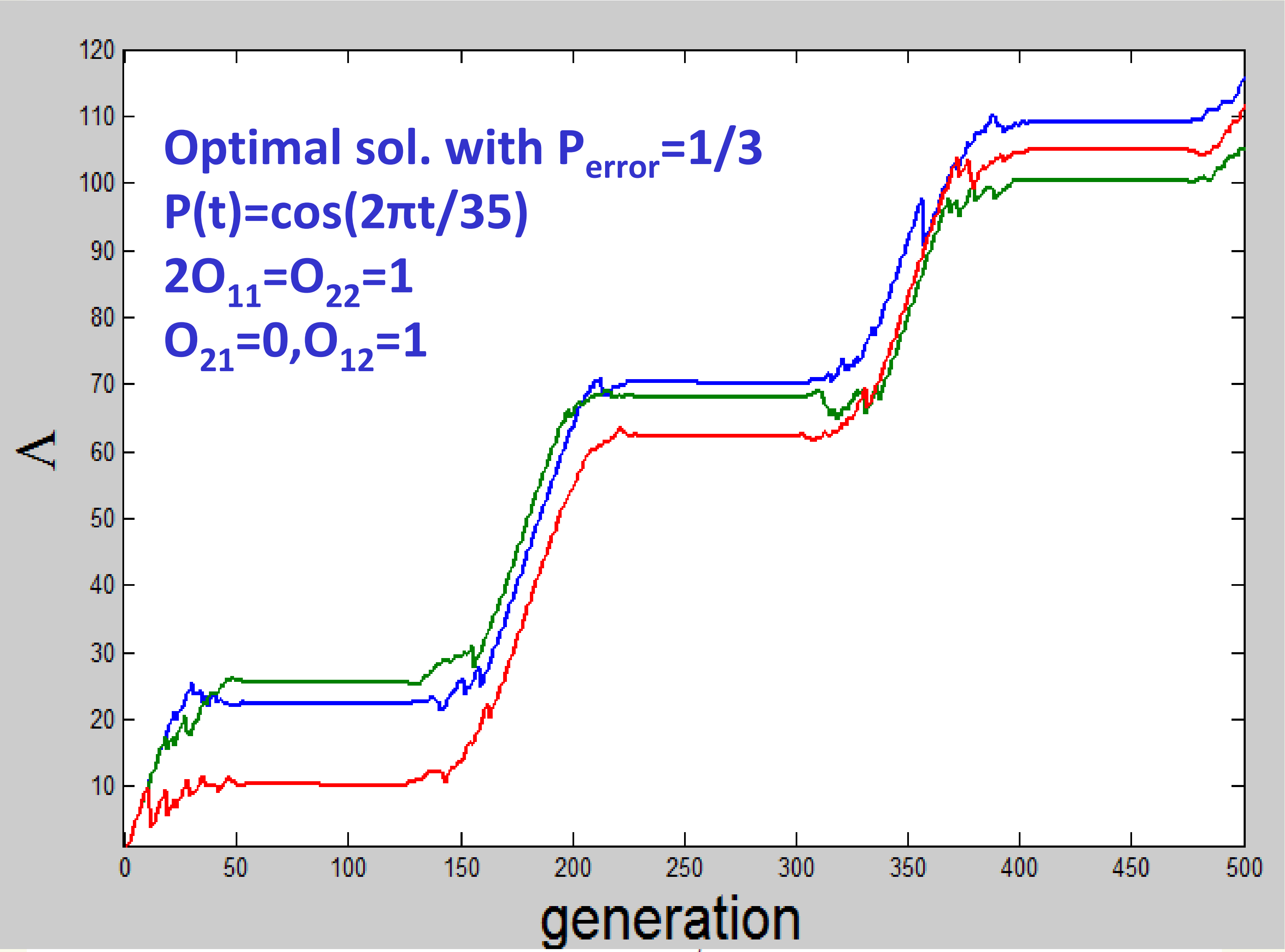
$$\sum_j p_j D(p_{i|j}(t_l)||\bar{p}_{i|j}(t_l)) = \sum_j p_j D(p_{i|j}(t_l)||\bar{p}_{i|j}(t_{l+1}))$$



to what ?: adjusted time average

$$\sum_k S_{ik}^{-1} b_{k|j} = \frac{1}{t_{\nu+1}-t_\nu} \int_{t_\nu}^{t_{\nu+1}} dt p_{i|j}(t)$$

Monte-Carlo (binary symmetric channel)



Conclusion:

If there is a dilemma – an increase of 1 bit in the side information rate can potentially increase the doubling rate by 1 bit/generation.