A geometric approach to information processing

Gabriel Jurado*
Department of Physics, College of Arts and Science, Florida State University, Tallahassee FL, 32310, USA. and
Condensed Matter Theory, National High Magnetic Field Laboratory (NHMFL), Florida State University, Tallahassee, Florida, 32310, USA.
(Dated: March 4, 2021)

Information has a natural thermodynamic interpretation in terms of the amount of energy dissipated during the erasure of a bit of information; this places a bound of $kTln(2)$ on the efficiency of information processing operations. Biological systems process information in order to build models of their environment. The efficiency of information processing in biological systems serves as a model for information processing in computational systems.

I. INTRODUCTION

Rolf Landuaer based his physics research on a simple rule: information is physical. That is, information is registered by physical systems such as strands of DNA, neurons and transistors; in turn, the ways in which systems such as cells, brains and computers can process information is governed by the laws of physics. Virtually every organism gathers information about its noisy environment and uses it to build models that dictate its behavior. Information processing is ubiquitous in biological systems, from single cells measuring external concentration gradients to large neural networks performing complex control operations. Landauer was able to characterize two types of information processing; logically reversible and logically irreversible operations. He then showed that logically irreverisble operations necessarily require the dissipation of energy; erasure transforms information from an accessible form to an inaccessible form, from memory to entropy, respectively. Landauer's principle transformed the physics of information by identifying the basic trade-off between information and physical thermodynamic quantities.

The essence of learning is to be able to reproduce the results of the teacher without prior experience and typically with incomplete information about the required task. In biological systems, learning is the construction of a model environment from external sensory data of the environment. If one considers learning as the particular state of information processing that remains stable and efficient despite a noisy environment, then one can begin to use the thermodynamic properties of information to construct a framework that investigates the physical nature of learning. The study of energetic efficiency of information processing in biological systems relies heavily on the emerging framework of stochastic thermodynamics. It is an integrated framework to study the relation between information processing and dissipation in interacting systems far from equilibrium and has been used to study the thermodynamic properties of learning.

* juradogabriel93@gmail.com

II. THERMODYNAMICS OF MACHINE LEARNING

The recent work of Alemi and Fischer [1] can be considered a step in this direction in the context of machine learning. They introduce a formal correspondence between the model framework of representation learning and thermodynamics by considering the mapping between a real-world data distribution, the learned model, and its corresponding stochastic representation, the noisy environment. The relative information between two distributions p and q,

$$KL(p||q) = \int dx p(x)\ log\frac{p(x)}{q(x)} = \left\langle log\frac{p(x)}{q(x)} \right\rangle$$

measures how much information is gained when a measurement reveals an element in world P, when an element in world Q was expected. In addition, defining the mutual information between the joint distribution for X and Y, and the product of their marginal distributions

$$I(X;Y) \equiv \left\langle log\frac{p(x,y)}{p(x)p(y)} \right\rangle$$

allows one to quantify how much information is gained about X by observing Y, or vice versa.

Typically, a vanishing KL divergence can only be achieved in theory, $KL(p||q) = 0$, and corresponds to the ideal scenario where the representation model constructed from the unknown data-generating model, is exactly able to predict the unknown distribution . In practice, the relative entropy between world P and world Q will always be non-zero, $KL(P||Q) \neq 0$. This is due to the inherent complexity of the unknown model and stems from the intrinsic entropy of the distribution, which cannot be controlled [1].

Thus, the optimal value for the KL divergence between the two worlds is found by the difference in multi-informations of each model, largely known as information projection [5]. In the present context, the minimal relative information is,

$$KL(P||Q) = minKL(p||q) = I_P - I_Q$$
$$= I(\Theta;\{X,Y\}) + \textstyle\sum_i I(X_i;\Phi) + I(Y_i;X_i,\Phi)$$
$$+ \textstyle\sum_i I(Z_i;X_i,\Theta) - I(X_i;Z_i) - I(Y_i;Z_i)$$

where $\{X, Y\}$ is a set of paired data, e.g., X is the raw high-dimensional data and Y is the corresponding low-dimensional label, and Φ is the governing parameter for the unknown joint distribution $p(\{x, y\}, \phi)$. Similarly, Z is a stochastic representation of X, defined by some arbitrary parametric distribution with parameters Θ, such as, $p(z_i|x_i, \theta)$.

The terms $I(X_i; \Phi)$ and $I(Y_i; X_i, \Phi)$ represent the intrinsic complexity of the data distribution and are outside the control of the representation model. These terms can be set to constants without loss of generality. The remaining terms are described below;

$I(X_i; Z_i)$ - measures how much information the representation Z contains about the input X,

$I(Y_i; Z_i)$ - measures how much information the representation Z contains about the input labels Y,

$I(Z_i; X_i, \Theta)$ - measures how much information the parameters Θ and input X determine about the representation Z,

$I(\Theta; \{X, Y\})$ - measures how much information is stored in the parameters Θ from the training set $\{X, Y\}$.

A. Functionals

Through an appropriate choice of functional, they were able to define variational bounds for the set of mutual informations.

$R \equiv \sum_i \left\langle log \frac{p(z_i|x_i, \theta)}{q(z_i)} \right\rangle \geq \sum_i I(Z_i; X_i, \Theta)$

The rate measures the complexity of the representation and measures how many bits or nats are encoded for each sample

$C \equiv -\sum_i \langle log\ q(y_i|z_i) \rangle \geq \sum_i H(Y_i) - I(Y_i; Z_i)$

The classification error measures how much information in Y is left unspecified in the representation Z.

$D \equiv -\sum_i \langle log\ q(x_i|z_i) \rangle \geq \sum_i H(X_i) - I(X_i; Z_i)$

The distortion measures how much information about X is left unspecified in the representation Z.

$S \equiv \left\langle log \frac{p(\theta|\{x, y\})}{q(\theta)} \right\rangle \geq I(\Theta; \{X, Y\})$

The entropy in the parameters measures the relative information between the distribution assigned to the parameters in world P, after learning from the data {X,Y}, relative to a data independent prior over the parameters $q(\theta)$. This is an upper bound on the mutual information between data and model parameters and measures the risk of overfitting.

In a previous paper [2], the authors argue that a better way to quantify the value of representation learning is to measure the mutual information between the observed and the latent, X and Z, respectively. By parameterizing the mutual information they show that varying the mutual information provides control over how much of the data is compressed in the latent representation, relative to how much is retained in the original representation. This relationship is visualized using a rate-distortion or RD curve, FIG. 1 [2].

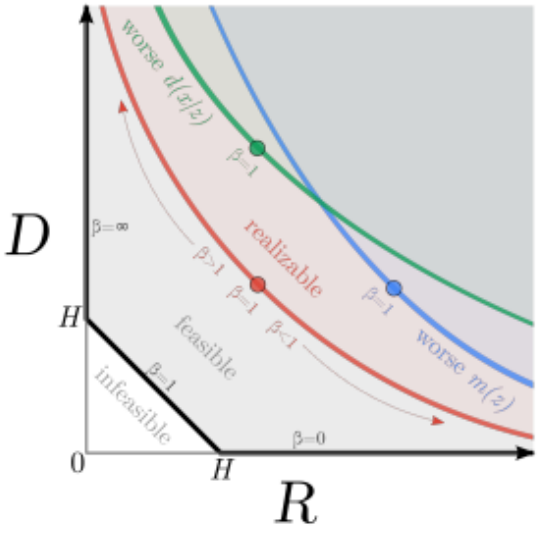

FIG. 1. Phase diagram in RD-plane. The thick black lines denote the feasible boundary in the idealized regime of infinite model capacity. $\beta = \frac{\partial D}{\partial R}$ and $d(x|z)$ and $m(z)$ are variational approximations, namely, the decoder and marginal, respectively.

The $R = 0$, $D = 0$, and $D = H - R$ lines all represent ideal bounds for the values of mutual information. As mentioned before, the KL divergence will always be non-zero for finite models, and any failure to achieve a vanishing divergence will result in optimal surfaces only asymptotically approaching the ideal limit.

B. Thermodynamics

By generalizing the framework that describes surfaces of optimal bounds between mutual informations, to include the four functionals R, C, D, and S, a formal correspondence to thermodynamics can be established. Letting the optimal surface satisfy some equation $f(R, C, D, S) = 0$, and noting that the optimal surface contains an equivalence class of states, i.e., the set of states minimizing the diffrenence introduced by projecting world P onto world Q, one can express any functional as a function of the others, e.g., $R = R_0(C, D, S)$. Since the function is smooth and convex, it allows for partial

derivatives to be expressed by differentials of the functionals

$$dR = \Big(\frac{\partial R}{\partial C}\Big)_{D,S} dC + \Big(\frac{\partial R}{\partial D}\Big)_{C,S} dD + \Big(\frac{\partial R}{\partial S}\Big)_{C,D} dS.$$

When the partials are defined as;

$$\gamma \equiv -\Big(\frac{\partial R}{\partial C}\Big)_{D,S}$$

$$\delta \equiv -\Big(\frac{\partial R}{\partial D}\Big)_{C,S}$$

$$\sigma \equiv -\Big(\frac{\partial R}{\partial S}\Big)_{C,D}$$

the total differential for the optimal surface can be expressed in a form analogous to the *First Law of Thermodynamics*

$$dR = -\gamma dC - \delta dD - \sigma dS$$

where the partials γ, δ, and σ, are analogous to Lagrange multipliers which impose constraints [1]. Continuing to push the thermodynamic analogy, the functionals R, C, D, and S become the systems extensive thermodynamic variables, e.g., volume, entropy, particle number, and energy, which change as the system changes. On the other hand, the partial derivatives γ, δ, and σ represent the intensive variables which are generalized thermodynamic forces of corresponding state variables, such as pressure, temperature, chemical potential, and tension.

The extensive functionals and corresponding differentials are well defined for any system state however, the intensive variables and corresponding partial derivatives are only well defined for equilibrium states. For example, states contained on the optimal surface.

Considering again the differential $dR = -\gamma dC - \delta dD - \sigma dS$, an intuitively non-trivial result emerges from the commutivity of mixed-partials

$$\Big(\frac{\partial^2 R}{\partial D \partial C}\Big) = \Big(\frac{\partial^2 R}{\partial C \partial D}\Big) \implies \Big(\frac{\partial \delta}{\partial C}\Big)_D = \Big(\frac{\partial \gamma}{\partial D}\Big)_C.$$

This expression equates the change in the derivative of the $R-D$ curve (δ) as a function of the classification error C for fixed distortion D, to the change in derivative of the $R-C$ curve (γ) as a function of the distortion (D) at fixed classification error C.

Just as naturally, one can introduce alternative thermodynamic potentials by preforming additional Legendre transformations. In particular, the *free rate*

$$F(C, D, \sigma) \equiv R - \sigma S$$
$$dF = -\gamma dC - \delta dD - S d\sigma$$

measures the rate of the system in terms of the intrinsic parameter σ and results in the following equality

$$\Big(\frac{\partial S}{\partial C}\Big)_\sigma = \Big(\frac{\partial \gamma}{\partial \sigma}\Big)_C.$$

In addition, taking second-order differentials produces appropriate quantities analogous to thermodynamic second-order differentials. For instance, the analog of heat capacity becomes the rate capacity

$$K_D \equiv \Big(\frac{\partial R}{\partial \sigma}\Big).$$

C. Equilibrium

In the framework constructed, the points on the optimal surface are analogous to equilibrium states, with well defined derivatives. When any subpart of a system is in thermal equilibrium with any other subpart, the system can be defined to be in an equilibrium state. Using The Second Law of Thermodynamics as motivation, a distinction between states on the optimal frontier and those off of it can be made, analogous to the distinction between equilibrium states and non-equilibrium states in thermodynamics.

Any equilibrium distribution can be expressed by a Boltzmann distribution, that is, the solution to an optimization problem subject to some constraints. For example, the entropy term S measures the relative entropy of the parameter distribution in the representation with respect to some unknown prior $p(\theta)$. Thus, the form of the distribution as a function of the functionals and corresponding partials is

$$p_0(\theta|\{x, y\}) = \frac{q(\theta)}{Z} e^{-(R+\delta D+\gamma C)/\sigma}$$

where Z is the normalization constant or partition function of the distribution

$$Z = \int d\theta q(\theta) e^{-(R+\delta D+\gamma C)/\sigma}.$$

The term in the exponential defined as

$$J(\gamma, \delta, \sigma) \equiv R + \delta D + \gamma C + \sigma S$$

takes the form

$$minJ(\gamma, \delta, \sigma) = J_0(\gamma, \delta, \sigma) = -\sigma log Z(\gamma, \delta, \sigma)$$

for an equilibrium distribution. Having a well defined equilibrium serves as a necessary reference state for the relative measurements of non-equilibrium states. Thus, the KL divergence between some non-equilibrium distribution and the equilibrium distribution is

$$KL(p(\theta)||p_0(\theta)) = \frac{\Delta J}{\sigma}$$

where

$$\Delta J \equiv J^{noneq} - J_0$$

and

$$J^{noneq} = \langle R + \delta D + \gamma C + \sigma S \rangle_{p(\theta)}$$

It is here that The Second Law of Thermodynamics can be formulated within the context of a Markov process. For a stationary Markov process whose stationary distribution is the equilibrium distribution, the KL divergence relative to equilibrium must decrease monotonically. By letting the non-equilibrium distribution define the initial value with the prior $q(\theta)$

$$J_{t=0} = \langle R + \delta D + \gamma C \rangle_{q(\theta)}$$

and the equilibrium distribution be the final state value, implicitly weighted by $q(\theta)$

$$J_{t=\infty} = -\sigma log Z$$

we have the following *Law of Decreasing Relative Entropy*

$$J_{t=0} \geq J_t \geq J_{t+1} \geq J_{t=\infty}.$$

The aforementioned analogies to thermodynamics provide the groundwork to develop new quantitative measures and relationships among models within the learning framework. The architecture constructed begins with the **five**-dimensional space of distributions $p(z|x,\theta), p(\theta|\{x,y\}), q(z), q(x|z), q(y|z)$. Once chosen these distributions define a **four**-dimensional space of optimal states given by the functionals R, C, D, S. The relationships between the functionals provides an optimal **three**-dimensional surface represented by δ, γ, σ. A single learning objective that targets points on the optimal surface expresses a large spectrum of learning objectives. This versatility is indicative of various identities and symmetries that may exist on the geometry of the optimal frontier. Is is from this geometry that the direct analogy to thermodynamic relations emerges, independent of any physical motivation [3].

III. GEOMETRY OF THERMODYNAMICS

A. Thermodynamic metric

The idea of introducing geometry into thermodynamic states is rooted as far back as Gibbs original work. Although, a formal correspondence of this idea was not presented until the work of Weinhold brought together the principle empirical laws of equilibrium thermodynamics, with the mathematical axioms of an abstract metric space [9]. This correspondence endowed the thermodynamic formalism with a geometric interpretation. First, consider the key empirical observations that an equilibrium thermodynamic theory must incorporate: (i) the observation that properties of an equilibrium system may be associated with low-order derivatives of a mathematical function U

$$R_i = \frac{\partial U}{\partial X_i}$$

(ii) the observation that the internal energy function U satisfies the requirement for an exact differential

$$\frac{\partial R_i}{\partial X_j} = \frac{\partial R_j}{\partial X_i}$$

(iii) and the observation that the internal energy is minimized in an isolated equilibrium state

$$\frac{\partial R_i}{\partial X_i} \geq 0$$

Then, formally associate each field differential

$$dR_i$$

with an abstract vector

$$dR_i = |R_i\rangle$$

and scalar product.

$$\langle R_i | R_j \rangle \equiv \frac{\partial R_i}{\partial X_j}$$

In this way, the key mathematical requirement for an abstract metric space to form a scalar product with the following properties

i. $\langle R_i | \lambda R_j + \mu R_k \rangle = \lambda \langle R_i | R_j \rangle + \mu \langle R_i | R_j \rangle$

ii. $\langle R_i | R_j \rangle = \langle R_j | R_i \rangle$

iii. $\langle R_i | R_j \rangle \geq 0$ (equality only if $|R_i\rangle = 0$)

could be readily realized in the space of thermodynamic states. Such a representation is mathematically isomorphic to Euclidean space and naturally invites the use of vector and matrix methods in lieu of partial differential equations.

However, Weinhold's construction is formally lacking a measure of distance and the ambiguous choice of the inner product does not present itself as a physically relevant quantity. In order to build a more general geometry in thermodynamic state space, Ruppeiner introduced the Riemannian geometric model of thermodynamics [6], by including the theory of fluctuations into the thermodynamic axioms. He focused on the curvature of these manifolds and found that the curvature for a manifold representing an ideal gas is zero. Consequently, interactions represent themselves in the Riemannian geometric model through curvature which is independent of any statistical mechanical model. Measuring interactions through curvature requires only thermodynamic quantities for input, is rooted in fluctuations, and provides an effective measure of interaction strength for each state of the thermodynamic system [3].

In order to introduce a metric structure compatible with the convexity of thermodynamic surfaces, Gilmore introduced a metric on potential surfaces, as opposed to the equation-of-state surface, which was the approach

taken by Weinhold and Ruppeiner. In this construct, a thermodynamic potential is introduced by

$$U(x^i, x_0^i) = U(x^i) - \left.\frac{\partial U}{\partial x^i}\right|_0 (x^i - x_0^i)$$

where x^i is a set of extensities and x_0^i the corresponding equilibrium values. The partial

$$\frac{\partial U}{\partial x^i} = \lambda_0^i$$

is the set of conjugate intensities at thermodynamic equilibrium. This is a naturally occurring potential and describes the probability distribution for fluctuations around equilibrium, i.e., $p \sim e^{-\beta U}$. The distance induced can have various interpretations. It can be thought of as the number of steps between two statistically distinct states, as the number of fluctuations needed to trace the length of a path, and the square-root of the minimum dissipation multiplied by the number of relaxations in some finite-time process [3].

B. Thermodynamic length

Special interest was given to the entropy form D^2S of the metric derived by Ruppeiner for its potential towards extending the notion of thermodynamic length into statistical systems and information theory [6]. The question was whether the probabilistic definition for entropy, i.e., maximum entropy, could be used to induce an equivalent metric as that of thermodynamic entropy D^2S. A proof by Salamon, Nulton, and Berry demonstrated that under suitable conditions, the probabilistic entropy S_p gives a length in the space of probability distributions equal to the length calculated with the thermodynamic entropy S_x [8]. The thermodynamic length can be evaluated from time-varying probability distributions as a result of its intimate connection with dissipation and irreversibility within finite-time processes [7]. Considering the equality of the lengths induced by either form of entropy, and the subtle relationship between this length and macroscopic, thermodynamic processes, e.g., dissipation, suggests the existence of an analogous macroscopic process quantifying *information dissipation* and *information irreversibility*.

The notion of distance in thermodynamic state space does not clearly present itself as a readily measurable quantity. Nearly 30 years after the introduction of thermodynamic geometry, Crooks showed how thermodynamic length for a system defined by equilibrium statistical mechanics could be measured with a computer simulation [4]. As a consequence, several non-trivial relations between thermodynamic length and information theory of non-equilibrium systems were revealed. Consider a conditional Gibbsian distribution of the form

$$p(x|\lambda) = \frac{1}{Z} e^{\beta H(x,\lambda)}$$

where β is inverse temperature, Z is the partition function, and H is the Hamiltonian. The Hamiltonian can be split into a set of collective variables X_i and conjugate generalized forces λ^i, such that,

$$p(x|\lambda) = \frac{1}{Z} e^{-\lambda^i(t) X_i(x)}.$$

The partition function that normalizes the probability distribution is directly related to the free energy F, the free entropy Ψ, and the entropy S [4]

$$lnZ = -\beta F = \Psi = S - \lambda^i \langle X_i \rangle .$$

The first derivatives of the free entropy give the first-order moments of the collective variables

$$\frac{\partial \Psi}{\partial \lambda^i} = -\langle X_i \rangle$$

and the second derivatives give the covariance matrix of the collective variables

$$g_{ij} = \frac{\partial^2 \Psi}{\partial \lambda^i \partial \lambda^j} = -\frac{\partial \langle X_i \rangle}{\partial \lambda^j}$$

$$= \langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \rangle .$$

Because the covariance matrix is positive semi-definite and varies smoothly from point to point, it can be used as a metric tensor to naturally equip the manifold of thermodynamic states with a Riemannian metric. Consequently, the Riemannian metric allows the distance along curves connecting different points to be measured [4]. For example, suppose the length of a curve is parametrized by t, from 0 to τ as

$$L = \int_0^\tau \sqrt{\frac{d\lambda^i}{dt} g_{ij} \frac{d\lambda^j}{dt}}\, dt$$

then, distance between two points is the length of the shortest curve, i.e., the geodesic. By introducing the thermodynamic divergence of the path

$$J = \tau \int_0^\tau \frac{d\lambda^i}{dt} g_{ij} \frac{d\lambda^j}{dt} dt$$

and using the Cauchy-Schwarz inequality

$$\int_0^\tau f^2 dt \int_0^\tau g^2 dt \geq \Big[\int_0^\tau f g dt \Big]^2$$

between the length and divergence results in the inequality

$$J \geq L^2.$$

Thermodynamic length and divergence are the fundamental quantities controlling dissipation of finite-time thermodynamic transformations that maintain an internal equilibrium [7]. Consequently, by measuring and optimizing for the thermodynamic length the minimal dissipation path connecting two thermodynamic states can

be found. The thermodynamic length and dissipation are related through

$$lim_{N\to\infty} N \sum_{t=1}^{N-1} \sqrt{\Delta\lambda^i \Delta \langle X_i \rangle} = L$$

where the total average dissipation is given by

$$w = \sum_{t=1}^{N-1} \Delta\lambda^i \Delta \langle X_i \rangle .$$

Furthermore, the connection between thermodynamic length and dissipation suggests a potential scheme for measurement. By computing equilibrium simulations at a series of points along the thermodynamic length then examining the scaling of dissipation with the number of steps [4]. The dissipation can be measured as the maximum likelihood of the free entropy change and Jensen-Shannon divergence between adjacent samples. The Jensen-Shannon divergence is the mean of the relative entropy of each distribution relative to the mean distribution,

$$JS(p;q) = \frac{1}{2}\sum_i p_i ln \frac{p_i}{1/2(p_i+q_i)} + \frac{1}{2}\sum_i q_i ln \frac{q_i}{1/2(p_i+q_i)}$$

such that, the maximum likelihood free free entropy change is given by

$$l(\Delta\Psi_{12}) \simeq 2K(JS(p^1;p^2) - ln2).$$

The total Jensen-Shannon metric along the path defines a lower bound for both the thermodynamic length and divergence

$$L = \sqrt{8}\int d\sqrt{JS}$$

and

$$J = 8\int dJS,$$

respectively [4]. Measurements are then repeated for smaller and smaller discretizations of the path until the thermodynamic length and divergence converge.

[1] Alexander A. Alemi and Ian Fischer. TherML: Thermodynamics of Machine Learning. 2018.

[2] Alexander A. Alemi, Ben Poole, Ian Fischer, Joshua V. Dillon, Rif A. Saurous, and Kevin Murphy. Fixing a Broken ELBO. 2017.

[3] Bjarne Andresen, R. Stephen Berry, Robert Gilmore, Ed Ihrig, and Peter Salamon. Thermodynamic geometry and the metrics of Weinhold and Gilmore. *Physical Review A*, 1988.

[4] Gavin E. Crooks. Measuring thermodynamic length. *Physical Review Letters*, 2007.

[5] Imre Csiszár and František Matúš. Information projections revisited. *IEEE Transactions on Information Theory*, 2003.

[6] George Ruppeiner. Thermodynamics: A Riemannian geometric model. *Physical Review A*, 1979.

[7] P. Salamon, J. Nulton, and E. Ihrig. On the relation between entropy and energy versions of thermodynamic length. *The Journal of Chemical Physics*, 1984.

[8] Peter Salamon, James D. Nulton, and R. Stephen Berry. Length in statistical thermodynamics. *The Journal of Chemical Physics*, 1985.

[9] F. Weinhold. Metric geometry of equilibrium thermodynamics. *The Journal of Chemical Physics*, 1975.