LOSSY IS LAZY

Susanne Still

Information and Computer Sciences, University of Hawaii at Mānoa,
1680 East-West Road, Honolulu, HI 96822, USA, sstill@hawaii.edu

ABSTRACT

Shannon's rate-distortion curve characterizes optimal lossy compression. I show here that the optimization principle that has to be solved to compute the rate-distortion function can be derived from a least effort principle: minimizing required thermodynamic effort necessitates the minimization of information (compatible with a given fidelity). Retaining less information costs less physical effort. In that sense, lossy compression is energy efficient, in other words, lossy is lazy.

1. INTRODUCTION

Rate distortion theory [1, 2, 3, 4] underlies much practical work in signal processing. It quantifies the rate at which data can be transmitted, given a tolerable level of fidelity. Shannon considered [1] "the set of messages of a long duration, say T seconds. The source is described by giving the probability density, in the associated space, that the source will select the message in question $[p(x)]$. A given communication system is described (from the external point of view) by giving the conditional probability $[p(y|x)]$ that if the message x is produced by the source the recovered message at the receiving point will be y."

Input messages, or input data, x, are then compressed into a representation, y, such that a certain desired level of fidelity is achieved, rather than perfect reconstruction. In other words, information is lost. In this process, some average distortion, $D[X,Y] := \langle d(x,y)\rangle_{p(x,y)}$ is encountered. The most efficient encoding compatible with a given quality of reproduction minimizes the mutual information[1] $I[X,Y] := \left\langle \ln\left[\frac{p(x,y)}{p(x)p(y)}\right]\right\rangle_{p(x,y)}$ under the constraint of fixed average distortion, D.

Shannon thus defined the rate, $R(D)$, of generating information compatible with a given distortion as the minimum of the mutual information under this constraint: [2]

$$R(D) := \min_{p(y|x)} I[X,Y] \quad \text{s.t. } D[X,Y] = D. \tag{1}$$

The minimum is taken over all possible communication systems, i.e. probabilistic assignments $p(y|x)$. The optimal rate is achievable, and algorithms exist for computing the rate-distortion function [3, 4].

This problem has a simple physical motivation, which I will now develop.

2. EFFORT OF CODING

Output messages are distributed according to $p(y)$.[3] However, when a specific input message, x, is given, then the corresponding code messages are distributed according to $p(y|x)$. Imagine a physical system which is changed from a state described by the distribution $p(y)$ to one described by $p(y|x)$. This change requires effort. How much effort?

The second law of thermodynamics tells us that we need to put in at least as much work as the resulting free energy difference, which is, on average over input x,

$$\Delta F[X,Y] := \langle F[p(y|x)]\rangle_{p(x)} - F[p(y)]\,, \tag{2}$$

where $F[p]$ denotes the *generalized*, or *nonequilibrium* free energy, $F[p] = \langle E\rangle_p + k_B T\langle \ln[p]\rangle_p$,[4] which has been used by a number of authors to describe nonequilibrium systems (see e.g. [5, 6, 7, 8, 9, 10, 11, 12, 13, 14], and references therein). It reduces to the thermodynamic equilibrium free energy when evaluated on the equilibrium distribution.

3. LEAST EFFORT PRINCIPLE

Typically, a representation of a quantity of interest is produced for some purpose, e.g. communication and reproduction of the original source data [1, 2, 4], or work extraction from a physical system [15, 16]. Let us define the function $u(x,y)$ to measure the general usefulness of a specific data representation. Its average value, $U[X,Y] := \langle u(x,y)\rangle_{p(x,y)}$ then quantifies the utility of the representation.

We are now in a position to state a *least effort principle* demanding that input data should be represented in such a way that the average free energy change (which is a lower bound on the effort) is minimized. Define the *least effort*,

[1] For simplicity, we measure information in units of the natural logarithm (nats). The shorthand $\langle\cdot\rangle_p$ denotes the average taken over the distribution p.

[2] Notation uses the convention in [4]: capital letters X and Y denote random variables. For visual clarity, all optimization problems appear without the constraints that ensure normalization and positivity of $p(y|x)$.

[3] Keep in mind that $p(y) = \langle p(y|x)\rangle_{p(x)}$.

[4] k_B is the Boltzmann constant, and T the temperature of a heat bath surrounding the system. The assumption is that the system exchanges only heat with the surroundings, and that the heat bath is large compared to the system which may be driven arbitrarily far from equilibrium by a change in external parameters. These parameter changes allow for doing work on and extracting work from the system.

$L(U)$, involved in representing x as y by the minimum free energy change compatible with utility U:

$$L(U) := \min_{p(y|x)} \Delta F[X,Y] \quad \text{s.t. } U[X,Y] = U. \tag{3}$$

The least effort function quantifies how conservative one can be with the expenditure of energy while achieving the intended utility. In other words, it measures how lazy one can afford to be.

Observations related to least effort coding have previously come up in the context of language [17, 18]. The effort of the speaker was modeled as the entropy of the code signals, while the effort for the listener was modeled as conditional entropy of the objects of reference, given the signal [18]. The combined effort was minimized, and the relative importance of the two terms was controlled by a parameter. At a critical value, Zipf's law [17] was retrieved at a phase transition [18]. While related in general spirit, the measure used in [18] is not the same as the physical effort discussed here.[5]

4. RATE-DISTORTION CURVE IS A LEAST EFFORT FUNCTION

Let a physical system that is in a state described by $p(y)$ have internal energy $E(y)$, and let the energy associated with the state described by $p(y|x)$ be denoted by $E_x(y)$. Write the difference as $\mathcal{E}(x,y) := E_x(y) - E(y)$, and denote its average by $E[X,Y] := \langle \mathcal{E}(x,y) \rangle_{p(x,y)}$. Then the least effort involved in the change $p(y) \to p(y|x)$, averaged over all x, is given by

$$\begin{aligned} \Delta F[X,Y] &= \langle E_x(y) \rangle_{p(y|x)p(x)} - k_B T H[Y|X] \\ &\quad - \langle E(y) \rangle_{p(y)} + k_B T H[Y] \qquad (4) \\ &= E[X,Y] + k_B T I[X,Y] \,. \qquad (5) \end{aligned}$$

Now consider the case that the average energy does not change, i.e. $\langle E_x(y) \rangle_{p(y|x)p(x)} = \langle E(y) \rangle_{p(y)}$, in other words, $E[X,Y] = 0$. A simple example is given by a particle in a double well potential. For simplicity of the exposition, make the potential rectangular, having an energy barrier of infinite energy between two wells of identical width and identical energy, E_0. Coarse grain the position of the particle so that $y = 0$ ($y = 1$) denotes the particle in the left (right) well. Then $E(y=0) = E(y=1) = E_0$, and hence $\langle E(y) \rangle_{p(y)} = E_0$. The particle can be forced into either well by deformation of the potential. Let $x \in \mathbb{R}$, and let the protocol that achieves this preparation of y depend on x, so that, at the end of the protocol, $y = \theta(x)$. Let, e.g.,

$$E_x(y) = \begin{cases} E_0 & \text{if } y = \theta(x) \\ \infty & \text{else} \end{cases} \,, \tag{6}$$

and

$$p(y|x) = \begin{cases} 1 & \text{if } y = \theta(x) \\ 0 & \text{else} \end{cases} \,. \tag{7}$$

Then $\langle E_x(y) \rangle_{p(y|x)} = E_0$, which is independent of x, and therefore $\langle E_x(y) \rangle_{p(y|x)p(x)} = E_0$, for any $p(x)$.

For classical systems and measurements, things can often be set up in such a way that the assumption $E[X,Y] = 0$ is valid. It could, however be violated by quantum entanglement, and also possibly in living, metabolizing agents. Both of these areas are outside the scope of this paper.

Under the assumption that the average energy does not change, the free energy change is proportional to mutual information:

$$\Delta F[X,Y] = k_B T I[X,Y] \,. \tag{8}$$

The least effort principle thus dictates minimization of mutual information.

The optimization problem in Eq. (3) can be solved using the method of Lagrange multipliers. The constraint is added to the objective function, with a Lagrange multiplier that effectively controls the trade-off between minimal effort and achieved utility. For data compression, utility is related to fidelity and can be identified with negative distortion.

A least effort data compression then has to solve

$$\min_{p(y|x)} \Big(\Delta F[X,Y] + \lambda D[X,Y] \Big) \,. \tag{9}$$

Comparison with Eq. (8) reveals that this is equivalent to $\min_{p(y|x)} \big(I[X,Y] + \lambda \bar{D}[X,Y] \big)$, where $\bar{D} = D/k_B T$ is the distortion measured in units of $k_B T$. The solution to this problem lies on the rate-distortion curve, $R(\bar{D})$, as we can see from comparison with the optimization problem in Eq. (1). This shows that the rate-distortion curve is a least effort function.

This finding is similar, but not identical to the formal mapping of the rate-distortion function onto free energy minimization in multiphase chemical equilibrium [3], and to the statements in [19, 20], where large deviations theory was used to show a formal analogy between the rate-distortion function and the free energy of a chain of particles, i.e. the minimum amount of work needed to compress the chain. These formal analogies are based on identifying the distortion function with physical aspects of a corresponding system, e.g. its energy. It was pointed out in [20] that these formal analogies have some interpretational freedom. Specifically, the interpretation of the Lagrange multiplier that effectively controls the trade-off between distortion and compression depends on the details of the analogy. In the mechanical analogy, it can be interpreted either as inverse temperature [19], or as a conjugate force [20]. In contrast, the derivation given above retains explicitly the distortion constraint and shows that physical temperature adjusts the units by rescaling the distortion measure, or, alternatively, by rescaling the trade-off parameter.

[5] Written in the notation used here, the effort in [18] was quantified by $\lambda H[X|Y] + (1-\lambda) H[Y]$, where the parameter λ weights how much listener and speaker contribute to the total effort. $H[Y] = -\langle \log[p(y)] \rangle_{p(y)}$ denotes Shannon entropy, and $H[X|Y] = -\langle \log[p(x|y)] \rangle_{p(x,y)}$ conditional entropy.

5. CHANNEL CAPACITY

The output messages y can also be interpreted as measurement outcomes. If the measurement is useful, then the observer learns something about x when given y. In the absence of y, the observer's best guess about x is expressed by the prior probability $p(x)$, but when the measurement is received, this changes to the posterior distribution $p(x|y) = p(y|x)p(x)/p(y)$ (Bayes' rule) [21]. Changing of the observer's knowledge state from prior to posterior comes at a cost; it takes a certain amount of effort to implement this change. By the same arguments as above, the minimum amount of work that has to be done (on average) is given by the free energy difference $\langle F[p(x|y)]\rangle_{p(y)} - F[p(x)]$.

This quantity also determines the maximum amount of work that can be *extracted* from a physical system which is (partially) described by x, by exploiting knowledge of y. By convention, energy flowing *into* the system is positive, while energy flowing *out* of the system has a negative sign. Hence, at most $F[p(x)] - \langle F[p(x|y)]\rangle_{p(y)}$ can be extracted as work. Assuming once again no change in average energy, i.e. $\langle E(x)\rangle_{p(x)} = \langle E_y(x)\rangle_{p(x|y)p(y)}$, we have

$$F[p(x)] - \langle F[p(x|y)]\rangle_{p(y)} = -k_B T I[X,Y] \,. \qquad (10)$$

Therefore, maximization of extractable work motivates maximization of mutual information.[6]

A simple example in which the condition $\langle E(x)\rangle_{p(x)} = \langle E_y(x)\rangle_{p(x|y)p(y)}$ holds is that of measuring the x-position of a particle in a box connected to a heat bath at temperature T. Let the length of the box be L. Then $p(x) = 1/L$ inside the box and zero outside. The energy of the particle does not depend on its x-position within the box, where it is given by the particle's kinetic energy, E_K, but the walls of the box pose an infinite energy barrier. Thus we may write:

$$E(x) = \begin{cases} E_K & \forall x \in [0,L] \\ \infty & \forall x \notin [0,L] \end{cases} \,. \qquad (11)$$

The average energy is $\langle E(x)\rangle_{p(x)} = E_K$.

Knowing that the particle is confined e.g. to the left side of the box results in a posterior of $p(x|y = \text{“LEFT”}) = 2/L$ for x between 0 and $L/2$, and zero outside that range (similarly for $y =$ “RIGHT”). This distribution describes a particle in a box of half of the length, but otherwise the same as the original box. The particle's energy is then E_K inside the range of the smaller box, and infinite outside that range:

$$E_{y=\text{“LEFT”}}(x) = \begin{cases} E_K & \forall x \in [0,L/2] \\ \infty & \forall x \notin [0,L/2] \end{cases} \,, \qquad (12)$$

and

$$E_{y=\text{“RIGHT”}}(x) = \begin{cases} E_K & \forall x \in (L/2,L] \\ \infty & \forall x \notin (L/2,L] \end{cases} \,, \qquad (13)$$

with an expected value of $\langle E_y(x)\rangle_{p(x|y)p(y)} = E_K$.[7]

Now, assume that the distribution $p(y|x)$, which describes the data representation method, or, alternatively, the measurement apparatus, be fixed. Then ask for the physical system that best matches the measurement apparatus in the sense that it allows for maximum work extraction, given the measurement (on average). Eq. (10) tells us that the answer is given by Shannon's channel capacity:

$$C = \max_{p(x)} I[X,Y] \,. \qquad (14)$$

For a fixed channel, one chooses the source which would allow for exploiting the knowledge obtained from the messages y towards maximum work extraction.

These are two sides of a coin: communicating more information costs more effort, but the more informative a measurement is about the state of a physical system, the more work that can be extracted from the system given the measurement outcome.

6. LEAST EFFORT MAXIMUM WORK EXTRACTION

Imagine two correlated systems, $\mathcal{X}$ and $\mathcal{Z}$ with mutual information $I[X,Z]$. An observer measures x, and represents it by y, which is obtained with probability $p(y|x)$. This representation, or measurement, can then be used to extract work from system $\mathcal{Z}$.

Knowledge of system $\mathcal{Z}$, given y, is expressed by the probability distribution $p(z|y)$. By the same arguments as above, the maximum amount of extractable work (averaged over all measurements) is given by the free energy difference $F[p(z)] - \langle F[p(z|y)]\rangle_{p(y)}$. Under the assumption that the average energy does not change, this is given by $-k_B T I[Y,Z]$.

The least effort data representation method which maximizes extractable work thus solves

$$\min_{p(y|x)} \left(I[X,Y] - \alpha I[Y,Z] \right) \,, \qquad (15)$$

The Lagrange multiplier α controls the trade-off between work extractable from system $\mathcal{Z}$ (which one wants to maximize) and necessary effort to represent system $\mathcal{X}$ (which one wants to minimize). We recognize Eq. (15) as the *Information Bottleneck* (IB) method [22], hereby lending IB a new thermodynamic motivation: it finds the least effort representation of system $\mathcal{X}$ that allows for maximum work extraction from a correlated system $\mathcal{Z}$.

If $\mathcal{X}$ and $\mathcal{Z}$ are kept at two different temperatures, T_X and T_Z, then α can be interpreted as the ratio T_Z/T_X: the larger the temperature difference, the more beneficial it is to keep relevant information, as it can be traded off for more extractable work.

7. SUMMARY

Least effort coding leads to data representations that lie on the rate distortion curve. Least effort is measured by

[6] Be reminded of the sign. $I[X,Y]$ is a non-negative quantity. Extracted work comes with a negative sign. Thus, more work can be extracted when $I[X,Y]$ is larger.

[7] This holds for all $p(y)$, because $\langle E_y(x)\rangle_{p(x|y)} = E_K$, which is independent of y.

the average free energy difference, quantifying the least amount of physical work necessary to change a system described by the average output distribution to one described by the specific output distribution necessary to produce output messages when the input is known. Information loss has to do with thermodynamic efficiency: least effort is proportional to mutual information (under the assumption that there is no average energy change). Codes that are efficient in a rate-distortion sense are also energetically efficient. In that sense, lossy compression is lazy compression, because it minimizes physical effort.

Channel capacity, on the other hand, represents the maximum amount of work that could be extracted from a source (on average) given the channel's output messages. In contrast to the above, where the source is fixed and the optimization is over encoding schemes, here the channel is given. The maximum is then taken over all possible sources, thus optimizing over physical systems for the best match in terms of possible work extraction.

The Information Bottleneck method provides the means of finding a minimum effort compression (or data representation) that allows for maximum work extraction from another system by exploiting correlations.

8. ACKNOWLEDGEMENTS

I am grateful for support from the Foundational Questions Institute (FQXi). I thank Arne Grimsmo and Rob Shaw for inspiring discussions that initiated this research, and Toby Berger, Antonio Celani, Gavin Crooks, David Sivak and Elan Stopnitzky for extremely valuable comments on the manuscript.

9. REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[2] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec*, vol. 4, no. 142-163, pp. 1, 1959.

[3] T. Berger, "Rate distortion theory: A mathematical basis for data compression," 1971.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, 2nd edition, 2006.

[5] F. Schlögl, "On stability of steady states," *Zeitschrift für Physik*, vol. 243, no. 4, pp. 303–310, 1971.

[6] J. Schnakenberg, "Network theory of microscopic and macroscopic behavior of master equation systems," *Reviews of Modern physics*, vol. 48, no. 4, pp. 571, 1976.

[7] R. Shaw, *The dripping faucet as a model chaotic system*, Aerial Press, 1984.

[8] B. Gaveau and L. S. Schulman, "A general framework for non-equilibrium phenomena: The master equation and its formal consequences," *Phys. Lett. A*, vol. 229, no. 6, pp. 347–353, 1997.

[9] H. Qian, "Relative Entropy: Free Energy Associated with Equilibrium Fluctuations and Nonequilibrium Deviations," *Phys. Rev. E*, vol. 63, pp. 042103, 2001.

[10] G. E. Crooks, "Beyond Boltzmann-Gibbs statistics: Maximum entropy hyperensembles out-of-equilibrium," *Phys. Rev. E*, vol. 75, pp. 041119, 2007.

[11] M. Esposito and C. Van den Broeck, "Second law and landauer principle far from equilibrium," *EPL (Europhysics Letters)*, vol. 95, no. 4, pp. 40004, 2011.

[12] S. Still, D. A. Sivak, A. J. Bell, and G. E. Crooks, "Thermodynamics of prediction," *Physical Review Letters*, vol. 109, pp. 120604, 2012.

[13] S. Deffner and C. Jarzynski, "Information processing and the second law of thermodynamics: An inclusive, hamiltonian approach," *Phys. Rev. X*, vol. 3, pp. 041003, Oct 2013.

[14] B. Gaveau, L. Granger, M. Moreau, and L. S. Schulman, "Relative entropy, interaction energy and the nature of dissipation," *Entropy*, vol. 16, no. 6, pp. 3173–3206, 2014.

[15] L. Szilard, "On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings," *Z. Phys.*, vol. 53, pp. 840–856, 1929.

[16] J. V. Koski, V. F. Maisi, T. Sagawa, and J. P. Pekola, "Experimental study of mutual information in a Maxwell Demon," *arXiv preprint arXiv:1405.1272*, 2014.

[17] G. K. Zipf, *Human Behavior and the Principle of Least Effort*, Addison-Wesley, Cambridge, MA, 1949.

[18] R. F. i Cancho and R. V. Solé, "Least effort and the origins of scaling in human language," *Proceedings of the National Academy of Sciences*, vol. 100, no. 3, pp. 788–791, 2003.

[19] N. Merhav, "An identity of chernoff bounds with an interpretation in statistical physics and applications in information theory," *Information Theory, IEEE Transactions on*, vol. 54, no. 8, pp. 3710–3721, 2008.

[20] N. Merhav, "Another look at the physics of large deviations with application to rate-distortion theory," *arXiv:0908.3562*, 2009.

[21] H. Jeffreys, *Theory of Probability*, Oxford University Press, third edition, 1998.

[22] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annual Allerton Conference*, B. Hajek and R. S. Sreenivas, Eds. 1999, pp. 368–377, University of Illinois, Available at http://xxx.arXiv.cornell.edu/abs/physics/0004057.