Supplementary Materials: The More You Know, the More You Can Grow: An Information Theoretic Approach to Growth in the Information Age

Martin Hilbert

1. Short Information Theory Primer

In essence, information theory defines information in terms of uncertainty and frames uncertainty in terms of probability theory ([1,2]). The less uncertainty, the more information and vice versa. The two basic metrics of information theory, entropy and mutual information, have a tight relation to the more well-known metrics of variance and covariance (e.g. [3]). Variance and entropy both measure diversity, while covariance and mutual information both measure the difference between joint and independent distributions. In information theory, entropy and mutual information are the answers to the two fundamental questions Shannon solved in his seminal 1948 paper.

Entropy is the result of his 'source coding theorem' and is the answer to the question of the purest state of information. It asks how many symbols are minimally needed to represent all information (after the elimination of redundant data). Formally it is calculated as $H(X) = -\sum_x p(x) * \log p(x)$, with $p(x)$ being the probability of all realizations x of the random variable X. It can be understood as a measure of uncertainty. With a logarithm of base 2 it measures uncertainty in bits, or, more specifically, it indicates how often uncertainty is reduced by half in order to reveal a true state of a random variable. To show its fundamental role Shannon used the asymptotic properties of long sequences of symbols, which requires the assumptions of ergodicity and stationarity. In essence entropy quantifies the growth of possible messages in the typical set, which indicates the rate of increase message diversity.

Mutual information is the result of Shannon's 'noisy channel coding theorem' and answers the question of the achievable communication rate over a noisy channel (the channel capacity). It is the uncertainty the sending and the receiving variables have in common. In information theory this is often depicted with the help of a Venn-diagram, such as in Figure 3 in the main text and in Figure S1 below. The mutual information is the shared intersection of the circles, which represent the entropies (here of the sender and receiver). Shannon's reasoning is that the mutual information between both is the uncertainty in the receiver, minus the uncertainty that remains in the receiver after the communication: $I(S;R) = H(R) - H(R|S)$. Since the arising common ground is supposed to be mutual, it has to be the same for the sender and the receiver and is therefore symmetrical. Figure S1 presents the case where the mutual information between S and R is calculated as the the uncertainty of R, minus the remaining uncertainty of R when knowing S.

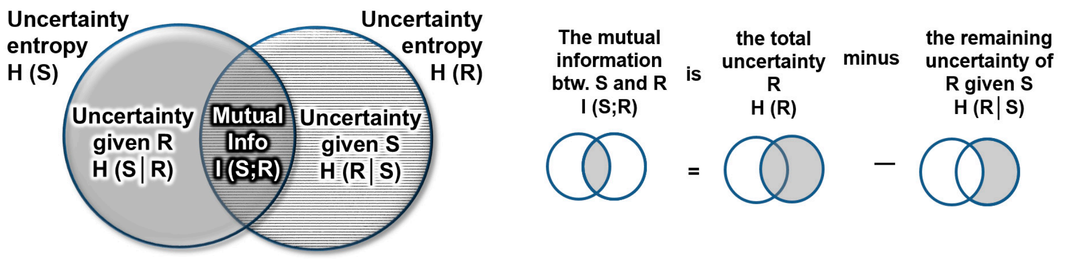
Uncertainty entropy H (S)
Uncertainty entropy H (R)
Uncertainty given R H (S│R)
Mutual Info I (S;R)
Uncertainty given S H (R│S)
The mutual information btw. S and R I (S;R)
is
the total uncertainty R H (R)
minus
the remaining uncertainty of R given S H (R│S)
=
—

Figure S1. Venn-diagrams of mutual information (intersection) and entropies (circles).

Mutual information is calculated as the ratio between the joint and its independent distribution. $I(X;Y) = \sum_{x,y} p(x,y) * \log\frac{p(x,y)}{p(x)*p(y)}$. It is a special case of the more general Kullback-Leibler relative entropy, which is calculated as the ratio between any two distributions of the same variable:

$D_{KL}(P||Q) = \sum_{x,y} p(x,y) * \log \frac{p(x,y)}{q(x,y)}$. One typical way of explaining the role of mutual information in information theory textbooks is through the Gedankenexperiment of the "noisy typewriter" (see [1,2]). The noisy typewriter starts with 26 equally likely symbols as input $p(In) = 1/26$. The uncertainty of receiving a specific one as output is therefore $H(In) = \log 26$. The noisy channel either sends input correctly or transforms it into the next letter of the alphabet with probability 0.5 (see Figure S2). This implies that receiving a certain output, there is an uncertainty of 1 bit about the original input: $H(In|Out) = -(0.5 * \log 0.5 + 0.5 * \log 0.5\ ) \ = 1$. The most straightforward way to avoid this is to only use a non-confusable subset of 13 uniquely identifiable inputs (see Figure S2). The capacity of this channel in terms of its mutual information is $I(In; Out) = \log 13$. This can be calculated as $I(In; Out) = H(In) - H(In|Out) = \log 26 - 1 = \log 13$ (compare the representation in Venn diagrams in Figure 4).

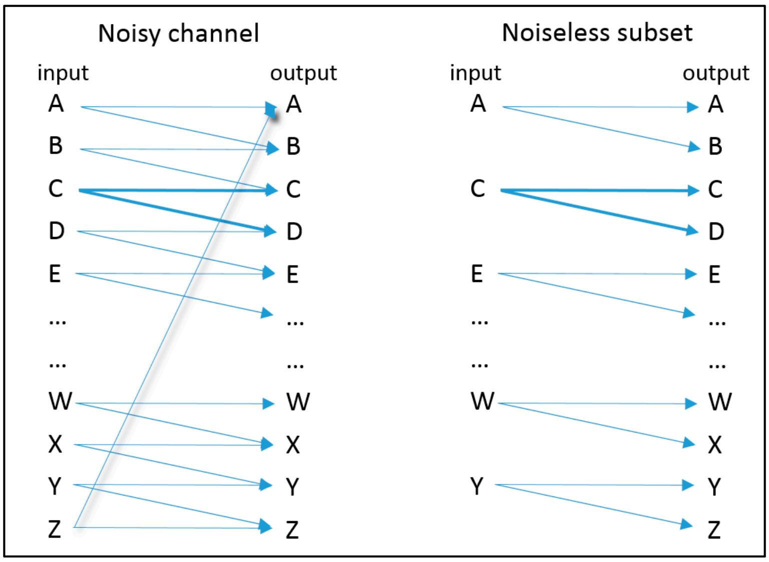

Figure S2. Noisy typewriter and its noiseless subset.

The logic of the noisy typewriter can be generalized, which is presented in Figure 2 of the main text. Following this logic, the fact that optimal growth turns $D_{KL}(P^+(g,e)\|P(g,e))$ into $I(G_s^+; E)$ (Equation (5) of the main text) shows that optimal population growth quantifies the amount of structure in the updated population that unequivocally comes from the environment through the mutual information between both. Or, in the normative sense, fitness can be optimized by searching for the channel constellation for which each channel output in the updated population can be assigned an unequivocal channel input from the environmental distribution.

Information theory is sometimes a bit inconsistent with notation. For example, the random variables embraced by an absolute entropy are represented by majuscule letters that omit the reference to the distribution, e.g. $H(E|G^+)$, while relative entropies use the opposite rule, e.g. $D_{KL}(P^+(g,e)\|P(g,e))$ (see [1]). As a result, our add-on notations like $...^+$ and $..._s$ are attached to different letters respectively.

2. Decomposing growth into information

Three steps are involved in the reformulations resulting in Equation (2). First, an expected value is taken on $\overline{\overline{W}}$ (or its log), which is justified by the fact that the expected value of a constant is the same constant (e.g. $\log \overline{\overline{W}} = E[\log \overline{\overline{W}}]$). Second, we employ a revers form of the so-called replicator equation to decompose average population fitness per environment into lower level type fitness:

$$\begin{aligned}\log\bar{\bar{W}} = E_{p^{+}(g^{+}|e)}[\log\bar{\bar{W}}] = \sum p^{+}(g^{+}|e)[\log\{\bar{W}(e)\}^{p(e)}] = \sum p^{+}(g^{+}|e)p(e)\left[\log\left\{w(g,e)\frac{p(g|e)}{p^{+}(g^{+}|e)}\right\}\right] \\ = \sum_{g,e} p^{+}(g^{+},e)[\log w(g,e)] - \sum_{g,e} p^{+}(g^{+},e)\log\frac{p^{+}(g^{+},e)}{p(g,e)}\end{aligned} \tag{S1}$$

Third, the reformulation of Equation (2) is obtained by replacing the true fitness values $w(g,e)$ with the weighted hypothetical diagonal fitness values (the share of the corresponding hypothetical fitness value: $w(g = i, e) = {}_{hyp}^{\;d}W(e) * m(e|g = i)$), and then expanding with the term $\frac{p^{+}(e|g^{+})}{p^{+}(e|g^{+})}$.

$$\log\bar{\bar{W}} = E_{p^{+}(g^{+},e)}\left[\log\left(\left\{{}_{hyp}^{\;d}W(e) * m(e|g)\right\} * \frac{p^{+}(e|g^{+})}{p^{+}(e|g^{+})}\right)\right] - D_{KL}(P^{+}(g^{+},e)\|P(g,e)) = \text{Equation(2)} \tag{S2}$$

3. Average Updating

We work with the joint probability distributions $P(g,e)$ and $P^{+}(g^{+},e)$. It is redundant to have the superscript $^{+}$ both on P^{+} and G^{+} (as only P changes), but it reminds of the fact that updating affects P but not E, and aims at integrating different notational habits from different disciplines. We calculate average updating, as $p^{+}(g^{+},e) = p(e) * p^{+}(g^{+} \mid e) = p(e) * p(g|e) * \frac{w(g,e)}{\bar{W}(e)}$. Note that on contrary to many traditional game theoretic setups, the initial generation $P(G)$, and $P(E)$ are not naturally independent, because $P(G|E)$ arises from the empirically detected fitness values $\bar{W}(e)$ and $w(g,e)$.

During each new period (and therefore at each environmental state e), the bet-hedging strategy assures that $p(g|e)$ is constant. This is the essence of bet-hedging: since we do not know when which environmental state will occur (only its probability), we look for a population distribution $P(G|e)$ that is hold constant at each step /environmental state. This implies that there is a proactive strategy $(\ldots_{s})$ to counteract the natural selection processes ongoing between each step (see main text). In practice this is done through constant redistribution from winning to loosing types during each step.

When calculating the force of selection through average updating, $P^{+}(g^{+},e)$, we consider selection pressure without such redistribution (the perspective with redistribution would not be insightful, as it would simply reflect the result of the redistribution strategy: $P(g,e) = P^{+}(g^{+},e)$). So in the case where we analyze bet-hedging, our equations evaluate the underlying average evolutionary selective pressures ongoing during bet-hedging. Let's analyze the following decomposition:

$$D_{KL}(P_{s}^{+}(g^{+},e)\|P_{s}(g,e)) = D_{KL}(P_{s}^{+}(g^{+})\|P_{s}(g)) + D_{KL}(P_{s}^{+}(e|g^{+})\|P_{s}(e|g)) \tag{S3}$$

The effect of bet-hedging on this equation is as follows: strategy based bet-hedging fixes $P_{s}(G|e)$ for each environment e (per definition of bet-hedging). This leads to the fact that $P_{s}(E|g) = P(E)$ (Fixing $P_{s}(G|e)$ over all e results in $P_{s}(G|E) = P_{s}(G)$. Therefore $p_{s}(e|g) = \frac{p_{s}(g,e)}{p_{s}(g)} = \frac{\cancel{p_{s}(g|e)}*p(e)}{\cancel{p_{s}(g)}}$.)and sets $D_{KL}(P_{s}^{+}(e|g^{+})\|P_{s}(e|g)) = D_{KL}(P_{s}^{+}(e|g^{+})\|P(e))$ on the right hand side of the equation. As stated in the main text, optimized bet-hedging aims for a distribution that results in a fixed point in which the distribution before updating in every environment $P_{s}(G)$ is the same as the average distribution over all environmental states after updating $P_{s}^{+}(G^{+})$, eliminating $D_{KL}(P^{+}(g^{+})\|P(g))$.

Note that while $P_{s}(G|e) = P_{s}(G)$ (since $P_{s}(G|e)$ is held constant by bet-hedging for all e), $P_{s}^{+}(G^{+}|e) \neq P_{s}^{+}(G^{+})$. This is because selection still acts on the population through replicator dynamics based on our empirically observed growth values $\bar{W}(e)$ and $w(g,e)$, even so—in practice—resources might get redistributed (in parallel or thereafter) to assure that the distribution before updating in the next environmental state during the time series again maintains fixed shares to enter the next round of updating. In essence we start with a fixed distribution $P_{s}(G|e)$ for all environments e and use the replicator equation to obtain our average distributions after general updating $P_{s}^{+}(G^{+}|e)$.**Error! Bookmark not defined.** This implies that $P_{s}(G) = P_{s}(G|e) \neq P_{s}^{+}(G^{+}|e)$. In other words, $P_{s}(G)$ and $P(E)$ are independent, but $P_{s}^{+}(G^{+})$ and $P(E)$ are not (they are dependent as they are affected by average updating). Optimal bet-hedging now looks for a population distribution over all environments that fulfills the condition $P_{s}(G) = P_{s}^{+}(G^{+})$. In words: selection acts during average updating per environmental state, while optimality implies that the average updated population distribution over all environmental states stays constant.

4. Optimality of $D_{KL} = I$

To show the two way relation between the appearance of the mutual information and optimal fitness, we make use of the fact that optimal growth is achieved for Kelly's case with a diagonal fitness matrix (for a proof see for example Chapter 6 in [1]), and combine it with the fact that the same result can be achieved by expressing existing type fitness as a combination of (hypothetical) fitness values from a diagonal matrix: $w(g = i, e) = \sum_e p(e|m = i) * {}_{hyp}^{d}W(e)$ (see [4,5]). In other words, we work with the weighted fitness matrix that achieves optimal fitness in the region of bet-hedging.

The condition that $D_{KL}(P^+(g^+, e)||P(g, e)) = \sum_{g,e} p^+(g, e) \log \frac{p^+(g,e)}{p(g,e)} = \sum_{g,e} p^+(g, e) \log \frac{p^+(g^+,e)}{p^+(g^+)*p(e)} = I(G^+; E)$ is fulfilled for $p(g, e) = p^+(g^+) * p(e)$, since the mutual information is defined as the relative entropy between the joint distribution and the corresponding independent distribution. With the help of the replicator equation, this condition can be rewritten as $p(g|e) = p^+(g^+) = \sum_e p(e) * p^+(g^+|e) = \sum_e p(e) * p(g|e) * \frac{w(g,e)}{\bar{W}(e)}$. We include the assumption of stable shares of types in all environments: $p(g|e = i) = constant$ for all i. In our case this is achieved either through proportional bet-hedging or by betting all resources on one type, but can be achieved by any other kind of stable equilibrium in the population shares. This cancels out $p(g|e)$ and we obtain the condition that the time-average of relative fitness is equal to 1:

$$1 = \sum_e p(e) \frac{w(g, e)}{\bar{W}(e)} \tag{S4}$$

Given that $p(g|e)$ is constant, we can expand with it:

$$\begin{gathered} p(g|e) * 1 = p(g|e) * \sum_e p(e) = p(g|e) * \sum_e p(e) \frac{w(g, e)}{\bar{W}(e)} = \sum_e p(e) p(g|e) \frac{w(g, e)}{\bar{W}(e)} \\ \sum_e p(g, e) = p(g) = \sum_e p(e) * p^+(g^+|e) = p^+(g^+) \end{gathered} \tag{S5}$$

From the above equality $p(g, e) = p^+(g^+) * p(e)$, it follows that ~~$p(g)$~~ $* p(e|g) =$ ~~$p^+(g^+)$~~ $* p(e)$. We start the two-way proof by showing that optimal fitness implies the existence of mutual information in our decomposition: $optimality \Rightarrow D_{KL} = I$. We express both the numerator and denominator of Equation (S5) with their equivalent expressions from the noiseless channel fitness matrix:

$$\sum_e p(e) \frac{w(g, e)}{\bar{W}(e)} = \sum_e p(e) \frac{p(e|m_i) * {}_{hyp}^{d}W(e)}{p(e) * {}_{hyp}^{d}W(e)} = \sum_e p(e|m_i) = 1 \tag{S6}$$

where the last step follows from stochasticity of the weighting matrix (in this case of optimal bet-hedging in mixed fitness landscapes $P(E|m_i) = P(E|g_i)$). This shows that in the case of optimal growth the relative entropy term D_{KL} turns into mutual information I.

One additional assumption is required for the complementary proof that: $D_{KL} = I \Rightarrow optimality$. It is that the fitness matrix of population types and environmental shares is linearly independent (referring to independence in the sense of linear algebra, not to probabilistic independence). This seems to be a reasonable demand, as redundant types or environmental states should be merged. We begin by reformulating the stochastic weighting matrix $1 = \sum_e p(e|m_i) = \sum_e \frac{w(g_i,e)}{{}_{hyp}^{d}W(e)}$. We include the restriction of the region of bet-hedging, which is $0 \leq p^+(g^+|e) \leq 1$, and replace $w(g_i, e)$ with the reverse form of the replicator equation. $\sum_e \frac{p^+({g^+}_i|e)}{p(g_i|e)} \bar{W}(e) \frac{1}{{}_{hyp}^{d}W(e)} = \sum_e \frac{p^+({g^+}_i,e)}{p(g_i|e)} \frac{\bar{W}(e)}{p(e)\, {}_{hyp}^{d}W(e)} = 1$. We introduce our starting condition, $D_{KL} = I$, which implies $p(g_i, e) = p(g_i|e) * p(e) = p^+({g^+}_i) * p(e)$, or $p(g_i|e) = p^+({g^+}_i)$. We then introduce our first assumption, that the shares of types $p(g_i|e)$ are fixed for a specific type i, which allows us to bring this term to the right hand side. This leaves us with $\sum_e p^+({g^+}_i, e)\frac{\bar{W}(e)}{p(e)\ {}_{hyp}^{d}W(e)} = p^+\left({g^+}_i\right) = \sum_e p^+({g^+}_i, e)$. We can rewrite this in matrix form for all types i over all types and environmental states:

$$\mathbf{P} * \boldsymbol{w} = \boldsymbol{p}^+ \tag{S7}$$

$$\begin{bmatrix} p^+(g^+ = 1, e = 1) & p^+(g^+ = 1, e = 2) & \dots \\ p^+(g^+ = 2, e = 1) & \dots & \dots \\ \dots & \dots & \dots \end{bmatrix} \begin{bmatrix} \bar{W}(e = 1)/[p(e = 1)\ {}_{hyp}^{d}W(e = 1)] \\ \bar{W}(e = 2)/[p(e = 2)\ {}_{hyp}^{d}W(e = 2)] \\ \dots \end{bmatrix} = \begin{bmatrix} \sum_e p^+(g^+ = 1, e) \\ \sum_e p^+(g^+ = 2, e) \\ \dots \end{bmatrix} \tag{S8}$$

If the rank of the coefficient matrix $\mathbf{P}$ is equal to the rank of the respective augmented matrix $\mathbf{P}^{\#}$, the system is consistent and must have at least one solution (Rouché–Capelli theorem). This is the case here, since the last column of the augmented matrix, $\boldsymbol{p}^+$, can easily be set to 0s through column operations of $\mathbf{P}$ (which do not affect the rank; i.e. subtracting each column once). So whatever the rank of $\mathbf{P}$ will be the rank of $\mathbf{P}^{\#}$**.** We furthermore know that the solution is unique if the rank is equal to the number of variables. Otherwise we have infinitely many solutions. The trivial case for the condition of a unique solution is Kelly's diagonal fitness matrix, with non-zero values only in the diagonal, which is already in reduced echelon form. In the case that either different types or different environments are linearly dependent we obtain infinitely many solutions (again, dependence refers to the concept from linear algebra here, not to random variables). If these redundant states and types are merged, the number or variables is equal to the rank. To identify the unique solution, we employ a method that works for many such problems: guess and verify. Plugging in $\frac{\bar{W}(e)}{p(e)\ {}_{hyp}^{d}W(e)}$ for all environments shows that it is the unique solution to the system. This confirms that in the case that the relative entropy term D_{KL} turns into mutual information I, optimal growth is achieved.

References

1. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley-Interscience: Hoboken, NJ, USA, 2006.
2. MacKay, D.J.C. *Information Theory, Inference and Learning Algorithms*, 1st ed.; Cambridge University Press: Cambridge, UK, 2003.
3. Garner, W.R.; McGill, W.J. The relation between information and variance analyses. *Psychometrika* **1956**, *21*, 219–228.
4. Donaldson-Matasci, M.C.; Lachmann, M.; Bergstrom, C.T. Phenotypic diversity as an adaptation to environmental uncertainty. *Evol. Ecol. Res.* **2008**, *10*, 493–515.
5. Donaldson-Matasci, M.C.; Bergstrom, C.T.; Lachmann, M. The fitness value of information. *Oikos* **2010**, *119*, 219–230.