Constrained information optimization – a seventh rule of life?

Janik Schüttler

June 5, 2020

In the course we have discussed six possible physical rules or characteristics of living systems: 1) the principle to trade energy for order, 2) even though some of the order comes for free; 3) dynamics are neither trivial nor too complex (usually scale sub-linearly with system size); 4) criticality to allow for dynamically jumping between states; 5) modular and 6) fractal structures. In this document I want to conceptually argue for a constrained information optimization, a notion I will define further into the text, as another principle of biological organization. I will discuss what information is, how it is related to thermodynamics, why information is important for biological systems, but can be optimized only constrained, introduce the information bottleneck methods to formalize constrained information optimizations, and present two examples of living systems, where a notion of optimal information processing has been found.

Why would a living agent care about information? Imagine for example two bacteria that are looking for food. Bacterium A swims around randomly in the hope to come across food eventually. Bacterium B, in contrast, can "smell" food better and better the closer it gets to the food, an ability that the bacterium can use to direct its movement towards food. This behavior is called chemotaxis, it will make for a good example throughout this essay. Not sur-prisingly, using its chemotaxing abilities bacterium B will be more successful in foraging than bacterium A. The advantage bacterium B has over bacterium A is that B can utilize *information* more efficiently than A, information about the location of nearby food. There are many more examples one could name, from cellular to organism levels of biological organization, simply transfer the above example to finding and assessing possible mates, as an example. This demonstrates how efficient information processing can correlate with biological fitness.

What are the natural limits to information processing? One answer is noise. Most of us will have failed at some point in our lives at identifying objects using our eyes when there is simply not enough light, but even a tiny source of light might help us enough to do so. This is our natural limit of information acquisition due to noise. Just like us, microorganisms have to deal with such fundamental limits caused by noise, as the famous work by Berg and Purcell demonstrates [1]. Imaging again a bacterium sensing molecule X. One of the Berg-Purcell arguments concerns a cell's limit to measure the local concentration c of molecule X. To make our inner physicists happy, let's straightforwardly approximate the cell to be spherical with radius r. The total number of molecules X inside the cell will be proportional to the cell volume and

the X's concentration, $N \sim cr^3$, at any given time, which the cell can measure only up to a noise level $\delta N \sim \sqrt{N}$ that can be taken to scale with the square root, a fundamental scaling relation in statistics. Thus, the relative error for one measurement of molecule's X concentration is of order

$$\frac{\delta c}{c} \sim \frac{\delta N}{N} = \frac{1}{\sqrt{N}} = \frac{1}{\sqrt{cr^3}}.$$

But wait, the cell has been optimized for billions of years thanks to mother evolution and has the clever idea to now take n measurements in a time frame T to decrease the error. The cell has to be careful to not measure the same molecules twice, or otherwise the samples would not be statistically independent, which forces the cell to wait roughly the time the X molecules take to diffuse out of the cell, r^2/D, until a new measurement can be done[1]. With $n = \frac{T}{r^2/D}$ measurements, the limit becomes

$$\frac{\delta c}{c} \sim \frac{1}{\sqrt{nN}} = \sqrt{\frac{r^2/D}{T} \frac{1}{cr^3}} = \frac{1}{\sqrt{DTcr}}.$$

This is an approximation for the limit of concentration differences a cell can detect. It tells us that the only way for a cell to measure a concentration c more precisely is to increase the duration of measurement or to grow (how annoying).

This is the story of noise, but there is another part to the limits of information that concern energy. More measurements do not only take more time, they also burn energy, which physicists have noticed only in the 20th century. The classic thought experiment to demonstrate this connection is Maxwell's demon, an idea that dates back all the way to the year of 1867. Maxwell imagined an actor with access to infinite information, the demon,
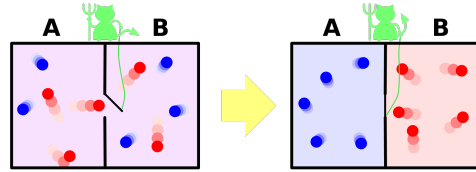

Here D denotes X's diffusion constant.

Figure 1: The Maxwell demon. In the left panel, the demon quickly opens the door between the box A and B to allow the fast (red) gas molecule to move to box B, after which the demon closes the door again. Many iterations of this procedure heat up the box B and cool down the box A, as illustrated in the right panel. Illustration from from Wikipedia.

that sits between two boxes filled with a gas and connected via an initially closed door. Using its information about the speed of the gas molecules, the demon would quickly open and close the door to allow a fast gas molecule to move from the left to the right box and similarly allow slow molecules to move from the right box to the left. After many repetitions of this procedure, fast molecules would accumulate in the right box and it warms up over time, while the gas in the left box cools down. This procedure increases entropy and thereby seems to violate the second law of thermodynamics. The other way round, if we agree on the second law of thermodynamics there must be some natural bounds on information coming from thermodynamics. The paradox has puzzled physicists until today and demon-like systems have even been experimentally demonstrated [2].

Landauer argued in 1961 that the information processing capabilities of the demon must be taken into account. He noticed that an erasure of information must be balanced by an entropic cost, i.e. it must release heat, thereby

making Maxwell's demon obey the second law of thermodynamics [3]. The erasure of one bit of information costs 0.7 the thermal heat $k_B T$,

$$E = k_B T \ln 2.$$

In all fairness, I have to note that this energetic cost is absurdly small, roughly $0.0175\,\text{eV}$ and possibly of no outstanding role in optimizing information processing, but this paradox shows beautifully how deeply information is connected to fundamental physics. In modern processors with billions of operations per second, Landauer's principle is thought to become important, which has motivated modern research on, for example, reversible computing architectures [4].

If you have not learned about information previously, Landauer's principle might appear like magic – alone the fact that physicists seem to be able to talk about information quantitatively. At this point we must halt and ask the one question I assume would puzzle anyone that hasn't seen it answered yet: *what even is information*? We have a natural understanding of information, a university lecture is supposed to be informative, scrolling through Facebook as well[2], I have the information how information can be formalized, which you might not yet have, and so on. Can this intuition be formalized? The important sentence here is the following: Information between two quantities represents how much knowing one quantity reduces your uncertainty about the other. Suppose you have fallen in love with a person, but you do not know what this person thinks about you – you are uncertain about this person's attitude toward you. To reduce your uncertainty, you could ask his/her out and chat to assess whether you two might have a future. Let H denote an uncertainty function. Then your information the date, D, has provided you on how much your love likes you, denote this by A, is the difference of your uncertainty before the date, $H(A)$, and your uncertainty after the date, $H(A|D)$,

$$I(A; D) = H(A) - H(A|D).$$

On a more abstract level, the information content between two random variables X and Y is given by

$$I(X; Y) = H(X) - H(X|Y).$$

Great, we have shifted the problem to finding an uncertainty function. A great intuition for uncertainty is the "peakedness" of the probability distribution associated with its random variable. Imagine a game, in which one of your four best friends, two male, two female, are randomly picked and you must find out by asking yes/no questions who was picked. In the beginning you are uncertain, the probabilities which friend could have been chosen is uniformly $\frac{1}{4}$. Suppose you ask whether the friend is male, answered with yes, such that only two friends remain. In other words, the answer has reduced your uncertainty about which friend was picked. Now, your probability distribution is more peaked, with two probabilities zero and the other $\frac{1}{2}$. We learn: more peaked probability distributions lower your uncertainty. This might sound trivial, but comes in handy also when dealing with highly involved settings. Remarkably, under reasonable assumptions, Shannon proved in 1948 that there is a unique measure of uncertainty, the entropy [5]

$$H(X) = - \sum_x P(x) \log P(x),$$

---

[2]Although this might make for a perfect segue into a debate on information quality...

where P is the probability distribution of the random variable X. But what is this unit of bits we saw in the Landauer's principle? Information and uncertainty are most frequently measured in units of bits, in which the logarithm in the above equation is taken with respect to basis 2. To illustrate a bit, imaging a fair coin. The uncertainty you have about the outcome before you flip the coin is

$$H(\text{coin}) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1 \, \text{bit}.$$

Similarly, you are twice as uncertain about the outcome of two coin flips, two coin flips carry 2 bits of uncertainty.

After this digression into information theory, we can return to our main question. We have learned that information is on the one hand highly valuable for a living agent, but on the other hand costly and noisy. This motivates me to postulate a constrained information optimization as a fundamental principle of life. This principle could be formulated as:

> *Life is tuned to maximize information content under the constraint of minimal energy consumption for information processing.*

In the following I will expand on this by introducing some theoretical concepts and examples related to my proposed constrained information optimization principle.

A neat approach for thinking about constrained information optimization was proposed by Tishby, Pereira and Bialek in 1999 [6]. Applied to our setting, their idea was that an agent must learn about the world W by making imperfect[3] measurements M, which the agent stores in its memory, S, in order to help with future actions. Optimally, the agent would like to maximize the information about the world in its memory, $I(W; S)$, while minimizing the information it must store, $I(M; S)$. We can state the information bottleneck in equations as

$$\text{argmax}_{M \to S} \, I(W; S) - \lambda I(M; S),$$

where $M \to S$ is a rule on which parts of the measurement to store and which to discard. In other words, the agent tries to maximize accuracy and compression, by squeezing as much useful information about the world as possible into its representation of the world, but compress out any irrelevant information. The parameter λ is used to shift priority between accuracy and compression. This "squeezing of information" is the reason why the principle is termed information bottleneck. It has proven a useful framework to study all sorts of information-related topics. Noteworthy, Tishby has begun to successfully apply the information bottleneck to the theory of neural networks just a few years ago, by which he can explain neural network training in terms of two phases, information uptake and compression [7].

What if we associate information directly with rewards? A useful framework to treat such constrained optimizations are Markov decision processes (MDP). In a MDP, the agent navigates its environment by choosing an action a that might trigger a change in the system's (environment + agent) state s. State transitions may be associated with a reward[4]. Figure 2 illustrates our beloved example of chemotaxis in the language of MDP. The cell, represented by a black circle, is looking for food, the green tile, while trying to avoid dangers, the red tile. The cell's actions are to move around the grid, which changes its position and thereby

---

[3]Noisy, deficient measurement apperati etc.

[4]Rewards can be positive and negative, whereby negative rewards are rater penalties than rewards.
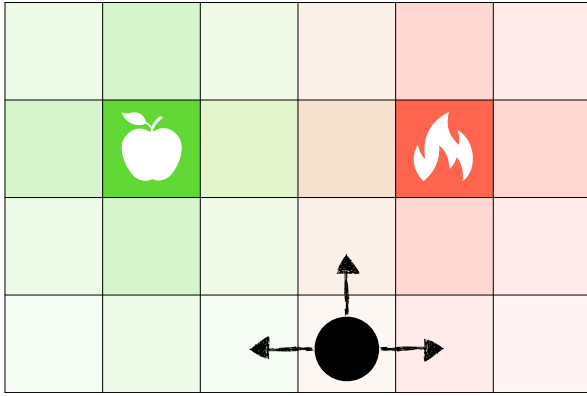
Figure 2: Example of chemotaxis as Markov decision process. An agent (black circle) can take actions and sense traces of reward (food, green tile) or danger (fire, red tile) to move around the grid. The setup was motivated by the setting of reference [8].

the state of the system. If it hits the green or red tile the cell will be rewarded or punished, respectively, all other state transitions are associated with small penalties that could be interpreted as energetic costs. To achieve its goal, the cell may sense gradients of information at its current position, otherwise the cell is blind. *Agents capable of processing information more efficiently are rewarded more.* Therefore, the cell would like to find a policy, i.e. a probability distribution over possible actions given the current state, that maximizes its utility (food). In the example of figure 2, the optimal paths walk in four steps from the bottom right corner to the food source. In the language of the information bottleneck, the cell is blind and must thus sense the world by taking a step in either direction and measuring the concentrations of reward and danger at this location, the information content of which it wants to maximize and which allows it to build an internal representation of the world over time that is, however, associated with a cost of storage the agent

aims to minimize.

But enough of the abstract talk, how well does biology perform in such a problem? Andrews and Iglesias analyzed *D. discoideum* cells' information optimality in chemotaxis [9]. In their model, a cell senses chemoattractants and responds by changing its direction of movement. They use the previously developed local-excitation, global-inhibition (LEGI) biochemical model for chemotaxis to compare their theoretical predictions. The LEGI model was developed to explain chemotaxis, with no motivation to maximize information processing. It has one parameter, a hill coefficient, that allows adjustment of the accuracy of the chemosensing. In the framework of information bottlenecks, the cell would like to minimize information it has to extract from the sensation in order to decide for a response (compression) under the constraint that the sensation should carry as much information about the chemoattractant field as possible (accuracy). The authors computed the theoretical optimal trade-off curve between information compression and accuracy and compared this theoretical prediction with the LEGI model. Amazingly, LEGI's accuracy curve agrees well with the author's optimal trade-off curve[5] [9]. This result may again be taken as a hint for a notion of information optimality in biological systems, but I presume more examples are needed.

Another beautiful example for how nature seems to have implemented such information optimal systems is the positional gene system in the early development of *Drosophila*. During development, a cell uses the concentrations of so-called gap genes as a marker for its position relative to the organism. Along the ante-

---

[5]At least for unbiased cells that are new to a chemoattractant field.

rior/posterior axis of the embryo, the expression levels of the four gap genes, *hunchhback (Hb), krüppel (Kr), giant (Gt), knirps (Kni)*, all vary such that each combination of expression levels allows cell to deduce its own relative position in the embryo with high precision. A visualization of the expression pattern can be found in the lower left plot of figure 3, where position is plotted on the x-axis versus expression levels on the y-axis for the four gap genes (different colors). How much positional information, i.e. the information the gap gene concentration offers about the position of the cell in the embryo, do these gap genes carry? Interestingly, the gap genes each provide about 2 bits of positional information, twice as much as expected from an on/off switch [10]. This suggests that the intermediate expression levels[6] provide significant additional information. Taken together, however, the overlaps in expression patterns of the four gap genes reduces their total information to about 4.2 bits. An important result is that the interplay of gap genes comes as close as 2% to the information theoretic optimal positional system – a hint for information optimality [10]? The authors were buzzed and decided to simulate alternative gap gene expression patterns in a biophysically involved, but accurate model of development [11][7]. They optimized the patterns with respect to positional information for various initial conditions and compared the opti-

mal patterns' positional information to the one of roughly 4.2 bits found in *Drosophila*. Some of the found optimal expression patterns are illustrated in figure 3 along with *Drosophila*'s pattern on the lower left. Strikingly, one of
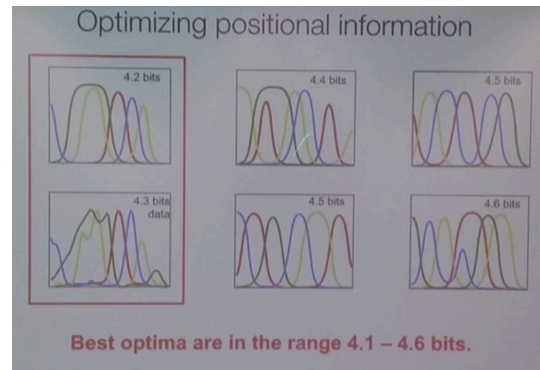

Figure 3: Real world expression patterns from *Drosophila* (lower left) compared to some of the optimized gap gene expression patterns from reference (all other). The top left expression pattern is remarkably close to *Drosophila*'s on the lower left. Image taken from a presentation by Tkacik [11].

the optimal simulated patterns (top left in figure 3) is in close agreement to the expression pattern observed in *Drosophila*, up to explainable deviations. The simulations also revealed distinct optimal expression patterns that are in parts more informative, ranging between 4.1 and 4.6 bit of positional information. However, the *Drosophila* expression pattern is the most resource-efficient in terms of associated protein costs. This is a remarkable result: life has maximized information content and minimized energetic costs, how cool is that?

Fortunately, I could have picked many more examples to discuss here. Think alone of neuroscience, the branch of science that almost screams to be studied with information theory and which is full of information-theoretic interesting optimization concepts. Just to name

---

[6]The expression levels between zero and maximal expression levels.

[7]The study has been running for many years now, but is still unpublished as of today. The here presented preliminary results were taken from a presentation by Tkacik at the 2019 Arnold Sommerfeld School on the Physics of Life. A recording of the presentation can be found at cast.itunes.uni-muenchen.de/clips/CCw7QHGdeA/vod/online.html.

some of the early examples, neurons were found to match the stimulus-response of the stimulus' cumulative probability density, which maximizes the information transmission capacity [12], and crickets are tuned close to the maximum information between the wind direction and the four neurons encoding the wind direction [13]. Evolutionary biologists are also becoming increasingly interested in biological information as a fitness-increasing resource [14, 15].

Constrained optimization of information is a nice contender for a physical principle of biological organization. However, research is still far from understanding the role of information in biology, just like most high-level functionalities of living systems. This is why one has to treat the presented concepts with care, just like anywhere in science. Sill, I find the so far reported examples of constrained information optimization wonderful. Note alone the interdisciplinarity of this essay: we have touched upon information theory, which originated from the theory of communication, the fundamental physics of energy in Landauer's principle, statistics and machine learning through Markov decision processes and the information bottleneck, and of course biology. This reflects the modern shifts in research towards interdisciplinarity over the course of the past decades – an exciting opportunity for science.

References

[1] Berg, H. & Purcell, E. Physics of chemoreception. *Biophysical Journal* **20**, 193–219 (1977).

[2] Koski, J. V., Maisi, V. F., Pekola, J. P. & Averin, D. V. Experimental realization of a Szilard engine with a single electron. *Proceedings of the National Academy of Sciences* **111**, 13786–13789 (2014).

[3] Landauer, R. Irreversibility and Heat Generation in the Computing Process. *IBM Journal of Research and Development* **5**, 183–191 (1961).

[4] Wille, R., Soeken, M., Groß e, D., Schönborn, E. & Drechsler, R. Designing a RISC CPU in Reversible Logic. In *2011 41st IEEE International Symposium on Multiple-Valued Logic*, 170–175 (IEEE, Tuusula, Finland, 2011).

[5] Shannon, C. E. A Mathematical Theory of Communication. *Bell System Technical Journal* **27**, 379–423 (1948).

[6] Tishby, N., Pereira, F. C. & Bialek, W. The information bottleneck method. *arXiv:physics/0004057* (2000). `physics/0004057`.

[7] Tishby, N. & Zaslavsky, N. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, 1–5 (IEEE, Jerusalem, Israel, 2015).

[8] Polani, D. Information: Currency of life? *HFSP Journal* **3**, 307–316 (2009).

[9] Andrews, B. W. & Iglesias, P. A. An Information-Theoretic Characterization of the Optimal Gradient Sensing Response of Cells. *PLoS Computational Biology* **3**, e153 (2007).

[10] Dubuis, J. O., Tkacik, G., Wieschaus, E. F., Gregor, T. & Bialek, W. Positional information, in bits. *Proceedings*

of the National Academy of Sciences **110**, 16301–16308 (2013).

[11] Tkacik, G. Deriving the Drosophila gap gene system ab-initio from an optimization principle. Presentation at Arnold Sommerfeld School on the Physics of Life, LMU Munich (2019).

[12] Laughlin, S. A Simple Coding Procedure Enhances a Neuron's Information Capacity. *Zeitschrift für Naturforschung C* **36**, 910–912 (1981).

[13] Theunissen, F. E. & Miller, J. P. Representation of sensory information in the cricket cercal sensory system. II. Information theoretic calculation of system accuracy and optimal tuning-curve widths of four primary interneurons. *Journal of Neurophysiology* **66**, 1690–1703 (1991).

[14] McNamara, J. M. & Dall, S. R. X. Information is a fitness enhancing resource. *Oikos* **119**, 231–236 (2010).

[15] Bergstrom, C. T. & Lachmann, M. The fitness value of information. *arXiv:q-bio/0510007* (2005). `q-bio/0510007`.