From Collective Inference to Gravity and Strings

Jonathan J. Heckman

Jefferson Physical Laboratory, Harvard University, Cambridge, MA 02138, USA

jheckman@physics.harvard.edu

Abstract

This essay is a less technical account of recent results announced by the author. The starting point for our considerations is a collective of agents which has been exposed to "The It" of external events. These events are used by the collective to construct a family of statistical models which depend on continuous fitting parameters. From this premise, we show how to recover "The It" of classical gravity and a theory of strings from "The Bits" of stable inference schemes by the collective.

1 Introduction

A stream of incoherent data – be it digital or analog – is useless without the further sieve of interpretation. Indeed, information is a useful construct insofar as it can be interpreted by an agent.

From this standpoint, asking whether "it comes from bit" or "bit comes from it" is a false choice: It is akin to asking whether the chicken or egg came first. One cannot build a framework for interpretation without external phenomena, but conversely, such external phenomena are meaningless until they are interpreted.

In this essay we shall therefore take a different tack. Our aim will be to study the interplay between "The It" of experience and "The Bits" of information. The system we study is that of a collection of agents – or collective for short – which seeks to produce a robust probabilistic model of events experienced by the collective.

There is something of a juggling act in coming up with a robust statistical model. We make three competing demands: Accuracy, Simplicity, and Stability. An accurate model is of course desirable, but adding many baroque features clearly diminishes its explanatory power. This is where simplicity, i.e. Occam's razor enters in the list of demands. Finally, the requirement of stability, i.e. the ability to hold on to an inference in the face of "jolts" to the system is also clearly important.

The really big surprise is that these very natural conditions have something to do with "The It" of gravity and string theory [1]. Indeed, by analyzing the conditions for a collective to produce a stable inference scheme, we show first off, that certain geometric arrangements of agents are more stable, and more surprisingly, how the equations of gravity emerge from the condition of stable statistical inference. Further, we show how to generalize these considerations to arrive at a string theory.

Much of the technical content of this essay is contained in the research article of reference [1]. Our aim here will be to offer a more conceptual and informal treatment.

2 Information and Inference

The starting point for our analysis is Bayes' rule [2]. This rule provides a recipe for revising a prior estimate for the odds of an event given new data. In equations, here is the recipe:

$$\Pr(A|B) = \frac{\Pr(A)}{\Pr(B)} \Pr(B|A). \tag{1}$$

The lefthand side $\Pr(A|B)$ tells us the probability of an event A assuming event B has occurred, while $\Pr(B|A)$ tells us the probability of B assuming A has occurred. The probabilities $\Pr(A)$ and $\Pr(B)$ can be read as telling us about the prior beliefs for the odds of A and B, respectively.

Here is an example of updating an inference: If we believe we are drawing a card from a fair deck, finding a queen means our subsequent odds of drawing a queen compared to a king have now gone down. In this and related cases, we can calculate the precise change using Bayes' rule.

From a philosophical standpoint, Bayes' rule provides an attractive starting point for statistical inference since it explicitly acknowledges the notion of a prior belief. On the other hand, this also makes it somewhat "subjective" since we need to give a weight to prior odds of the event occurring. This is sometimes held in contrast with a frequentist approach to statistics, where one assumes a notion of being able to repeat an experiment with arbitrary precision. In some sense, this is simply a specific choice of prior beliefs, but with the sheen of more objectivity. We do not need to get into these fine points here, and so refer the reader to the (Bayesian slanted) reference [3] for further discussion.

Clearly, to update an inference there must be some agent doing the updating. What then shall we view as a viable agent? We shall take a rather broad view that anything which can adjust its behavior in reaction to a change in its environment should count as an agent. Here are some familiar examples of agents:

$$\text{A human;} \quad \text{A monkey;} \quad \text{A tree.} \tag{2}$$

Nothing requires us to select an agent with any semblance of "life-force". Indeed, so long as an entity can react to its environment, we view it as an agent. Here are some less lively examples of agents:

$$\text{A machine which can learn;} \quad \text{An electron;} \quad \text{A D-instanton.} \tag{3}$$

The example of a machine exhibits the idea of training a program or machine on some set of data, and then exposing it to new data. The last example is a theoretical construct from string theory which specifies the location for the endpoint of a string.

Regardless of how we designate an agent, forming a reliable inference strategy is typically a rather complicated process. Faced with this state of affairs, the best an agent will typically be able to do is develop some model of the world. Within this model, the agent can try to find an optimal fit.

To illustrate, suppose that an agent is exposed to a set of N independent events $E = \{e_1, ..., e_N\}$ generated by a probability distribution $p_{\text{true}}(x)$ for x a real number. We shall always take N to be a very large number in our discussion. The agent might have a "Gaussian model of the world":

$$p_{\text{guess}}(x, \{y, \alpha\}) = \frac{1}{\sqrt{2\pi\alpha^2}} \exp\left(-\frac{(x-y)^2}{2\alpha^2}\right) \tag{4}$$

which gives a probability distribution for the event x depending on both the center of the distribution y, and its overall width α. Assuming this "Gaussian worldview", the agent can make a best fit to the parameters y and α.

If the only goal of the agent is to fit the observed data, then it could in principle simply adopt N localized bumps as its model of the world. This is clearly overkill, though, because it requires introducing a new parameter for each new observed event. This example brings out the tension between balancing the competing demands of accuracy and simplicity.

Bayes' rule again points the way forward. If an agent adopts a family of statistical models A which depend on M continuous parameters $\{y\} \equiv \{y^{(1)}, ..., y^{(M)}\}$, the generalization of Bayes' rule now requires us to sum over all the different statistical models obtained from different values of the parameters:

$$\Pr(A|E) = \frac{\Pr(A)}{\Pr(E)} \int d^M y \; w(y) \; \Pr(E|\{y\}). \tag{5}$$

Here, the lefthand side should be viewed as the probability of the family of models given by A assuming prior events E. On the righthand side, $\Pr(A)$ refers to the prior belief for a "model of type A". In the quantity $\Pr(E|\{y\})$ we are considering a particular choice of parameters y for a model of type A, and assigning it weight $w(y)$. The weight $w(y)$ reflects the strength of a prior belief that within the realm of models of type A, a specific choice of parameters is more likely. If the agent has two competing models of the world, say A and B, we can compute $\Pr(A|E)$ and $\Pr(B|E)$. The model with higher posterior probability is clearly more desirable.

Now the really interesting point is that in the limit where the agent has sampled the true distribution a large number of times N, $\Pr(A|E)$ is proportional to a partition function of the sort encountered in statistical mechanics [4, 5]:

$$\Pr(A|E) \propto \frac{\Pr(A)}{\Pr(E)} \int d^M y \; w(y) \; \exp(-N \; \mathcal{E}(y)) \tag{6}$$

where

$$\mathcal{E}(y) = D_{KL}(p_{\text{true}}||p_{\text{guess}}) \equiv \int_X d\mu(x) \;\; p_{\text{true}}(x) \ln \frac{p_{\text{true}}(x)}{p_{\text{guess}}(x; \{y\})}, \tag{7}$$

measures the proximity of the guess to the true distribution. The quantity $\mathcal{E}(y)$ is known as the relative entropy or Kullback-Leibler divergence. It has the natural interpretation as quantifying the amount of information in the sense of Shannon [6] an agent would gain by adjusting its original guess to the true distribution. Indeed, for a very good guess, $D_{KL}(p_{\text{true}}||p_{\text{guess}})$ is small. By contrast, a poor guess will have a very large value of $D_{KL}(p_{\text{true}}||p_{\text{guess}})$.

Here, the information content is measured in "nats" rather than "bits". This is indeed the natural choice to make when dealing with continuous probability distributions. An important point is that in the continuous context, the relative entropy is actually better behaved than the continuum version of the Shannon entropy, which is known as the differential entropy [6]:

$$S[p] = -\int_X d\mu(x) \; p(x) \ln p(x). \tag{8}$$

This can be either positive or negative, but by contrast, D_{KL} is always positive.

What makes equation (6) so striking is that out of almost nothing, we have arrived at the physically rich situation of a statistical mechanical model, with the number of sampled events playing the role of an inverse temperature $N \sim 1/T$. In the language of statistical mechanics, the energy corresponds to how far from the true distribution the guess happens to be. Minimizing this quantity is clearly desirable. On the other hand, if we have a statistical model A which only attains a good fit for a narrow set of parameters y, we might deem the model rather contrived. This is where the ability to explore many nearby models comes in. In the language of statistical mechanics, Occam's razor, i.e. the overall simplicity of the model translates to maximizing the entropy [4, 5]. The balancing act between minimizing energy and maximizing entropy in statistical mechanics corresponds to the juggling of accuracy and simplicity.

Something even more striking happens when we restrict to guesses which are nearby the correct one. Expanding in small fitting parameters:

$$p_{\text{true}} = p_{\text{guess}} + \delta y^I \frac{\partial p_{\text{guess}}}{\partial y^I}, \tag{9}$$

one arrives at an infinitesimal line element $ds_{\text{parameters}}$ on the space of fitting parameters:

$$\mathcal{E}(y) \simeq ds^2_{\text{parameters}} = G_{IJ}\ \delta y^I \delta y^J, \tag{10}$$

where G_{IJ} is known as the information metric:

$$G_{IJ} = \int_X d\mu(x)\ \ p_{\text{guess}}(x;\{y\}) \frac{\partial \ln p_{\text{guess}}(x;\{y\})}{\partial y^I} \frac{\partial \ln p_{\text{guess}}(x;\{y\})}{\partial y^J}. \tag{11}$$

In statistics, this is commonly referred to as the Fisher score for the statistics. Geometrically, it defines a notion of infinitesimal distance between nearby statistical models. For further details on the interplay between information theory and differential geometry, we refer the interested reader to reference [7].

In other words, the behavior of the agent as it updates its fitting parameters can be visualized as traversing a minimal length arc, with distance dictated by the information metric G_{IJ}. This is our first hint that inference has something to do with "Bits of It".

3 Collective Inference

But even if our agent is a genius, it is quite likely that it will fail to uncover the exact form of the true distribution $p_{\text{true}}(x)$. Indeed, once an agent picks a model of the world, it is stuck with it, and has to live with the consequences. By the same token, a collection of agents working together might arrive at a better picture of what is going on compared with a single agent. Surprisingly enough, making the agents cooperate produces a rather different perspective on the statistical mechanical interpretation of inference.

We define a collective of agents as sweeping out a d-dimensional subspace inside all possible fitting parameters. So, rather than fix a single value of the fitting parameters, the collective gets to explore a whole geometry of choices.

We have already encountered the case of a single agent, that is, the case $d = 0$. In the case $d = 1$, we have a line of points. One can visualize this as a grid of points

on the edge of a ruler, or as sprinkled over a circle. In the case $d = 2$, we have many choices for the geometry of our agents: We could sprinkle the agents over the surface of a floor; the surface of a globe; the surface of a doughnut; or the surface a doughnut with extra holes. In $d > 2$, it is harder to visualize the geometry of the agents, but we can still use general methods from Riemannian geometry to treat this case as well.

To illustrate, consider a collective of agents sprinkled over a two-dimensional globe. We can specify a location on the globe in terms of a latitude and longitude, which we collect as two numbers σ_{latitude} and $\sigma_{\text{longitude}}$. For each such σ, we produce a choice of fitting parameter $y(\sigma)$. We view this as the choice made by an agent sitting at this position on the globe. The key difference from the $d = 0$ case is that now, there are many agents, each of which gets to make their own statistical model, and their own choice of fitting parameters.

To be part of the collective, however, the guesses of nearby agents have to be correlated. This is reflected in the condition that a variation in the guess of one agent is linked to the variation of its neighbors:

$$\delta y^I \delta y^J \to h^{ab} \frac{\partial y^I}{\partial \sigma^a} \frac{\partial y^J}{\partial \sigma^b}. \tag{12}$$

Here, σ denotes a local coordinate on the d-dimensional submanifold, and h^{ab} specifies the inverse metric on Σ_{agent}. The local distance between agents is measured via the line element ds_{agent}:

$$ds^2_{\text{agent}} = h_{ab}\ d\sigma^a d\sigma^b. \tag{13}$$

The entire collective should be viewed as defining a "meta-agent", that is, an agent which is not confined to a given statistical model. Rather, it attempts to sample a large number of nearby worldviews, and by doing so, could potentially arrive at a more stable picture of events as they unfold.

To make these considerations more concrete, in [1] we studied the success of the collective by calculating the product of posterior probabilities for all of the agents in the collective:

$$Z(A_{\text{coll}}|E_{\text{coll}}) = \Pr(A_{(1)}|E_{(1)}) ... \Pr(A_{(K)}|E_{(K)}). \tag{14}$$

The idea is that each of the K agents in the collective samples N events. The sample set for the i^{th} agent is $E_{(i)}$ and the family of statistical models for this agent is $A_{(i)}$. The geometric mean of their posterior probabilities is simply $[Z(A_{\text{coll}}|E_{\text{coll}})]^{1/K}$.

Repeating all of the steps for the single agent reviewed in section 2, but now for our whole collection of agents, one again gets a statistical mechanical partition function, but now for a much bigger set of possible configurations: Schematically, this is a sum over all choices of fitting parameters for the collective:

$$Z(A_{\text{coll}}|E_{\text{coll}}) \sim \sum_{\text{config}} \exp(-\mathcal{E}_{\text{config}}/T), \tag{15}$$

where the temperature $T \sim 1/N$ is the same as in the case of the single agent. Here, the sum over configurations is a shorthand for summing over the parameters of the agents:

$$\sum_{\text{config}} \sim \prod_{\text{agents}} \int d^M y_{\text{agent}}. \tag{16}$$

Expanding to leading order in the fluctuations δy, one finds that the "energy" $\mathcal{E}_{\text{config}}$ is the sum of two terms, a kinetic energy and a potential energy:

$$\mathcal{E}_{\text{config}} = \mathcal{E}_{\text{kinetic}} + \mathcal{E}_{\text{potential}}. \tag{17}$$

The kinetic energy is the generalization of equation (10) to a higher dimensional space:

$$\mathcal{E}_{\text{kinetic}} = \int\limits_{\Sigma_{\text{agent}}} d^d\sigma \ \sqrt{\det h}\ G_{IJ}\ h^{ab} \frac{\partial y^I}{\partial \sigma^a} \frac{\partial y^J}{\partial \sigma^b}. \tag{18}$$

The potential energy $\mathcal{E}_{\text{potential}}$ constitutes the prior beliefs of the collective. If the agents have a strong prior belief about the values of the fitting parameters, they can drag the parameters to those values.

One of the main tools available for analyzing these sorts of systems is quantum field theory in "Euclidean signature". The name derives from the difference with the study of field theory in "Lorentzian signature" where a notion of time is singled out in the theory. In Euclidean signature, space and time are on an equal footing in the geometry of the agents. Once space and time are on an equal footing, the famous path integral measure of Feynman's approach to quantum mechanics simply becomes a sum over all the different configurations of fitting parameters (c.f. equation (15)):

$$\prod_{\text{agents}} \int d^M y_{\text{agent}} \sim \int [\mathcal{D}y]\,. \tag{19}$$

Interpreted as a physical theory, a kinetic term of the type in equation (18) leads to a quantum field theory known as a non-linear sigma model. Non-linear sigma models have found wide-ranging application, from the theory of pions, to supersymmetric field theories and string theories. Adding prior beliefs to the collective deforms the sigma model by a potential energy contribution. A helpful analogy is to view the potential energy as the effect of gravity pulling a ball down a hill.

There is a beautiful geometric interpretation of the non-linear sigma model when the potential energy $\mathcal{E}_{\text{potential}} = 0$. In the inference scheme this corresponds to the natural assumption that all nearby families of models are a priori equally plausible. In this case, the sigma model tells us how to embed the space Σ_{agent} for the agents in the space of parameters Y. The choice of all possible ways of performing this embedding is captured by the summing operation of equation (15). This summing process is captured more precisely by the path integral of a quantum field theory in Euclidean spacetime. It literally instructs the collective to try out all possible fitting parameters, and to weight each choice according to the suppression factor $\exp(-\mathcal{E}_{\text{config}}/T)$.

4 Stable Inference and Gravity

Just as in the case of a single agent, there is no guarantee that our collective will reach an accurate inference as to the true nature of the distribution. However, there are some things the collective can do to improve its chances of a successful guessing strategy.

Our main focus here will be on the desirable condition that the collective has reached a stable inference scheme. This means first of all, that a small tap or perturbation in the nature of the true distribution $p_{\text{true}}(x)$ should not completely destabilize the worldview

of the collective. On the other hand, the inference scheme ought to be flexible enough to adapt to big changes.

Here is a qualitative example of why both notions of stability are desirable: If we take a trip to Hawaii and it is sunny for the first twenty days of our visit, we might conclude: "Hawaii is always sunny". If on the twenty-first day we encounter some rain, we should still be able to hold on to some approximate notion that Hawaii is a sunny place. Of course, if we experience many rainy days in a row (say if we visit in a rainy season), then we should be able to revise our original inference.

In the context of statistical inference, our agent can produce a model depending on some continuous parameters. Once it takes this step, it ought to be able to adjust its model, and the values of various continuous parameters in the face of changing events.

Our main focus will be on the growth or decline of a perturbation in the original guess as we average over successively bigger neighborhoods in the collective. Our interest will in particular be in the case where the true distribution receives "a jolt", or equivalently, when the collective "changes its mind". These sorts of perturbations are reflected in the substitution:

$$p_{\text{guess}} \to p_{\text{guess}} + \text{perturbation}. \tag{20}$$

Since these perturbations also implicitly feed into the value of the Fisher information metric in equation (11), and in turn the overall energy in equation (18), we see that these perturbations are something recognizable to the field theory: They deform the kinetic energy of the collective.

The main intuition is that there is a tension between what the collective has agreed on, and what some contrarian agent may find. If there are enough rogue agents, then they can spoil the inference of the collective. Conversely, if the collective is powerful enough, it can suppress dissenting opinion.

Our aim will be to study the effects of these perturbations as a function of how closely we pack the agents on Σ_{agent}. Essentially, we ask how well the inference capabilities of the collective fare as we take bigger and bigger averages over neighbors, then next to nearest neighbors, and then next to next to nearest neighbors. This block averaging over successively bigger neighborhoods of agents provides a way to track the long distance behavior of the collective.

A full analysis of this sort is clearly a complicated matter, but thankfully, we can simply borrow known results from related questions which have received close scrutiny in the physics literature. The averaging procedure we have just summarized is simply one way to track the renormalization of parameters in the effective quantum field theory defined by the agents. "All" we are asking for is the behavior of the perturbations as a function of length scales on Σ_{agent}.

Quite surprisingly, the effects of such perturbations are quite sensitive to the geometry of the agents. While we refer the interested reader to [1] for the technical explanations, here we would like to instead offer a more intuitive discussion.

If an agent is in contact with only a few other agents, then every perturbation can strongly influence its neighbors. Each perturbation ripples throughout the entire collective, causing a dramatic change in the final inference. In one dimension, each contrarian wields enormous influence.

Conversely, if an agent is in communication with many neighbors, the "collective wisdom" of the neighbors can drown out the voices of a few contrarians. In other

words, the group usually wins. This occurs when the agents are in high coordination with one another, that is, for $d = 3$ or more.

In $d = 2$, something new happens: There is enough contact with other agents to influence a neighbor, but the neighbors still exert an influence over a single agent.

Something particularly special happens for the case of a two-dimensional space of agents with uniform prior beliefs, that is, $\mathcal{E}_{\text{potential}} = 0$. The notion of a stable inference can then be analyzed in terms of the condition that in the field theory, a perturbation to the collective will not explode or completely die away in the limit where many agents are making many nearby guesses.

In other applications of non-linear sigma models there has been extensive work on the conditions necessary for such perturbations to lead to a stable perturbation of the original system. These stable perturbations define "conformal fixed points" of the non-linear sigma model.

The interplay between the scaling behavior over successively larger patches of a two-dimensional non-linear sigma model has a rather remarkable connection to the Einstein field equations of classical gravity in vacuum [8–10]. For excellent reviews on the application of these methods in string theory, see [11] as well as the volumes [12, 13]. These equations are encapsulated in a set of interlocking conditions for how the geometry of fitting parameters link together:

$$R_{IJ} - \frac{1}{2} G_{IJ} R = 0. \tag{21}$$

The terms R_{IJ} and R are the Ricci tensor and scalar, respectively, and should be viewed as telling us about how the space of parameters has become curved by the inference strategy. The condition that these contributions balance to zero reflects the special circumstance that there is no explicit source of external information (i.e. stress energy) in the system. Adding such a term T_{IJ} brings us all the way to the Einstein field equations:

$$R_{IJ} - \frac{1}{2} G_{IJ} R = 8\pi G_{\text{Newton}} T_{IJ}. \tag{22}$$

In the statistical inference setting, this means we have exposed the collective to new information. Indeed, as is well known from the principles of general relativity, once we add a source of stress energy, the path taken by test particles will be consequently bent. This bending is nothing but the agents reacting to new information.

Summarizing, we have arrived at the following list of implications:

$$\text{Stable Inference} \Longrightarrow \text{2D Conformal Fixed Point} \Longrightarrow \text{Classical Gravity}. \tag{23}$$

5 Onwards to Strings

For the purposes of inference, it is natural to ask how sensitive the value of $Z(A_{\text{coll}}|E_{\text{coll}})$ will be to the choice of a reference metric $ds^2_{\text{agent}} = h_{ab}\ d\sigma^a d\sigma^b$ which measures the infinitesimal proximity of agents. Rather than making a specific choice, we might instead sum over all possible choices of metric h_{ab}. Extending the path integral sum by this enlarged sum is encapsulated in the generalization:

$$Z(A_{\text{coll}}|E_{\text{coll}}) \to \sum_{h_{ab}\text{'s}} \sum_{\text{config}} \exp(-\mathcal{E}_{\text{config}}/T). \tag{24}$$

In the path integral formulation of Euclidean gravity, this amounts to coupling the space of agents to two-dimensional gravity. In a roundabout way, we have somehow landed at the starting point for the Polyakov approach to string theory!

In this broader context, stability of an inference scheme now leads to some rather strong conditions. For example, one of the famous constraints from perturbative string theory is that the number of fitting parameters is no longer an unconstrained number. Returning to the original setting of statistical inference, there is of course no need to sum over the choice of metrics for the agents. We take this to mean that a stable inference is harder to achieve once the distances between agents are allowed to fluctuate. It would clearly be interesting to further explore the ramifications of this special class of statistical inference schemes.

6 Discussion

In this essay we have explored the interplay between "The Bits" of information in statistical inference, and "The Its" of experience, gravity and string theory. In particular, we have seen that a two-dimensional collective of agents can reach stable inference schemes unavailable to collectives in other dimensions.

As we have repeatedly emphasized, trying to disentangle "The Its" from the "Bits" would require further assumptions beyond those made in this essay. Indeed, our main point is the surprising mileage we have gotten from *not* making wild extrapolations beyond the standard rubric of statistical inference and quantum field theory. All we have required is a collective working together towards a mutual inference. Summing over possible metrics between agents, the appearance of a string theory from such basic considerations is especially striking.

Having come this far, it seems reasonable to turn the tables and ask about the consequences of our analysis for candidate theories of quantum gravity, such as string theory. A major bottleneck in even getting started with the study of quantum gravity is that there is no sharp definition of a "local observable". This is in spite of the fact that great use has been made of this concept in many non-gravitational systems. Here, we have attempted the converse: Starting from an approximate worldview by a collective of agents, we have shown how to recognize the onset of gravity.

But there are also limits to what the collective could ever hope to learn. Indeed, the resolving power of the collective is limited by a precision of order $1/\sqrt{N}$, so arbitrarily sharp inferences cannot be made [1]. The answers to some pressing questions may remain forever shrouded in the murk of statistical noise.

We close with a quote [14]:

"What can be said at all can be said clearly; and whereof one cannot speak, thereof one must be silent."

Acknowledgements: We thank J. Barandes, J.J. Heckman Sr., D. Krohn and A. Murugan for helpful discussions, and ISCAP at Columbia University for hospitality. The work of JJH is supported by NSF grant PHY-1067976.

References

[1] J. J. Heckman, "Statistical Inference and String Theory," `arXiv:1305.3621 [hep-th]`.

[2] T. Bayes and R. Price, "An Essay towards solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M. A. and F. R. S.," *Phil. Trans. of the Royal Soc. of London* **53 (0)** (1763) 370–418.

[3] B. Efron, "Why Isn't Everyone a Bayesian?," *The American Statistician* **40, No. 1** (1986) 1–5.

[4] V. Balasubramanian, "A Geometric Formulation of Occam's Razor For Inference of Parametric Distributions," `arXiv:adap-org/9601001`.

[5] V. Balasubramanian, "Statistical Inference, Occam's Razor and Statistical Mechanics on the Space of Probability Distributions," *Neural Comp.* **9(2)** (1997) 349–368, `arXiv:cond-mat/9601030`.

[6] C. E. Shannon, "A Mathematical Theory of Communication," *Bell Syst. Tech. J.* **27** (1948) 379–423.

[7] S.-I. Amari and H. Nagaoka, *Methods of Information Geometry*, vol. 191. Translations of Mathematical Monographs; American Mathematical Society and Oxford University Press, Providence, RI, 2000.

[8] D. H. Friedan, "Nonlinear Models in $2 + \epsilon$ Dimensions," *Ann. Phys.* **163** (1985) 318.

[9] L. Alvarez-Gaume, D. Z. Freedman, and S. Mukhi, "The Background Field Method and the Ultraviolet Structure of the Supersymmetric Nonlinear Sigma Model," *Ann. Phys.* **134** (1981) 85.

[10] C. G. Callan Jr., E. Martinec, M. Perry, and D. Friedan, "Strings in Background Fields," *Nucl. Phys.* **B262** (1985) 593.

[11] C. G. Callan Jr. and L. Thorlacius, "Sigma Models and String Theory," *Providence 1988, Proc. of the Theor. Adv. Study Inst. in Elem. Part. Physics: Particles, Strings and Supernovae* **2** (1989) 795–878.

[12] J. Polchinski, *String Theory Volume I: An Introduction to the Bosonic String.* Cambridge University Press, Cambridge, U.K., 1998.

[13] J. Polchinski, *String Theory Volume II: Superstring Theory and Beyond.* Cambridge University Press, Cambridge, U.K., 1998.

[14] L. Wittgenstein, *Tractatus Logico-Philosophicus (trans. by F.P. Ramsey and C.K. Ogden).* Kegan Paul, Trench, Trubner & Co. Ltd., London, U.K., 1922.