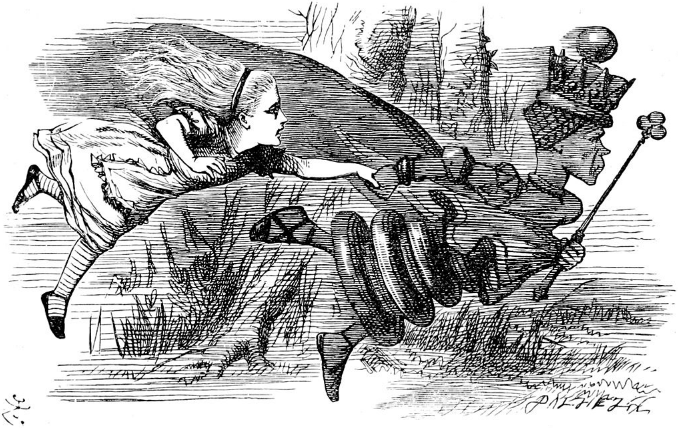
THE FUNDAMENTAL THEOREM OF NATURAL SELECTION



John Baez

slides and paper at tinyurl.com/fisher-theorem

In 1930, the famous biologist, statistician and eugenicist Ronald Fisher stated his ‘fundamental theorem of natural selection’:

> *The rate of increase in fitness of any organism at any time is equal to its genetic variance in fitness at that time.*

In 1972, George R. Price wrote:

> *It has long been a mystery how Fisher derived his famous 'fundamental theorem of Natural Selection' and exactly what he meant by it.*
>
> *He compared this result to the second law of thermodynamics, and described it as holding 'the supreme position among the biological sciences'. Also, he spoke of the 'rigour' of his derivation of the theorem and of 'the ease of its interpretation'. But others have variously described his derivation as 'recondite' (Crow & Kimura), 'very difficult' (Turner), or 'entirely obscure' (Kempthorne). And no one has ever found any other way to derive the result that Fisher seems to state.*

George R. Price explains:

> *In addition to the central confusion resulting from the use of the word fitness in two highly different senses, Fisher's three publications on his theorem contain an astonishing number of lesser obscurities, infelicities of expression, typographical errors, omissions of crucial explanations, and contradictions between different passages about the same point.*

Some say Fisher meant this:

> *The time derivative of the mean fitness of a population equals the variance of its fitness.*

But this is only true under very limited circumstances. We'll see why — and see what 'fitness' means here.

I'll also prove a *true* theorem correcting Fisher's. It's a simple, general result on dynamical systems and information theory.

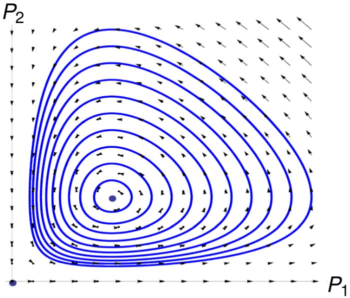
Ironically, the concept that saves the day is 'Fisher information'. I'll explain that too.

Fisher didn't use differential equations, but population biology often uses these, so I'll translate his ideas into that language.

The most famous example is this predator-prey model:

$$\frac{dP_1}{dt} = \alpha P_1 - \beta P_2 P_1$$

$$\frac{dP_2}{dt} = \gamma P_1 P_2 - \delta P_2$$



- $P_1(t)$ is the population of rabbits at time t.
- $P_2(t)$ is the population of wolves at time t.

THE GENERAL LOTKA–VOLTERRA EQUATION

Suppose we have self-replicating entities of different kinds:

- organisms belonging to different species
- genes of different alleles
- restaurants belonging to different chains
- people with different beliefs
- game-players with different strategies
- etc.

I'll call them **replicators** of different **types.**

Let $P_i \colon \mathbb{R} \to (0, \infty)$ be the **population** of the *i*th type, as a function of time.

The **Lotka–Volterra equation** says

$$\frac{d}{dt} P_i = f_i(P_1, \dots, P_n)\, P_i$$

where $f_i(P_1, \dots, P_n)$, the **fitness** of the *i*th type, can depend on the populations of all the types.

Here $f_i \colon (0, \infty)^n \to \mathbb{R}$ is any continuous function.

The probability that a randomly chosen replicator belongs to the *i*th type is

$$p_i = \frac{P_i}{\sum_j P_j}$$

The **mean fitness** is

$$\langle f \rangle = \sum_j f_j \, p_j$$

where f_j is short for $f_j(P_1, \ldots, P_n)$.

The **variance in fitness** is

$$\mathrm{Var}(f) = \sum_j p_j \left(f_j - \langle f \rangle \right)^2$$

In this framework, Fisher's supposed claim:

> *The time derivative of the mean fitness of a population equals the variance of its fitness.*

amounts to this:

$$\frac{d}{dt}\langle f \rangle = \mathrm{Var}(f)$$

We'll see it's not true — but $\mathrm{Var}(f)$ equals something *else* that's interesting.

First let's figure out

$$\dot{p}_i = \frac{d}{dt} p_i$$

$$\begin{aligned}\dot{p}_i &= \frac{d}{dt}\frac{P_i}{\sum_j P_j} \\ &= \frac{\dot{P}_i}{\sum_j P_j} - \frac{P_i\left(\sum_j \dot{P}_j\right)}{(\sum_j P_j)^2}\end{aligned}$$

Using the Lotka–Volterra equation:

$$\dot{p}_i = \frac{f_i P_i}{\sum_j P_j} - \frac{P_i\left(\sum_j f_j P_j\right)}{(\sum_j P_j)^2}$$

Using the definition of p_i again:

$$\begin{aligned}\dot{p}_i &= f_i p_i - \big(\sum_j f_j p_j\big) p_i \\ &= \Big(f_i - \langle f \rangle\Big) p_i\end{aligned}$$

THE REPLICATOR EQUATION

This is called the **replicator equation**:

$$\dot{p}_i = \Big(f_i - \langle f \rangle \Big) p_i$$

For the fraction of people like you to increase, you don't need to be fit. You just need to be fitter than average!

Now let's calculate the time derivative of the mean fitness:

$$\begin{aligned} \frac{d}{dt}\langle f\rangle &= \frac{d}{dt}\sum_i f_i\, p_i \\ &= \sum_i \frac{df_i(P_1,\ldots,P_n)}{dt}\, p_i + f_i\, \dot{p}_i \end{aligned}$$

We'll show the *second* term is $\mathrm{Var}(f)$. So, Fisher's supposed claim

$$\frac{d}{dt}\langle f\rangle = \mathrm{Var}(f)$$

is true if and only if the *first* term vanishes, e.g. if each fitness function f_i is constant.

Let's compute that second term:

$$
\begin{aligned}
\sum_i f_i\,\dot{p}_i &= \sum_i f_i \Big(f_i - \langle f\rangle\Big) p_i \\
&= \sum_i f_i \Big(f_i - \langle f\rangle\Big) p_i \;-\; \overbrace{\langle f\rangle \sum_i \Big(f_i - \langle f\rangle\Big) p_i}^{\text{this is zero}} \\
&= \sum_i \Big(f_i - \langle f\rangle\Big)\Big(f_i - \langle f\rangle\Big) p_i \\
&= \mathrm{Var}(f)
\end{aligned}
$$

So,

$$\frac{d}{dt}\langle f\rangle \neq \mathrm{Var}(f)$$

except in very special cases, but

$$\sum_i f_i\,\dot{p}_i = \mathrm{Var}(f)$$

whenever the Lotka–Volterra equation holds.

The problem: what's so great about

$$\sum_i f_i \dot{p}_i \quad ?$$

Here *information theory* enters the stage.

IT'S ALL RELATIVE — EVEN INFORMATION!

When you learn something, how much information do you gain?

It depends on what you believed before!

We can model hypotheses as probability distributions. When you update your prior hypothesis p to a new one q, how much information have you gained?

This much:

$$I(q,p) = \sum_{i=1}^{n} q_i \ln\left(\frac{q_i}{p_i}\right)$$

This is the **information of *q* relative to *p***, also called the "information gain" or "Kullback–Leibler divergence".

$$I(q,p) \geq 0 \qquad \text{and} \qquad I(q,p) = 0 \iff q = p$$

For example, suppose we flip a coin you think is fair. Your prior hypothesis is this:

$$p_H = \frac{1}{2} \qquad p_T = \frac{1}{2}$$

Then you learn it landed heads up:

$$q_H = 1 \qquad q_T = 0$$

The relative information is 1 bit:

$$I(q,p) = 1 \ln\left(\frac{1}{1/2}\right) + 0 \ln\left(\frac{0}{1/2}\right) = \ln 2$$

where we define $0 \ln 0 = 0$. You have gained 1 bit of information.

But suppose you think there's only a 25% chance of heads:

$$p_H = \frac{1}{4} \qquad p_T = \frac{3}{4}$$

Then you learn the coin landed heads up:

$$q_H = 1 \qquad q_T = 0$$

Now the relative information is higher:

$$I(q,p) = 1 \ln\left(\frac{1}{1/4}\right) + 0 \ln\left(\frac{0}{3/4}\right) = \ln 4 = 2 \ln 2$$

You have gained 2 bits of information!

THE FISHER INFORMATION METRIC

How can we quantify the *rate of learning?*

Here we face a “paradox”:

For any probability distribution $p(t)$ that changes with time in a differentiable way, we have

$$\frac{d}{dt} I(p(t), p(t_0))\Big|_{t=t_0} = 0$$

for all times t_0.

“To first order, you’re never learning anything new.”

However, as long as the velocity $\dot{p}(t_0)$ is nonzero, we have

$$\frac{d^2}{dt^2} I(p(t), p(t_0))\Big|_{t=t_0} > 0$$

"To second order, you're always learning something new... unless your opinions are fixed."

This lets us define a "rate of learning" — that is, the "speed" of the changing probability distribution $p(t)$.

Namely, define the length of the vector $\dot{p}(t_0)$ by

$$\|\dot{p}(t_0)\|^2 = \frac{d^2}{dt^2} I(p(t), p(t_0))\Big|_{t=t_0}$$

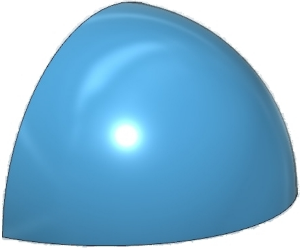
This notion of length defines a Riemannian metric on the space of probability distributions: the **Fisher information metric.**

We can calculate the Fisher information metric and show

$$\boldsymbol{g(v, w) = \sum_i \frac{v_i w_i}{p_i}}$$

where v, w are tangent vectors to the space of probability distributions at the point p: that is, n-component vectors with $\sum_i v_i = \sum_i w_i = 0$.

This makes the space of probability distributions round:



So, the square of the “rate of learning” is

$$\|\dot{p}\|^2 = g(\dot{p}, \dot{p}) = \sum_i \frac{\dot{p}_i^2}{p_i}$$

Now suppose p obeys the replicator equation:

$$\dot{p}_i = \Big(f_i - \langle f \rangle\Big) p_i$$

Then

$$\begin{aligned} \|\dot{p}\|^2 &= \sum_i \frac{\dot{p}_i^2}{p_i} \\ &= \sum_i \Big((f_i - \langle f \rangle\Big)^2 p_i \\ &= \mathrm{Var}(f) \end{aligned}$$

THE FUNDAMENTAL THEOREM OF NATURAL SELECTION

Theorem. If the populations $P_i \colon \mathbb{R} \to (0,\infty)$ obey

$$\frac{d}{dt} P_i = f_i(P_1, \dots, P_n)\, P_i$$

for some continuous functions $f_i \colon (0,\infty)^n \to \mathbb{R}$, then the probabilities

$$p_i = \frac{P_i}{\sum_j P_j}$$

obey

$$\|\dot{p}\|^2 = \mathrm{Var}(f)$$

The square of the rate of learning is the variance of the fitness!

SOME LESSONS

Fisher was close to stating a true result. The biggest missing ingredient was Fisher information, which he himself had invented earlier in 1922.

Often we can make progress just by combining ideas we already have. Often we hold the keys to our own problems, but just don’t notice it.

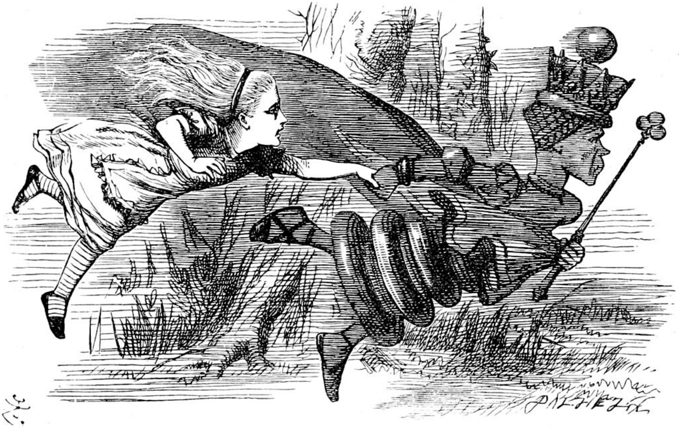
On the other hand, Fisher’s idea that nature progresses toward an optimum was wrong.

In the simple but general model we considered, it’s not *the rate of increase in mean fitness* that equals the variance in fitness. It’s the *rate at which information is updated*.

Diversity may not lead to improvement, but it does lead to change.

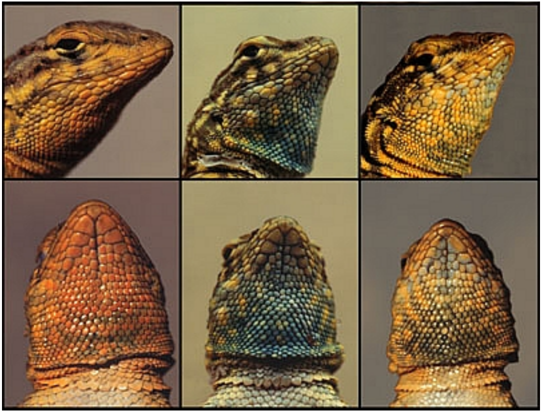
The **Red Queen Hypothesis** says that replicators must keep changing simply to survive amid other changing replicators.

THE RED QUEEN HYPOTHESIS

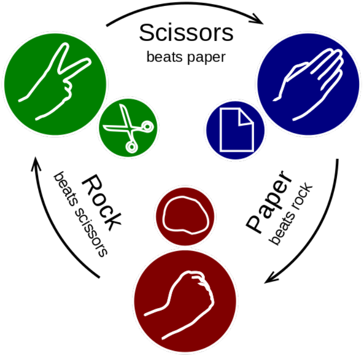


"Now, here, you see, it takes all the running you can do, to keep in the same place."

For example, in males of the common side-blotched lizard, orange beats blue, blue beats yellow, and yellow beats orange:

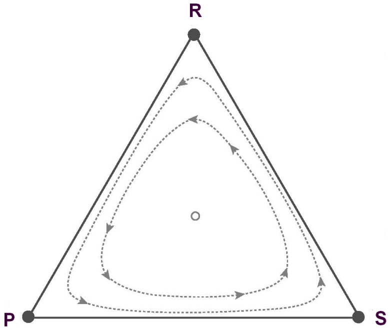


We can model this using the replicator equation by assuming lizards play randomly chosen opponents in the rock-paper-scissors game, with fitness determined by the game's outcome:



- Sinervo and Lively, The rock-paper-scissors game and the evolution of alternative male strategies, *Nature* **380** (1996).

It’s possible for the replicator equation to give this dynamics for the probability distribution of strategies ‘rock’, ‘paper’ and ‘scissors’:



There is a steady state, but it is not an attractor. Mean fitness is not always increasing. In general, *the population never stops learning new information!*

For more, try:

- John Baez, The fundamental theorem of natural selection.
- John Baez and Blake Pollard, Relative entropy in biological systems.
- Marc Harper, The replicator equation as an inference dynamic.
- Marc Harper, Information geometry and evolutionary game theory.