MACHINE LEARNING
Science and Technology

PAPER • OPEN ACCESS

Bayesian renormalization

To cite this article: David S Berman *et al* 2023 *Mach. Learn.: Sci. Technol.* **4** 045011

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
27 May 2023

REVISED
17 September 2023

ACCEPTED FOR PUBLICATION
6 October 2023

PUBLISHED
18 October 2023

PAPER

Bayesian renormalization

David S Berman[1], Marc S Klinger[2,*] and Alexander G Stapleton[1]

[1] Centre for Theoretical Physics, Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom
[2] Department of Physics, University of Illinois, Urbana, IL 61801, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: marck3@illinois.edu

Keywords: renormalization, Bayesian inference, diffusion learning, information geometry, data compression, Fisher metric

Abstract

In this note we present a fully information theoretic approach to renormalization inspired by Bayesian statistical inference, which we refer to as Bayesian renormalization. The main insight of Bayesian renormalization is that the Fisher metric defines a correlation length that plays the role of an emergent renormalization group (RG) scale quantifying the distinguishability between nearby points in the space of probability distributions. This RG scale can be interpreted as a proxy for the maximum number of unique observations that can be made about a given system during a statistical inference experiment. The role of the Bayesian renormalization scheme is subsequently to prepare an effective model for a given system up to a precision which is bounded by the aforementioned scale. In applications of Bayesian renormalization to physical systems, the emergent information theoretic scale is naturally identified with the maximum energy that can be probed by current experimental apparatus, and thus Bayesian renormalization coincides with ordinary renormalization. However, Bayesian renormalization is sufficiently general to apply even in circumstances in which an immediate physical scale is absent, and thus provides an ideal approach to renormalization in data science contexts. To this end, we provide insight into how the Bayesian renormalization scheme relates to existing methods for data compression and data generation such as the information bottleneck and the diffusion learning paradigm. We conclude by designing an explicit form of Bayesian renormalization inspired by Wilson's momentum shell renormalization scheme in quantum field theory. We apply this Bayesian renormalization scheme to a simple neural network and verify the sense in which it organizes the parameters of the model according to a hierarchy of information theoretic importance.

1. Introduction

Perhaps the central question in data science is the following: how does our understanding of a system improve as we obtain more data? The natural language for formulating this question is through statistical inference [1–3]. From the perspective of statistical inference, our understanding of a system is encoded in the probability we assign to different plausible explanations for how the system works. These explanations are formalized as probability models for observable data specified in terms of various parameters. The probability assigned to each of these models is subsequently encoded in an object called the Bayesian posterior distribution, which can be thought of as a probability distribution over all possible probability distributions for observable data. In terms of these concepts, [4] presents an answer to the aforementioned question by deriving an explicit equation governing the evolution of the posterior distribution as a function of the amount of collected data. We refer to this equation and more broadly to the idea of dynamically updating one's beliefs in light of new data using Bayesian inference as dynamical Bayesian inference, or dynamical Bayes (DB). A central observation from DB is that as new data is collected the 'current' most likely model flows through the space of possible models toward the probability distribution truly responsible for generating observed data.

The idea that learning induces a flow in the space of models is immediately quite evocative of a different kind of 'meta'-theory: the renormalization group (RG). RG is a set of ideas and strategies broadly concerned with formalizing the role of scale in our understanding and formulation of physical theories. In its original form, as conceived of by Kadanoff and Wilson [5–7], the RG consists of taking a system described by a large number of degrees of freedom and performing a coarse-graining operation in which subsets of degrees of freedom are combined together and averaged over to form new collective variables. In physics applications, coarse-graining neighborhoods are determined based on considerations related to locality—that is, degrees of freedom that are nearby in physical space are joined together. For this reason, the RG takes a theory which describes behaviors of a system down to arbitrarily small scales to a new theory which describes behaviors only up to a distance scale that is constrained by the size of the typical coarse graining neighborhood. In the terminology of physics, we say that an RG flow takes a UV theory (a theory that is valid at arbitrarily small scales, or equivalently arbitrarily high energies) to an IR theory (a theory that is valid only at relatively large distances, or equivalently relatively low energies). From the data science perspective, one may think of an IR theory as corresponding to a naive model in which a large amount of data, namely the fine grained description of the system at length scales smaller than the prescribed cutoff, have yet to be incorporated. By contrast, a UV theory has incorporated most if not all of the available data about the system and therefore corresponds to the complete data generating model or 'ground truth'. In what follows we shall often make use of this interpretation of the UV and IR. For our purposes, we will be interested in a relatively modern incarnation of renormalization that typically goes under the name of the exact RG (ERG) [8–17]. ERG seeks to formalize the ideas of renormalization in a more mathematically rigorous fashion, by formulating an ERG flow as a one parameter family of theories governed by a (functional) differential equation.

Viewing a physical theory as a probability distribution for whatever observable degrees of freedom make up the system, the effect of an RG coarse-graining scheme is therefore to induce a flow through the space of possible theories—just like in DB. However, in contrast to the case of learning in which one flows through the space of models toward the data generating model, an RG flow begins at the data generating model (a UV complete theory) and flows away toward some less complete model which remains accurate only for a subset of the original degrees of freedom. This observation motivates the idea that the RG flow may be regarded as a procedure 'inverse' to that of DB, with the former taking a data generating model down to an approximate model and the latter taking an approximate model back to the data generating model. This idea was formalized in [18], in which we showed that the equation governing DB is formally equivalent to an ERG flow if we invert the direction of the flow. More explicitly, performing DB *in reverse* by discarding data as opposed to observing data *defines* an ERG scheme which we refer to as the dynamical Bayesian RG (DB-RG) scheme or simply *Bayesian renormalization*.

In this note we aim to flesh out Bayesian renormalization. In particular, we would like to stress how the DB-RG scheme frees renormalization from its reliance on physical locality. As we alluded to above, when performing renormalization for physical systems one implements a coarse-graining scheme that is directly motivated by physical locality e.g. either one defines collective variables by pooling together degrees of freedom that are contained in a common spatial neighborhood, or one integrates out degrees of freedom that have support on momentum shells above a particular high energy cutoff. In either case, the existence of a hierarchy of physical scales and their role in defining the RG scheme ensures that we can interpret an ERG flow as beginning from a UV theory and ending at some IR fixed point. But what are we to do if the physical system we are interested in renormalizing has non-local interactions? Or worse yet, what if we are interested in renormalizing a model that does not have a physical interpretation at all? Such a situation presents itself in recent work which seeks to import the machinery of renormalization into data science contexts as a tool for performing data compression and improving the interpretability and performance of high dimensional models [19–29].

Bayesian renormalization overcomes the apparent lack of 'real' scale by coming equipped with its own *emergent* scale—the distinguishability of models. Put differently, the space of models has a natural information geometric structure [30–32] endowed by the Fisher metric, which is an infinitesimal measure of the relative entropy between a pair of probability distributions. As we will demonstrate, the DB-RG scheme automatically coarse-grains in a way that respects locality in the space of models, as dictated by the Fisher metric. This fact is particularly serendipitous in touting the utility of DB-RG for renormalizing data science models. In data compression tasks and model building considerations the Fisher metric is used to distinguish between so-called 'sloppy' and 'stiff' parameters. The former covary only weakly with the model output and therefore correspond to small eigenvalues of the Fisher metric, while the latter covary very strongly with the model output and therefore correspond to large eigenvalues of the Fisher metric. To hone the interpretability and generalizability of a model, one may therefore be interested in a scheme which systematically discards sloppy parameters in favor of a model that depends only on strict ones. From the perspective of the Fisher geometry, this can explicitly be thought of as a 'UV regularization scheme'; a consistent method for dealing with the fact that we cannot resolve up to the arbitrarily small distances in model space necessary to identify between models that differ only along directions coordinatized by sloppy parameters [33, 34]. Thus, one way of interpreting the DB-RG scheme is as an automated data compression algorithm which sequentially integrates out 'high energy' parameters (e.g. parameters associated with small eigenvalues of the Fisher metric) in the same way that a physical RG integrates out large momentum shells. In section 4 we provide an explicit demonstration of this fact.

Putting the preceding discussion into more physical language, stiff and sloppy parameters are the data science answer to the idea of relevant and irrelevant operators. In the case where the model of interest *does* possess a 'real' scale, this analogy becomes explicit. For example in the physics literature [35] observed that in the space of conformal field theories, the Fisher metric coincides with the Zamolodchikov metric and thus the hierarchy of relevant and irrelevant operators as dictated by the spectrum of the latter coincides with the hierarchy of stiff and sloppy parameters as dictated by the former. In a similar vein, but working in the opposite direction [36] found that the parameters that were regarded as most significant by the information bottleneck formalism for data compression [37] coincide with the most relevant operators in the conventional RG sense provided the model that is being compressed is given by a local statistical field theory.

In light of these observations and the information theoretic character of the emergent scale in DB-RG, another central goal of this note is to encourage the reader to think about renormalization as a manifestly information theoretic procedure. For example, consider two theories that differ only in terms of modes that exist above some observable momentum scale (ie the UV cutoff). For all intents and purposes these theories are equivalent, as there is no existing experiment which can be conducted to differentiate them. As we have now established, there is a clear analog to this in broad data science contexts: one has two models that differ only along 'sloppy directions' in the model manifold. Sloppy parameters cannot be tuned except with a quantity and/or precision of observation that is not achievable due to experimental limitation. Thus, again, such models should be regarded as practically equivalent. Ultimately, this line of thinking suggests that an RG universality class should correspond to the set of all models/theories which yield equivalent predictions below a threshold set by the amount of useful information that can be collected about the system in question. From this perspective, the relevant notion of scale is always the distinguishability in model/theory space, it just happens that the same information can be communicated in terms of an energy scale in the physical case because such scales bound our experimental capability[3].

The organization of the paper is as follows. In section 2.1 we review exact renormalization in its original physical context. We stress the perspective that a useful subclass of ERG schemes constitute functional diffusion equations, as was originally touted by [41], and recognize the role played by the physical scale in defining such diffusive ERGs. Building on the viewpoint that renormalization and diffusion are equivalent we identify diffusion as a useful device for renormalizing data models even without physical scales in section 2.2. This picture of diffusion based renormalization for general data models is closely related to the influential diffusion learning paradigm [42] in which intractable distributions are run through diffusion channels in order to produce tractable models for data generation tasks. However, in the absence of a physical scale one cannot control the information which is coarse-grained out of the model, in contrast with physical renormalization schemes which always remove information in a hierarchy of real energy scales. This motivates the Bayesian renormalization scheme which is introduced in section 3. After reviewing the Fisher geometry underpinning Bayesian inference (section 3.1), and the dynamical Bayesian inference scheme (section 3.2), we explicitly derive the DB-RG scheme in sections 3.3 and 3.4. The DB-RG scheme can be understood as a specific form of diffusive renormalization in which the information that is lost to diffusion is governed by the Fisher metric. Hence, DB-RG is a precisely the refined form of diffusion learning we sought in which the inverse of the distance measure induced by the Fisher metric plays the role of an energy scale for the purpose of the ERG. In section 4 we illustrate the usefulness of our approach by applying the philosophy of Bayesian renormalization to a simple autoencoder. We conclude with discussion in section 5, in which we review our new perspective on renormalization and suggest future directions both for applications of DB-RG to data science tasks, and as a new theoretical tool.

2. Renormalization and diffusion

In this section we review exact renormalization in its original physical context with an emphasis placed on the relationship between renormalization and diffusion. We establish that an RG flow corresponds to a semigroup of conditional expectation operators acting on a sample space of random variables appropriate to a given theory. These conditional expectations generate a Markov process that can be associated with a

[3] We should note that similar ideas to this have been communicated before in the physics literature by [38, 39], and in the data science literature by [40].

stochastic differential equation, or equivalently with a partial differential equation of the Fokker–Planck form. The stochastic differential equation dictates a coarse-graining scheme in the usual sense, while the partial differential equation absorbs the impact of this coarse graining at the level of the probability distribution describing the variables of interest.

In a physical theory, an RG flow is seeded with information about the hierarchy of momentum scales through the so-called ERG kernel, which ensures that the information coarse grained away by diffusion is associated with high energy data. By contrast, naive diffusion based renormalization discards information indiscriminately in systems that do not possess a physical scale. This motivates the Bayesian renormalization scheme discussed in section 3, which establishes a meaningful scale for arbitrary systems of random variables that can be seeded into the analog of the ERG kernel. This scale is associated with the distinguishability of probability distributions in the space of models/theories, and reproduces a physical scale whenever one is present.

2.1. Exact renormalization is diffusion

Following the lead of [41], we take the perspective that an ERG flow can be understood as a one parameter family of probability distributions, $\{P_\Lambda[\phi]\}_{\Lambda\in\mathbb{R}}$. Here Λ is a physically meaningful RG scale (typically associated with a momentum cutoff), and $\phi\in\mathcal{F}$ corresponds to the field configuration relevant to a given theory. $P_\Lambda[\phi]$ should therefore be read as the probability density assigned to the field configuration ϕ at the scale Λ. Schematically, we regard

$$P_\Lambda[\phi]\propto \mathrm{e}^{-S_\Lambda[\phi]}, \tag{1}$$

where $S_\Lambda[\phi]$ is the renormalized action at scale Λ. The guiding principle of the ERG is that the flow $P_\Lambda[\phi]$ must be chosen in such a way that the partition function is preserved[4]:

$$\frac{\mathrm{d}}{\mathrm{d}\ln\Lambda}\int_{\mathcal{F}}\mathcal{D}\phi\; P_\Lambda[\phi]=0. \tag{2}$$

The ERG principle (2) ensures that all correlation functions below the scale Λ are preserved over the course of the ERG flow.

The most familiar form of exact renormalization is the so-called Polchinski scheme [10]. In Polchinski's ERG, one writes the probability distribution over fields in the form

$$P_\Lambda[\phi]\propto \mathrm{e}^{-\frac{1}{2}\int\frac{\mathrm{d}^dp}{(2\pi)^d}\,\phi(p)G(p^2)K_\Lambda^{-1}(p^2)\phi(-p)}\,\mathrm{e}^{-S_{\mathrm{int},\Lambda}[\phi]}. \tag{3}$$

We recognize the first term as the Gaussian distribution associated with a free field theory with propagator $G(p^2)$, but with the incorporation of a function $K_\Lambda^{-1}(p^2)$ which plays the role of a smooth cutoff function in momentum space. In words, $K_\Lambda^{-1}(p^2)$ suppresses the contribution of momentum modes above the cutoff scale Λ. The second term in (3) is the exponential of the renormalized interacting action at the scale Λ.

In Polchinski's picture, $K_\Lambda(p^2)$ has a prescribed dependence on Λ, thus Polchinski's ERG equation arises by determining the equation which must be obeyed by $S_{\mathrm{int},\Lambda}[\phi]$ in order to satisfy the principle (2). By a straightforward computation, one can show that the resulting equation can be put into the form:

$$\frac{\mathrm{d}}{\mathrm{d}\ln\Lambda}P_\Lambda[\phi]=\int_{M\times M}\mathrm{d}^dx\mathrm{d}^dy\left\{C_\Lambda^{\mathrm{Pol.}}(x,y)\frac{\delta^2P_\Lambda[\phi]}{\delta\phi(x)\,\delta\phi(y)}+\frac{\delta}{\delta\phi(x)}\left(P_\Lambda[\phi]\,C_\Lambda^{\mathrm{Pol.}}(x,y)\frac{\delta V_\Lambda^{\mathrm{Pol.}}[\phi]}{\delta\phi(y)}\right)\right\} \tag{4}$$

$$\equiv\Delta P_\Lambda[\phi]+\mathrm{div}\left(P_\Lambda[\phi]\,\mathrm{grad}_{C_\Lambda^{\mathrm{Pol.}}}V_\Lambda^{\mathrm{Pol.}}[\phi]\right), \tag{5}$$

where

$$C_\Lambda^{\mathrm{Pol.}}(p^2)=(2\pi)^d\,G(p^2)^{-1}\frac{\partial K_\Lambda(p^2)}{\partial\ln\Lambda};\quad V_\Lambda^{\mathrm{Pol.}}[\phi]=\int\frac{\mathrm{d}^dp}{(2\pi)^d}\phi(p)\,G(p^2)\,K_\Lambda^{-1}(p^2)\,\phi(-p). \tag{6}$$

One might recognize (4) as the *Fokker–Planck* equation with diffusion governed by $C_\Lambda^{\mathrm{Pol.}}(p^2)$ and drift governed by the potential $V_\Lambda^{\mathrm{Pol.}}[\phi]$. This is the first indication of a deep relationship between exact renormalization and diffusion. Note that the equivalence between (4) and (5) is just a rewriting in terms of the functional (infinite dimensional) equivalent of vector operators. This is so one can identify (4) as a functional version of Fokker–Plank.

[4] Here $\mathcal{D}\phi$ is the path integral measure. For a review of path integral techniques and functional renormalization see [43], or for a more thorough treatment, see [44].

We refer to the Polchinski approach as an *ERG scheme* because it corresponds to a particular *choice* on how to renormalize the theory described by (3). The nexus of this choice can be traced back to the way Polchinski decided to regulate the action in (3) by introducing a smooth cutoff function $K_\Lambda^{-1}(p^2)$. This choice manifests itself in the particular form of the diffusion and drift aspects of (6) specifying the Fokker–Planck equation associated with the Polchinski ERG (4). Choosing different regulating functions corresponds to different ERG schemes, and as a result different Fokker–Planck equations specified by the data (6).

More abstractly, we can *define* an ERG directly by specifying the data (C_Λ, V_Λ) corresponding to the diffusivity and drift of a Fokker–Planck equation. Equation (4) still holds but with (C_Λ, V_Λ) replacing $C_\Lambda^{\text{Pol.}}(p^2), V_\Lambda^{\text{Pol.}}[\phi]$. This is then an ERG corresponding to a different scheme.

The Fokker–Planck equation corresponds to a bonafide ERG because it satisfies the ERG principle (2). To see that this is the case, let us now show that we can rewrite (4) in the form

$$-\frac{\mathrm{d}}{\mathrm{d}\ln\Lambda}P_\Lambda[\phi] = \int_M \mathrm{d}^d x\, \frac{\delta}{\delta\phi(x)}\left(\Psi_\Lambda[\phi;x]P_\Lambda[\phi]\right), \tag{7}$$

where M is the spacetime manifold on which the theory is defined [12]. Hopefully it is clear that any one parameter family $P_\Lambda[\phi]$ satisfying (7) also satisfies (2). This is because (7) specifies a *divergence flow*, that is the right hand side of (7) is a divergence in the space of field configurations. We can therefore employ the divergence theorem to observe that

$$\frac{\mathrm{d}}{\mathrm{d}\ln\Lambda}\int_{\mathcal{F}}\mathcal{D}\phi\ P_\Lambda[\phi] = -\int_{\mathcal{F}}\mathcal{D}\phi\ \int_M \mathrm{d}^d x\,\frac{\delta}{\delta\phi(x)}\left(\Psi_\Lambda[\phi;x]P_\Lambda[\phi]\right) = 0. \tag{8}$$

In order to write (4) in the form (7) we take

$$\Psi_\Lambda[\phi;x] = \int_M \mathrm{d}^d y\, C_\Lambda(x,y)\,\frac{\delta\Sigma_\Lambda[\phi;P_\Lambda]}{\delta\phi(y)}, \tag{9}$$

as has appeared previously in [12, 16, 17, 41, 45]. Here $C_\Lambda(x,y)$ is the *ERG kernel* appearing in the Fokker–Planck equation associated to the ERG, and $\Sigma_\Lambda[\phi;P_\Lambda]$ is called the *scheme functional* which is determined through the *ERG potential* V_Λ via the equation

$$\Sigma_\Lambda[\phi;P_\Lambda] = -\ln\left(\frac{P_\Lambda[\phi]}{\mathrm{e}^{-V_\Lambda[\phi]}}\right) = S_\Lambda[\phi] - V_\Lambda[\phi]. \tag{10}$$

Plugging (9) back into (7), we reconcile (4) with the diffusion and drift aspects given by (C_Λ, V_Λ), as desired. Together (C_Λ, V_Λ) therefore specify a consistent scheme for regulating the high energy degrees of freedom of the field theory, in analogy with the regulating function $K_\Lambda^{-1}(p^2)$ appearing in (3).

It is worth noting that (7) defines an approach to ERG that is more general than diffusion. All divergence flows, which can generically be written in the form (7), specify ERG flows as evidenced by (8). However, only the subset of divergence flows in which the *reparameterization kernel* Ψ_Λ is taken to be of the form (9) result in Fokker–Planck equations. Equation (7) is called the *Wegner–Morris equation*, and ERG schemes satisfying the Wegner–Morris equation are called *Wegner–Morris schemes*. The choice of Wegner–Morris scheme is encapsulated entirely in the reparameterization kernel. We shall refer to reparameterization kernels which are in the form of (9) as *Fokker–Planck schemes* to highlight their relationship with diffusion. A Fokker–Planck scheme is specified entirely by the data (C_Λ, V_Λ).

To conclude this section, let us briefly explore an alternative way to recognize that the Wegner–Morris equation specifies an ERG flow which allows us to supply a more concise and intuitive interpretation of the ERG. The effect of (7) on P_Λ can be absorbed by continuously reparameterizing the fields ϕ at each new scale according to the rule

$$\phi'(x) = \phi(x) + (\delta\ln\Lambda)\,\Psi[\phi;x]. \tag{11}$$

Equation (11) should be regarded as the integral curve of $\Psi_\Lambda[\phi;x]$ in the space of field configurations, where we are regarding $\Psi_\Lambda[\phi;x]\in T\mathcal{F}$ as a tangent vector to this space. Solving (11) results in a one parameter family of field configurations $\{\phi_\Lambda\}_{\Lambda\in\mathbb{R}}$, in which ϕ_Λ can be thought of as describing the relevant field degrees of freedom at scale Λ. In this way, exact renormalization can be connected with more familiar Wilsonian renormalization schemes by interpreting equation (11) as specifying a coarse graining procedure. The coarse graining map (11) is a diffeomorphism in the space of field configurations, which means it must leave the integral

$$\int_{\mathcal{F}}\mathcal{D}\phi\ P_\Lambda[\phi] \tag{12}$$

invariant. Thus, again, we find that the equation (7) specifies a meaningful renormalization scheme in the usual sense of satisfying (2).

Specializing to Fokker–Planck ERG schemes, we can expand on this discussion. As was introduced in detail in [18], a (functional) Fokker–Planck equation of the form (4) is associated with a (functional) stochastic differential equation (SDE):

$$\mathrm{d}\phi\left(x\right) = -\mathrm{grad}_{C_\Lambda} V_\Lambda\left[\phi\right]\left(d\ln\Lambda\right) + \sqrt{2}\int_M \mathrm{d}^d y\,\sigma_\Lambda\left(x,y\right)\mathrm{d}W_\Lambda\left(y\right). \tag{13}$$

Here, $W_\Lambda(x)$ is a function valued Weiner process, and σ_Λ is the diffusivity kernel defined by the property that it 'squares' to the covariance C_Λ:

$$\int_M \mathrm{d}^d z\,\sigma_\Lambda\left(x,z\right)\sigma_\Lambda\left(z,y\right) = C_\Lambda\left(x,y\right). \tag{14}$$

Equation (13) is the stochastic differential equation that arises from the deterministic gradient flow defined by (11) subject to noise with covariance governed by C_Λ. Thus, we have arrived at the punchline: an ERG flow specified by the data $(\mathcal{F}, C_\Lambda(x,y), V_\Lambda[\phi])$ can be understood as the stochastic coarse graining of the field degrees of freedom arising from the reparameterization of the fields under the gradient of the potential V_Λ subject to noise governed by C_Λ.

2.2. Diffusion is exact renormalization

In section 2.1, we illustrated how ERG flows a la Wegner–Morris correspond to a subclass of functional diffusion equation (7). These Fokker–Planck schemes are specified by the pair (C_Λ, V_Λ) which set, respectively, the diffusion and drift of a stochastic process underyling the Fokker–Planck equation associated with the ERG. The stochastic process (13) explicitly describes a coarse graining of the field degrees of freedom, while the related diffusion equation encodes the impact of such a coarse graining on the renormalized action, resulting in an effective field theory at each scale. One of the main insights of [18] is that viewing exact renormalization as a diffusion process suggests an approach to renormalization which is applicable to a wider class of systems modeled by probability distributions, beyond merely those which posses a physical interpretation. In this section we will outline this approach.

Consider a random variable Y with sample space S. For simplicity, let us assume that Y is a continuous random variable, and S is a Riemannian manifold. A general class of diffusion equations on S are then specified by a one parameter family of probability distributions $\{p_t(y)\}_{t\in\mathbb{R}}$ along with a metric[5] $g_t : TS \times TS \to \mathbb{R}$ and a potential function $V_t : S \to \mathbb{R}$ such that

$$\frac{\mathrm{d}p_t\left(y\right)}{\mathrm{d}t} = \Delta p_t\left(y\right) + \mathrm{div}\left(p_t\left(y\right)\mathrm{grad}_{g_t} V_t\left(y\right)\right). \tag{15}$$

Equation (15) is the Fokker–Planck equation associated with the stochastic differential equation

$$\mathrm{d}Y_t^i = -\left(\mathrm{grad}_{g_t} V_t\right)^i \mathrm{d}t + \sqrt{2}\left(\theta_t\right)^i_j \mathrm{d}W_t^j \tag{16}$$

where[6]

$$\delta^{kl}\left(\theta_t\right)^i_k\left(\theta_t\right)^j_l = g_t^{ij}. \tag{17}$$

It is worth noting here that the g_t plays a double role. From one perspective it is the metric in this process but interpreted in terms of the underlying statistics it is the inverse of the covariance matrix. Then equations (15) and (16) should be compared with (4) and (13) from the exact renormalization context.

As was the case in section 2.1, we can formally solve for the gradient flow of the potential V_t to determine a one parameter family of renormalized degrees of freedom. To be precise, let $\gamma : \mathbb{R} \to S$ be a one parameter family of points in S solving the gradient flow problem

$$\frac{\mathrm{d}\gamma_t}{\mathrm{d}t} = \mathrm{grad}_{g_t} V|_{\gamma_\tau}. \tag{18}$$

[5] Here, TS is the tangent bundle to the manifold S.

[6] Those who are familiar may notice that equation (17) identify $(\theta_t)^i_k$ (at each t) as the components of a vielbeins for the metric g_t.

In terms of γ_t we can now write (16) in the form

$$\mathrm{d}Y_t^i = -\dot{\gamma}_t{}^i\,\mathrm{d}t + \sqrt{2}\,(\theta_t)^i_j\,\mathrm{d}W_t^j. \tag{19}$$

where $\dot{\gamma}_t$ and θ_t correspond to the drift and diffusion, respectively.

The Fokker–Planck equation (15) has a schematic solution:

$$p_t(y) = \int_S \mathrm{d}^d y_0\; \pi(y, y_0; t)\, p_0(y_0)\,. \tag{20}$$

Here, $p_0(y)$ is the initial data, and $\pi(y, y_0; t)$ is the *heat Kernel*. Provided the drift and diffusion are constant over the full sample space the diffusion kernel will be a Gaussian of the form[7]

$$\pi(y, y_0; t) = \mathcal{N}\left(\gamma_t, t g_t^{-1}\right)(y - y_0) \tag{21}$$

Explicitly, $\pi(y, y_0; t)$ is the transition probability density for a sample point to start from y_0 and diffuse to y in a time t. We shall interpret the heat kernel as a stochastic map encoding the implicit coarse graining scheme associated with the diffusive RG flow.

This picture of renormalization is closely related to information theoretic approaches to renormalization from the high energy physics community that have been employed in the study of holography and operator algebras [46–48]. To see this, note that we can encode (20) as a semigroup of *conditional expectation operators* acting on the space of functions on S, $E_t : \Omega^0(S) \to \Omega^0(S)$. The conditional expectation operator E_t acts as[8]

$$E_t(f) = \mathbb{E}_{\pi_t}(f(Y_0)) = \int_S \mathrm{Vol}_S(y_0)\;\pi(y, y_0; t) f(y_0)\,, \tag{22}$$

so that, for example, the posterior predictive distribution is given by

$$p_t(y) = E_t(p_0)(y) = \mathbb{E}_{\pi_t}(p_0(Y_0))\,. \tag{23}$$

The set of operators $\{E_t\}_{t\in\mathbb{R}}$ form a semigroup in the sense that

$$E_{t_2} \circ E_{t_1} = E_{t_1 + t_2}. \tag{24}$$

In a more general context, a conditional expectation on a von Neumann algebra M is a projection of M to a subset $N \subset M$, $E : M \to N$, which retains the normalization of states such that $E(\mathbb{1}_M) = \mathbb{1}_N$ [49]. A very broad class of renormalization schemes accessible to quantum probabilities can subsequently be formulated as a semigroup of conditional expectation operators $\{E_\Lambda\}_{\Lambda\in\mathbb{R}}$ acting on the space of operators affiliated with a given system. In a recent work [48], this form of renormalization was given the name *code subspace renormalization* to reflects its relationship with error correction[9].

The relationship between renormalization and diffusion is very satisfying because we can think of a diffusion process as destroying some of the fine grained information stored in a very complicated, ‘UV complete’ probability distribution. Indeed, this is the point of view which is advocated for in the influential diffusion learning paradigm [42]. In diffusion learning, one begins with a highly complex and intractable distribution p_0 which is run through a forward diffusion process for a time t_f in order to arrive at an analytic distribution, p_f. One can then sample data from p_0 by first sampling data from p_f and subsequently solving a ‘reverse’ stochastic differential equation derived from (16) [53]. To be more precise, given a generic forward diffusion process specified by the SDE

$$\mathrm{d}Y_t^i = \mu_t^i(Y_t)\,\mathrm{d}t + \sigma_t{}^i{}_j(Y_t)\,\mathrm{d}W_t^j, \tag{25}$$

[7] Strictly speaking this form of the heat kernel is only approximate to leading order in the rate of change of g_t, however we can always control how fast g_t changes in order to make sure this form holds to arbitrary precision.

[8] In equation (22) and hereafter we shall use the following standard notation: when taking the expectation value of a function of a random variable, $X \sim p$, the random variable shall appear as a capital letter inside of the expectation and the probability distribution shall appear as a subscript e.g. $\mathbb{E}_p(f(X))$.

[9] In the case that the operator algebra in question is Abelian, the formal definition of a conditional expectation is in one to one correspondence with the set of conditional probability distributions in the ordinary measure theoretic sense. This suggests that one can obtain an explicit form for the renormalization of quantum states (viewed as states on non-Abelian operator algebras) in terms of generalized diffusion processes as discussed in [50–52]. We plan to explore these generalizations in forthcoming work.

the associated reverse SDE is of the form [54]

$$\mathrm{d}Y_s^i = \left\{ \mu_s^i - \frac{1}{2}\partial_j \left(\delta^{kl} {\sigma_s}^i{}_k {\sigma_s}^j{}_l \right) - \delta^{kl} {\sigma_s}^i{}_k {\sigma_s}^j{}_l \partial_j \ln p_s \right\}\Bigg|_{Y_s} \mathrm{d}s + {\sigma_s}^i{}_j \left(Y_s\right) \mathrm{d}W_s^j \tag{26}$$

Here, t, which runs from 0 to t_f, is the forward time, while s, which runs from t_f to the initial time 0, is the reverse time. The reverse SDE is determined by all of the same data as the forward SDE with exception of the score functions—$\partial_j \ln p_s(y)$. Here, $p_s(y)$ is probability distribution obtained by diffusing p_0 according to the forward SDE up to the reverse time s. Simulating the reverse SDE therefore amounts to efficiently reconstructing the scores. In the machine learning community, a resolution to this problem has been presented in terms of the training of a score based generative algorithm (see [53] for a comprehensive review). Interestingly, such a score-based model can be interpreted as a statistical inference problem in which the distribution p_s is *learned* by observing draws from intermediate diffused distributions. We will revisit this in section 3 where we provide an alternative point of view on the same fact by demonstrating that inverting a diffusion process can be thought of as a Bayesian inference experiment.

Based on the observations of the preceding paragraph, one may be tempted to say that diffusion learning can be thought of as a form of exact renormalization for data science models. In light of the correspondence between ERG and diffusion, such a statement is technically correct. However, naive diffusion lacks a very crucial feature which is at the core of a good RG flow; namely a meaningful notion of scale. Indeed, unlike physical renormalization which coarse grains hierarchically over length/energy scales, naive diffusion removes information in an essentially unstructured way. Thus, while basic diffusion can technically be regarded as a form of renormalization, it fails to provide real control over how a probability model changes as a function of any meaningful scale. More plainly, in physics based renormalization we know that the information lost to coarse graining correspond to high energy modes via the construction of the ERG kernel. By contrast, naive diffusion of a probability distribution apparently discards information indiscriminately. Ultimately, this is problem that the Bayesian approach to diffusion/renormalization will overcome.

3. Bayesian renormalization and information geometry

We now turn to an information theoretic approach to renormalization which will provide a meaningful scale that is applicable to arbitrary random variables, and reproduces the physical scale when the chosen random variable descends from a physical system. This notion of scale is defined in terms of the Fisher information metric, and therefore corresponds to the distinguishability between points in the space of probability distributions. Our new approach emphasizes the role of information and information geometry in renormalization, which we now understand broadly as a mechanism for identifying equivalence classes of probability distributions that are indistinguishable as predictive models at a level of precision fixed by the amount of accessible data. This viewpoint allows us to draw a very sharp analogy between renormalization and aspects of data compression [36, 55–58], data generation [59–61], data classification [40, 62–67], dimensional reduction and model selection [33, 34, 68] commonly studied in data science and machine learning. We present this approach to renormalization through its relationship with Bayesian inference to highlight its information theoretic origin.

Bayesian inference is an approach to reconstructing the probability distribution responsible for generating a sequence of observed data. Let Y be a random variable taking values in the sample space S. Given a series of independent draws, $\{y_t\}_{t=1}^T$, from the data generating distribution p_Y^*, the output of a Bayesian inference is a posterior predictive distribution, $p_T(y)$, which is the best approximation to p_Y^* given the data that has been observed. The idea of Bayesian renormalization is to perform Bayesian inference *in reverse* by discarding data rather than incorporating it. The posterior predictive distribution subsequently pools attributes from models that are similar but not equivalent to the data generating model. As we shall illustrate, this defines an RG scheme that automatically encodes an information theoretic notion of relevant and irrelevant degrees of freedom. It is specified by an explicit diffusion process that enforces the relevance criterion by coarse-graining in a way that sequentially 'integrates out' parameters in order of their relevance just like one would sequentially integrate over momentum shells in a Wilsonian renormalization.

3.1. Bayesian inference and information geometry

The first step in Bayesian inference is *model selection*. This corresponds to choosing a parametric family of probability distributions on S, $p_{Y|\Theta}(y \mid \theta)$. *A priori*, we should consider any allowed distribution for Y as a candidate for the data generating distribution. Thus, we might take

$$p_{Y|\Theta}\left(y \mid \theta\right) = \mathrm{e}^{-\theta^i S_i(y)} \tag{27}$$

where here $\{S_j(y)\}_{j\in\mathcal{J}}$ is the set of log-likelihoods corresponding to allowed probability models for Y. In principle the space of models might be infinite dimensional, however for simplicity let us assume that it is finite[10].

Let $\mathcal{M} = \{p_{Y|\Theta}(y \mid \theta) \mid \theta \in \mathbb{R}^n\}$. In words, $\mathcal{M}$ is a space whose points correspond to probability distributions for Y. By construction, this space has a local coordinate system given in terms of the parameters θ. A natural basis for the tangent space of $\mathcal{M}$, is given to us in terms of the *score vectors* $\underline{\ell}_i = \frac{\partial}{\partial\theta^i}\ln(p_{Y|\Theta}(y \mid \theta))$. It is easy to see that these vectors are linearly independent and spanning insofar as they are isomorphic to the coordinate basis $\underline{\partial}_i$. Moreover, when viewed as functions on S the score vectors have zero expectation value due to the normalization condition on p:

$$\mathbb{E}_\theta\left(\underline{\ell}_i\left(Y\right)\right) = \int_S \mathrm{d}^d y\, p_{Y|\Theta}\left(y \mid \theta\right) \frac{\partial \ln\left(p_{Y|\Theta}\left(y \mid \theta\right)\right)}{\partial\theta^i} = \frac{\partial}{\partial\theta^i}\int_S \mathrm{d}^d y\, p_{Y|\Theta}\left(y \mid \theta\right) = 0. \tag{28}$$

In terms of this basis, we define the Fisher information matrix:

$$\mathcal{I}_{ij}\left(\theta\right) = \mathbb{E}_\theta\left(\underline{\ell}_i\underline{\ell}_j\right) = \mathbb{E}_\theta\left(\frac{\partial \ln\left(p_{Y|\Theta}\left(Y \mid \theta\right)\right)}{\partial\theta^i}\frac{\partial \ln\left(p_{Y|\Theta}\left(Y \mid \theta\right)\right)}{\partial\theta^j}\right). \tag{29}$$

The Fisher matrix (29) should be interpreted as the components of a metric, the Fisher metric, on $\mathcal{M}$ in the basis $\{\underline{\ell}_i\}$.

The Fisher metric provides an infinitesimal measure of the similarity between two models in $\mathcal{M}$. This can be seen most clearly through the relationship between the Fisher metric and the KL-divergence. Recall,

$$D_{\mathrm{KL}}\left(\theta \parallel \theta'\right) = \mathbb{E}_\theta\left(\ln\left(\frac{p_{Y|\Theta}\left(Y \mid \theta\right)}{p_{Y|\Theta}\left(Y \mid \theta'\right)}\right)\right). \tag{30}$$

measures the relative entropy between two distributions. D_{KL} is an information divergence, which means that $D_{\mathrm{KL}}(\theta \parallel \theta') \geqslant 0$ and it is equal to zero if and only if $\theta = \theta'$. This makes D_{KL} a good measure of the distinguishability between models. In general, however, D_{KL} is not a symmetric function of θ and θ' and therefore cannot be regarded itself as an inner product. Nonetheless, in the immediate neighborhood of a point $\theta \in \mathcal{M}$ the KL-divergence can be expanded to quadratic order as

$$D_{\mathrm{KL}}\left(\theta \parallel \theta'\right) = \frac{1}{2}\mathcal{I}_{ij}\left(\theta\right)\delta\theta^i\delta\theta^j + \mathcal{O}\left(\delta\theta^3\right). \tag{31}$$

The Fisher metric plays a significant role in parameter estimation because it encodes the sensitivity of a model's output to changes in parameter values [40]. This is closely related to the task of distinguishing between model parameters that are 'sloppy' and 'stiff' [69, 70]. Strict parameters covary strongly with the model output and therefore correspond to large eigenvalues of the Fisher metric. Sloppy parameters, on the other hand, covary only weakly with the model output and therefore correspond to small eigenvalues of the Fisher metric. From a geometric perspective, this means that we can think of sloppy directions—that is directions coordinatized by sloppy parameters—as being highly compact in the space of models. In particular, two models that differ only along sloppy directions will be very hard to distinguish. As a consequence, sloppy parameters require an extensive amount or quality of observed data in order to be fit [71]. However, because these parameters only weakly impact the predictive power of the model it may also be possible to systematically remove them in favor of a reduced model that depends only a sufficient set of strict parameters.

This state of affairs strongly resembles the typical use case of the RG in physical theories. Our ability to meaningfully distinguish between theories is capped by our capacity to perform experiments which measure physics beyond particular scales. In this sense, two theories which yield equivalent predictions except beyond scales that cannot be experimentally probed must be regarded as practically equivalent. If we view a physical theory as parameterized by the Fourier modes of field degrees of freedom, it is precisely the high energy modes which correspond to the 'sloppy parameters' since one requires an extensive or even impossible set of measurements to distinguish between theories that differ only at very high energy scales. From this point of view, we recognize renormalization as a scheme for systematically regulating sloppy model parameters in order to arrive at an equivalence class of models that cannot be distinguished at the level of accuracy admitted by our present ability to observe data.

[10] The following analysis carries over formally to the infinite dimensional case.

In recent work [34] it has been suggested that a similar approach would be very useful in a more broad data science context. Quite serendipitously, this can be thought of as a 'UV regularization scheme' from the perspective of the Fisher geometry—it is a consistent method for dealing with the fact that we cannot resolve up to the arbitrarily small distances in model space necessary to identify between models that differ along directions coordinatized by sloppy parameters. As we introduce the Bayesian renormalization scheme, we will highlight how it *automatically* performs an information geometrically natural regularization to this end, that is oriented directly toward intelligently removing these sloppy degrees of freedom.

3.2. Dynamical Bayesian inference

The upshot of section 3.1 is that the starting point of a Bayesian inference experiment can be understood as the specification of a Riemannian information geometry $(\mathcal{M},\mathcal{I})$, where $\mathcal{M}$ consists of all possible probability models for Y, and $\mathcal{I}$ is an infinitesimal measure of the distinguishability between models. As we have discussed, adopting an information geometric approach to Bayesian inference already allows us to begin formulating the correspondence between learning in the space of models and renormalization. The next step in Bayesian inference is updating, which we regard as a specification of dynamics in the space of models.

In conventional Bayesian inference, the updating phase starts by specifying a *prior* distribution $\pi_0 : \mathcal{M} \to \mathbb{R}$ which acts as the initial data for the dynamical system. In light of observed data, the prior distribution is updated to the posterior distribution, $\pi_T(\theta)$, by using Bayes' law. That is

$$\pi_T(\theta) \propto \left(\prod_{t=1}^{T} p_{Y|\Theta}(y_t \mid \theta)\right) \pi_0(\theta). \tag{32}$$

The constant of proportionality can be deduced by enforcing that π_T be a normalized probability distribution on $\mathcal{M}$. One should interpret (32) as specifying that the probability the data generating model lives in a small neighborhood of the point $\theta \in \mathcal{M}$ is proportional to the probability that one would have observed the existing sample conditional on the underlying parameter value being in such a neighborhood, multiplied by the prior weight given to that neighborhood. In this way the posterior distribution is slowly trained around the region in model space where the data generating distribution lives.

In [4], we asked the question of what Bayesian inference would look like as a continuous time dynamical system. In other words, we think of data as being continuously observed so that the sample $\{y_t\}_{t=1}^{T}$ is growing as a function of the 'time' parameter T. We subsequently showed that the posterior distribution is governed by an integro-differential equation

$$\frac{\partial \pi_T(\theta)}{\partial T} = -\left(D_{\mathrm{KL}}(\theta^* \parallel \theta) - \mathbb{E}_{\pi_T}(D_{\mathrm{KL}}(\theta_* \parallel \Theta))\right)\pi_T(\theta). \tag{33}$$

Here θ_* is the parameter corresponding to the location of the data generating distribution in $\mathcal{M}$, and we have assumed that this value is unique[11]. The equation (33) has a schematic solution

$$\pi_T(\theta) \propto \mathrm{e}^{-TD_{\mathrm{KL}}(\theta_* \parallel \theta)} \tag{34}$$

which we interpret as a Boltzmann distribution with 'energy' given by the KL-divergence between the data generating model and a model at $\theta \in \mathcal{M}$. This observation is intimately related with the idea of treating the distinguishability of models as a kind of 'energy' scale.

At sufficiently late T the posterior distribution will be of the form

$$\pi_T(\theta) = \mathcal{N}\left(\mu_T, \frac{1}{T}\mathcal{I}(\mu_T)^{-1}\right)(\theta). \tag{35}$$

Here μ_T is the T-path of the maximum a posterior (MAP) estimate. We regard the specification of μ_T as defining a new dynamical element. Once the full time path of the MAP has been observed, one can construct a potential function $V : \mathcal{M} \to \mathbb{R}$ for which μ_T is *defined* to be a gradient flow

$$\frac{\mathrm{d}}{\mathrm{d}T}\mu_T = \mathrm{grad}_{\mathcal{I}} V|_{\mu_T}. \tag{36}$$

One can consider V as encoding the details of the sequence with which data was observed in relation to the sequential evolution of the answer to the question of which single model best approximates the data generating distribution. Eventually, as $T \to \infty$ we expect that $\mu_T \to \theta_*$.

[11] This assumption ensures that the posterior predictive distribution will converge to the data generating model almost surely.

One may interpret (35) as specifying that the posterior distribution π_T is localized around the MAP, μ_T, with a characteristic width given by $\frac{1}{T}\mathcal{I}(\mu_T)^{-1}$. Notice, the width of this distribution is shrinking as a function of *T*. In other words, as more and more data is observed, the posterior trains around a smaller and smaller neighborhood of the MAP. This means that the posterior predictive model at time *T* will become more and more precisely attuned to the specific characteristics of the data generating distribution.

We conclude that a dynamical Bayesian inference scheme is specified by the triple $(\mathcal{M}, \mathcal{I}, V)$, where $\mathcal{M}$ specified the set of allowed models, $\mathcal{I}$ endows this set with a notion of *scale* in terms of the distinguishability of nearby models, and V encodes the sequence with which data is observed and the trajectory of the MAP as it converges toward the data generating model. This set of data is cosmetically very similar to the data defining an ERG flow, $(\mathcal{F}, C_\Lambda, V_\Lambda)$. We will now demonstrate how one can *define* an information theoretic ERG flow in terms of the dynamical Bayesian inference scheme.

3.3. Backward inference and model space renormalization

The *posterior predictive distribution* is obtained by taking the convolution of the posterior and the likelihood model:

$$p_T(y) = \mathbb{E}_{\pi_T}\left(p_{Y|\Theta}(y \mid \Theta)\right) = \int_{\mathcal{M}} \mathrm{Vol}_{\mathcal{M}}(\theta)\, \pi_T(\theta)\, p_{Y|\Theta}(y \mid \theta). \tag{37}$$

Taking the posterior to be of the form (35), we can see that (37) is a weighted sum of probability models for Y in which the set of models outside a small neighborhood of the MAP are heavily suppressed. Notice that the posterior distribution is playing a very similar role to the cutoff function in (3), only in the opposite direction. The smooth cutoff suppresses high momentum modes or, in other words, information at short scales. Conversely, as T increases, the posterior distribution suppresses information at large distances as measured by the information geometry on $\mathcal{M}$. This makes sense, as we have dictated renormalization is a form of diffusion in which information is removed from the probability model. By contrast, Bayesian inference incorporates new information with the observation of each new data point. This discrepancy motivates the idea of considering Bayesian inference *in reverse* in which, rather than incorporating new data, an experimenter sequentially removes data from the reconstructed model. We refer to this process as *backward inference*.

Formally, backward inference corresponds to flowing along the inverse 'time' parameter $\tau = \frac{1}{T}$. In terms of this parameter,

$$\pi_\tau(\theta) = \mathcal{N}\left(\mu_\tau, \tau\mathcal{I}(\mu_\tau)^{-1}\right)(\theta). \tag{38}$$

As τ increases the width of the posterior distribution *increases* and a larger set of models are meaningfully incorporated into the posterior predictive distribution[12]. Over time the set of models which live in a small neighborhood of the data generating model will receive less and less weight in the posterior predictive distribution, as the reconstructed model becomes a weighted sum of a more diverse set of models. Inputting (38) into (37), we find the τ-dependent predictive distribution:

$$p_\tau(y) = \int_{\mathcal{M}} \mathrm{Vol}_{\mathcal{M}}(\theta)\ \mathcal{N}\left(\mu_\tau, \tau\mathcal{I}(\mu_\tau)^{-1}\right)(\theta)\, p_{Y|\Theta}(y \mid \theta), \tag{39}$$

which defines a semigroup of conditional expectation values acting in the space of models. Hence we regard (39) as defining an RG flow directly *in the space of models*.

3.4. Bayesian inversion and data space renormalization

Operationally, (39) defines a perfectly reasonable renormalization scheme which coarse grains in model space. However, there is utility to translating (39) so that it can also be understood as explicitly coarse graining in the space of data realizations, as is more standard in typical renormalization schemes. To accomplish this task, we will restrict our attention to spaces of models that are realized in the context of Bayesian inversion [72].

[12] Strictly speaking (38) corresponds to the Bayesian posterior only for sufficiently small τ. However, recall that the prior distribution is irrelevant in a Bayesian update. Thus, we can extend (38) to all values of τ, and interpret the distribution it flows to as $\tau \to \infty$ as some prior distribution that converges to the data generating distribution with sufficient observations. This is also preferable for interpreting (38) as an RG flow, since we are interested in what 'low energy' effective models we can get to at late τ.

The goal of a Bayesian inversion problem is to deduce the signal, θ, that predicated a measured output or data, y. The data and signal are related by a map, $y = G(\theta)$, which we regard as a deterministic model. In practice, either due to explicit stochasticity or limitations to measurement precision, we must regard the realized output as a random variable which depends conditionally on the signal as

$$Y \mid \Theta = G(\Theta) + N \tag{40}$$

where N is a random variable that is conditionally independent of Y and encodes the aforementioned noise. One may therefore read (40) as dictating that Y and Θ are related by a 'law' $Y = G(\Theta)$ but subject to some random fluctuations governed by N. Provided that the noise is distributed with some density $p_N(n)$, we can form the conditional density of Y given Θ by pulling this measure back by (40) to obtain

$$p_{Y|\Theta}(y \mid \theta) = p_N(y - G(\theta)). \tag{41}$$

The procedure described above is familiar from conventional approaches to Bayesian inference for modeling complex systems and constructing neural networks. For example, in the most simple case of a feed-forward neural network, the deterministic function $G(\theta)$ is given by $f(x; W, b)$ where f is the neural network architecture specified by its set of weights, W, and biases, b. In the simplest case of L^2 loss, we take

$$p_{Y|\Theta}(y \mid \theta) = \mathcal{N}\left(0, \sigma^2\right)(y - G(\theta)) \tag{42}$$

where σ^2 is a hyperparameter that sets a scale for the tolerated prediction error.

Given a parametric family of probability distributions for Y which can be written in the form (41), we are motivated to rewrite the posterior predictive distribution (39) as an integral over S by forming the pushforward measure via the mapping G[13].

$$p_\tau(y) = \int_S \mathrm{Vol}_S(y_0)\, \mathcal{N}\left(\gamma_\tau, \tau K_\tau^{-1}\right)(y - y_0)\, p_N(y_0). \tag{43}$$

Here $\gamma_\tau = G(\mu_\tau)$ is the data prediction associated with the MAP estimate, and

$$\left(K_\tau^{-1}\right)^{ab} = \frac{\partial G^a}{\partial \theta^i}\frac{\partial G^b}{\partial \theta^j}\mathcal{I}_\tau^{ij} \tag{44}$$

is the pushforward of the inverse Fisher metric by the map G to give the induced Fisher metric on the data space itself.

We can now compare (43) with (20) and recognize that the posterior distribution, when pushed forward into the sample space by the map G, is the heat kernel of an associated convection–diffusion equation. In particular, (43) can be read as the solution to a Fokker–Planck equation

$$\frac{\mathrm{d}p_\tau}{\mathrm{d}\tau} = \Delta p_\tau + \mathrm{div}\left(p_\tau \mathrm{grad}_{K_\tau} V_\tau\right) \tag{45}$$

or equivalently, as describing a stochastic process

$$\mathrm{d}Y_\tau^a = -\left(\mathrm{grad}_{K_\tau} V_\tau\right)^a \mathrm{d}\tau + \sqrt{2}\,(\theta_\tau)^a_b\, \mathrm{d}W_\tau^b. \tag{46}$$

As discussed in (36), V_τ is a potential function on the sample space for which γ_τ is a gradient flow, and $\delta^{cd}(\theta_\tau)^a_c(\theta_\tau)^b_d = (K_\tau^{-1})^{ab}$.

Equations (43), (45) and (46) complete an explicit mapping between a dynamical Bayesian inference scheme and an ERG flow which we can now represent schematically as the identification:

$$(\mathcal{M}, \mathcal{I}, V) \leftrightarrow (S, K, V). \tag{47}$$

The mapping (47) provides a novel information theoretic interpretation for exact renormalization as the inverse process dual to a Bayesian inference procedure governed by (33). In particular, we have realized our goal of demonstrating how performing the inverse of statistical inference, that is by discarding data rather than collecting data, implements an information theoretic renormalization scheme with emergent scale given by the distinguishability of probability distributions in model space. The role of this scale in coarse graining can be seen through the relationship between the ERG kernel, K_τ^{-1}, and the Fisher information metric.

[13] The pushforward operation is well defined at the level of measures even though the map G may not be invertible since we only need for the inverse image of G to be measurable.

3.5. Comparison with diffusion learning

For clarity, it is helpful to compare the governing equations of Bayesian renormalization and DB with the forward and reverse diffusion processes introduced in section 2.2. From the diffusion learning perspective, one begins by running the data generating distribution through a forward diffusion process specified by an SDE of the form (25). This results in a more tractable distribution which can subsequently be used to generate samples from the data generating distribution by appealing to the associated reverse SDE (26). To accomplish this step one must implement an algorithm which reconstructs the score functions; a task which can be formulated as a Bayesian learning problem. By contrast, our derivation of the Bayesian renormalization scheme has ensued in exactly the opposite direction. We began with a Bayesian learning problem governed by the dynamical Bayesian inference equation (33). Then, we constructed the reverse process (*relative to Bayes*) by studying how the learned distribution changes when data is removed as opposed to incorporated. The result was a diffusion process governed by the SDE (46).

One should identify the forward diffusion, (25), with the Bayesian diffusion process, (46). Similarly, one should identify the reverse diffusion process, (26), with the dynamical Bayesian learning process, (33). In summary, one way of interpreting Bayesian renormalization is as an information theoretic formulation of the observation that the reverse SDE associated with a diffusion process corresponds with a statistical learning problem. The advantage of this approach is that, opposed to a naive diffusion scheme as would be implemented by (20), the Bayesian diffusion scheme intelligently discards information according to a hierarchy of importance as dictated by the Fisher metric. In this respect we hope that Bayesian renormalization can inspire new approaches to information theoretic optimal diffusion learning while also providing insights into the underlying information theoretic character of both diffusion learning and generic renormalization.

To this end, although we have introduced Bayesian renormalization through its relationship with Bayesian inference, we should stress that it motivates a methodology that can be undertaken without the need to first perform a Bayesian inference experiment. In particular, we can renormalize a family of probability distributions by diffusing it through the space of possible models with the diffusion kernel $\mathcal{N}(\gamma_\tau, \tau K_\tau^{-1})$; one can understand this procedure as information theoretic form of diffusion learning where equation (46) provides an explicit proposal for a forward diffusion process which coarse grains information in a hierarchy of importance as specified by the Fisher metric. In turn, the reverse SDE associated with (46) through the identification (26) has an explicit interpretation as encoding the Bayesian learning of the data generating distribution in the time T corresponding to the reintegration of removed data. Here, K_τ^{-1} should be chosen by first computing the Fisher information matrix of the model space of interest, and subsequently pushing this matrix forward in the sample space by selecting a function $G : \mathcal{M} \to S$. The potential function V_τ can be chosen arbitrarily, and will govern the mean tendency of the diffusion process. Thus, the Bayesian renormalization scheme corresponds to the choices of G and V_τ, with $\mathcal{I}$ being fixed by the system. This is significant because $\mathcal{I}$ is what encodes the information theoretic scale of the problem.

The identification of the mapping G should be thought of as a form of data compression in which a deterministic model for the data realizations in terms of the parameters is postulated. Over the course of the Bayesian renormalization, there will be a flowing in these parameters resulting in the formulation of an effective model. This procedure is related to a variety of different data compression techniques such as the information bottleneck [37] and variational autoencoding [56]. In this way, the Bayesian renormalization scheme makes direct contact with previous insights into the relationship between model building/dimensional reduction for complex systems and the RG such as [19–21, 36].

4. Implementation of Bayesian renormalization

The preceding discussion has been largely abstract. Thus, in this penultimate section, we shall provide some insight into the overarching themes of our work through an explicit example. Rather than analyzing a Bayesian inference experiment, we apply the philosophy of Bayesian renormalization directly to a neural network in order to stress the general applicability of our approach. In particular, we will demonstrate how Bayesian renormalization can be used to systematically remove 'sloppy' parameters from a autoencoder network.

Before delving into our experiment, it is instructive to reflect on the relationship between Bayesian inference and machine learning. To quote [73], 'Statisticians and computer scientists often use different language for the same thing.' In this regard, it is helpful to provide a brief dictionary outlining the correspondence between Bayesian inference and the study of neural networks which can be found in table 1. For a more extensive review of how Bayesian statistics is related to neural networks, see [74].

Table 1. Approximate dictionary mapping equivalent quantities in Bayesian statistics to those in the study of neural networks.

Machine learner's approach	Bayesian's approach
Neural network	Statistical model
Weights w_i and biases b_i	Model parameter θ_i
Training sample	Observed data
Trained weights/biases	Parameter terminal distribution θ^*
Minimization of loss function	Maximization of log-likelihood $\ln p_{Y\|\Theta}(y\|\theta)$

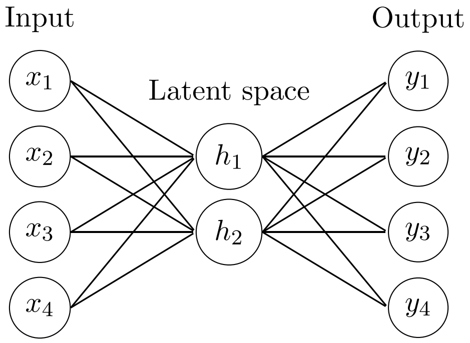


Figure 1. Highly degenerate model architecture with $n = 4, m = 2$. A simple MLP autoencoder with a pair of linear (dense) layers.

4.1. Autoencoders

An autoencoder is a type of (artificial) neural network trained using unsupervised learning which is used to create efficient encodings of a set of unlabeled input data. By definition, autoencoders are composed of two blocks: an encoder and a decoder; typically both blocks are trained simultaneously.

Assuming the encoded and decoded message spaces are Euclidean, let

$$\chi = \mathbb{R}^n \qquad\qquad \tilde{\chi} = \mathbb{R}^m. \tag{48}$$

where n denotes the dimension of the input data, and m the dimension of the *latent space* chosen when the network is initialized. For each data instance $y \in \chi$, we denote the encoded latent space representation as $\tilde{y} \in \tilde{\chi}$. Let E_θ be the encoder function, parameterized by a set of variables $\{\theta_i\}$ and mapping from the space of decoded (source) messages χ to the space of encoded messages $\tilde{\chi}$,

$$E_\theta : \chi \rightarrow \tilde{\chi}. \tag{49}$$

Similarly, D_ϕ is the decoding function, parameterized by a second set of variables $\{\phi_j\}$ and mapping from the space of encoded messages $\tilde{\chi}$ to the space of decoded messages χ,

$$D_\phi : \tilde{\chi} \rightarrow \chi. \tag{50}$$

As a choice of autoencoder architecture, we consider a simple dense, (linear) multi-layer perceptron network with **no biases**, n input neurons, and m latent space neurons (where the number of output neurons is thus constrained also to n). Selecting the architecture as a multi-layer perceptron network is intentional: since each node is fully connected with an associated weight parameter, one may expect the network to have many weights that covary comparatively weakly with the output of the network. A schematic summary of the network architecture is presented in figure 1.

In the case of the aforementioned network, we choose the element-wise encoder and decoder layers to be

$$\left(E_\theta\right)_i = \alpha_{\text{Enc}}\left(\theta_{ij}x^j\right) \qquad\qquad \left(D_\phi\right)_j = \alpha_{\text{Dec}}\left(\phi_{ij}x^j\right), \tag{51}$$

where $\alpha_{\text{Enc}}, \alpha_{\text{Dec}}$ are the encoder and decoder activation functions[14]. Each parameter of the encoder and decoder networks are elements of the corresponding weight matrix θ_{ij}—the machine learner may be more

[14] The activations are typically non-linear.

familiar with the usual form $\alpha(Wx+b)$ where W is the typical weight matrix and b is the corresponding bias vector. For all activations, we choose the *rectified linear unit activation* defined as

$$\alpha(x) = \max(0,x) := \begin{cases} x \text{ if } x>0 \\ 0 \text{ otherwise.} \end{cases} \tag{52}$$

Let $D \subset \chi$ denote the set of training data. Upon the introduction of each new item of training data, $y \in D$, the parameters of the autoencoder are updated[15] to minimize a given loss function $L(y \mid \theta, \phi)$. In the following numerical examples, we consider the typical *mean squared error (MSE) loss* defined by

$$L(y \mid \theta,\phi) = \frac{1}{n}\sum_{j=1}^{n}\left(y_j - D_\phi E_\theta (y)_j\right)^2 . \tag{53}$$

4.2. Information shell renormalization scheme

In the course of this note, we have stressed that a crucial question in any data science context is how we can arrive at an effective model for a given system using only the most relevant set of parameters. Indeed, it is typical for models to possess parameters that are comparatively less sensitive than others, and we have argued that these can be identified systematically via the Fisher matrix. As we have stressed in table 1 a neural network, like our autoencoder, can be understood as a family of statistical models quantified in terms of the various parameters used to initialize its architecture. In our case, each set of parameters (θ,ϕ) corresponds to a different autoencoder, with our best estimate of the 'true' or optimal autoencoder given by the value of these parameters which is realized after all of the training data have been utilized. We denote these optimal parameters by (θ^*,ϕ^*). Appealing again to table 1, we can give a more decidedly probabilistic viewpoint on the family of autoencoders by regarding each autoencoder (θ,ϕ) as initializing a probability distribution over data of the form

$$p_{Y|\Theta,\Phi}(y \mid \theta,\phi) \propto \exp(-L(y \mid \theta,\phi)) . \tag{54}$$

Then, the minimization of the loss can equivalently be regarded as a maximization of the log-likelihood specified in (54).

Regarding the set of all possible autoencoders as a family of statistical models, we can employ the machinery developed in section 3.1 to endow this space with an information geometry. In particular, the Fisher matrix of the **encoder layer** may be calculated explicitly from the **trained model** using the approach outlined in [75]. During this analysis the trained parameters of the decoder are left untouched—this allows comparison between pruning methods while eliminating inter-layer interactions. We denote by $\mathcal{I}_{ij}$ the components of the resulting Fisher matrix. Because it is computed after training one should regard this as the Fisher metric evaluated at the point (θ^*,ϕ^*).

As we have discussed, the eigenvalues of the Fisher metric serve as a proxy for the relevance of the various parameters which are included in the model class. From a physical perspective, one may think of each parameter θ_i as a mode of a particular theory, and of its associated diagonal element in the Fisher metric $\mathcal{I}_{ii}$ as corresponding to its characteristic 'length scale'. Parameters with small Fisher diagonals vary weakly under the acquisition of new data and therefore require either extremely fine measurements or very large amounts of data in order to substantiate even an incremental change in their trained value. For this reason, we can think of these parameters as changing only on very fine length scales in the space of models. Conversely, parameters with larger Fisher values vary strongly with the model and therefore possess features that can be distinguished more broadly with less or less fine-tuned data.

In light of these observations, we now propose the most simple-minded form of Bayesian renormalization. We introduce a sliding parameter, Λ, and divide the set of parameters into two subsets:

$$\Theta_\Lambda^> \equiv \{\theta_i \mid \mathcal{I}_{ii} > \Lambda\}, \qquad \Theta_\Lambda^< \equiv \{\theta_i \mid \mathcal{I}_{ii} \leqslant \Lambda\}. \tag{55}$$

The set $\Theta_\Lambda^>$ consists of those parameters that covary with the model output on 'length' scales greater than Λ. We refer to these parameters equivalently as relevant/stiff parameters relative to the scale Λ for obvious reasons. Conversely, the set $\Theta_\Lambda^<$ consists of those parameters that covary with the model output on 'length' scales below Λ. We refer to these parameters equivalently as irrelevant/sloppy relative to the scale Λ. The renormalization procedure we shall consider is simply to remove the parameters in $\Theta_\Lambda^<$. This realizes a reduced/renormalized model that depends only on the parameters $\Theta_\Lambda^>$. One should think of this as akin to a

[15] For simplicity, we choose the ADAM optimizer, but other choices such as stochastic gradient descent would work as well.

Table 2. Values of configurable properties of the autoencoder network.

Model property	Value
Input (output) dimension	784
Latent space dimension	20
Number of encoder parameters (pre-prune)	15680
Mean diagonal Fisher value	0.7148
Optimizer	ADAM
ADAM $(\alpha, \beta_1, \beta_2, \epsilon)$	$(0.001, 0.9, 0.999, 10^{-8})$
Cutoff Λ	$(0, 0.125, 0.25, \ldots, 2.375)$

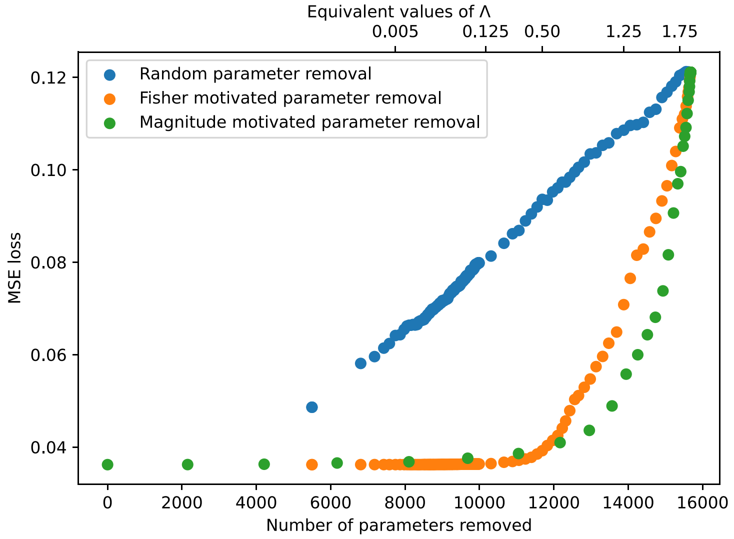


Figure 2. Mean square error loss between the true distribution and output of the pruned networks.

hard cutoff regularization scheme in a quantum field theory (QFT) in which modes that depend on physics at length scales smaller than a given cutoff Λ are systematically removed.

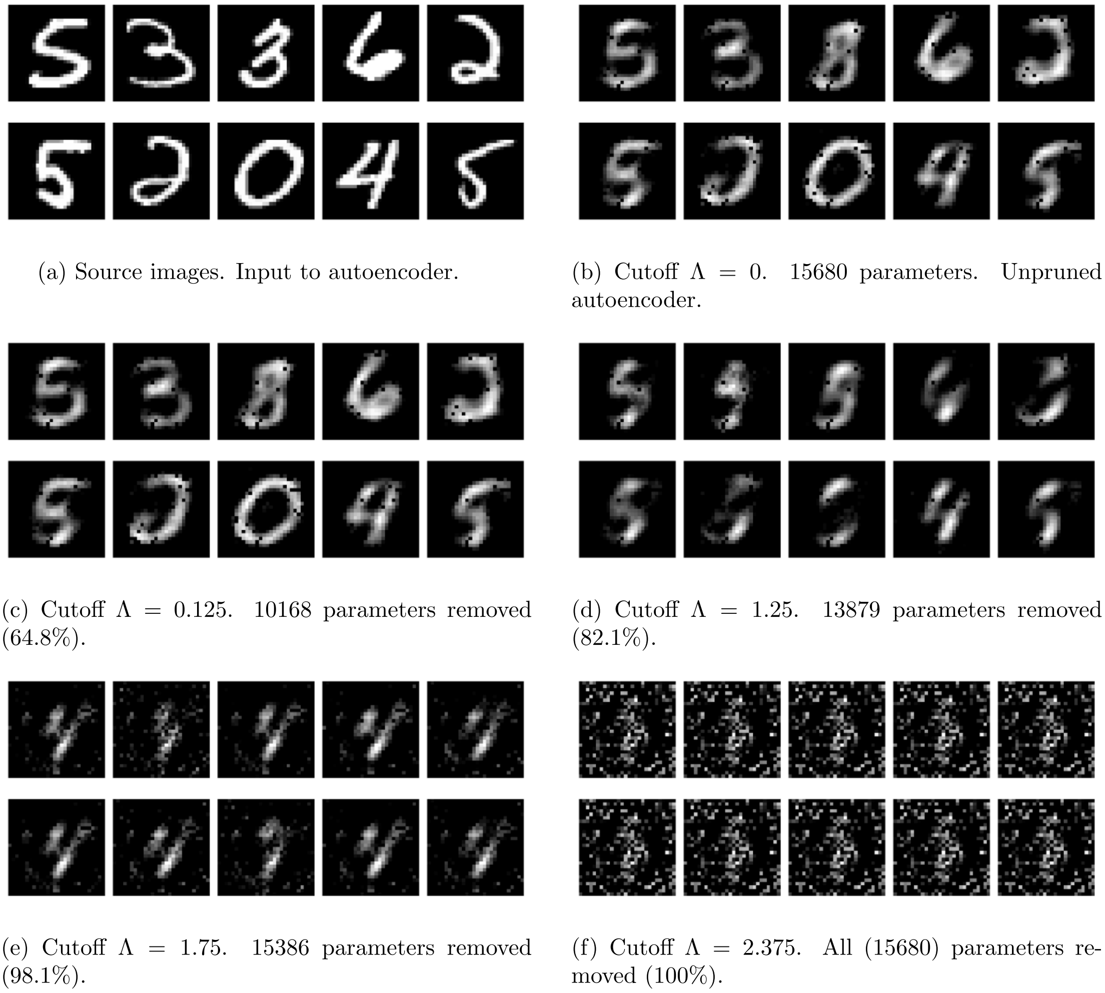
4.3. Experimental results

For our experiment, we begin by training the autoencoder described in section 4.1 on the MNIST dataset. Recall that MNIST is a series of approximately 50 thousand training samples and ten thousand testing samples consisting of 28×28 pixel (784 dimensional) images depicting handwritten digits 0 through 9. The network and optimizer[16] settings are outlined in table 2.

After having trained the model, we perform the Bayesian renormalization scheme outlined above. The result of this procedure at various values of the cutoff parameter Λ can be seen in figure 3. To illustrate the sense in which the Bayesian RG scheme more adequately coarse grains the model with respect to the information contained in its various parameters we have also provided two alternative schemes. In the first scheme, demonstrated in figure 4, the same number of parameters are removed in each iteration as in the Fisher-inspired scheme, but using a uniform distribution over the parameters of the model. The second alternative scheme is a popular form of pruning for neural networks which we refer to as 'magnitude inspired pruning'. In the magnitude inspired approach a fixed number of parameters are removed in each iteration corresponding to the remaining parameters with the smallest absolute value, hence the name.

As one can see in figure 3, the autoencoder remains remarkably effective even once more than half of its parameters have been removed, provided these parameters are removed using the Bayesian scheme. To understand this result it is instructive to look at the distribution of (diagonal) Fisher elements associated with the encoder—see figure 5. Here one can see that there is a large gap in the spectrum at $\Lambda \sim 0.1$—we regard this gap as identifying the genuinely 'sloppy' parameters in the model. In figure 2 we explore the loss landscape as a function of the number of removed parameters, comparing the Fisher-motivated scheme with random parameter removal and magnitude inspired pruning. In regards to this comparison there are two important points to be made. Firstly, the Fisher-motivated scheme produces minimal losses up to the

[16] ADAM parameters are defined in [76].



(a) Source images. Input to autoencoder.

(b) Cutoff $\Lambda = 0$. 15680 parameters. Unpruned autoencoder.

(c) Cutoff $\Lambda = 0.125$. 10168 parameters removed (64.8%).

(d) Cutoff $\Lambda = 1.25$. 13879 parameters removed (82.1%).

(e) Cutoff $\Lambda = 1.75$. 15386 parameters removed (98.1%).

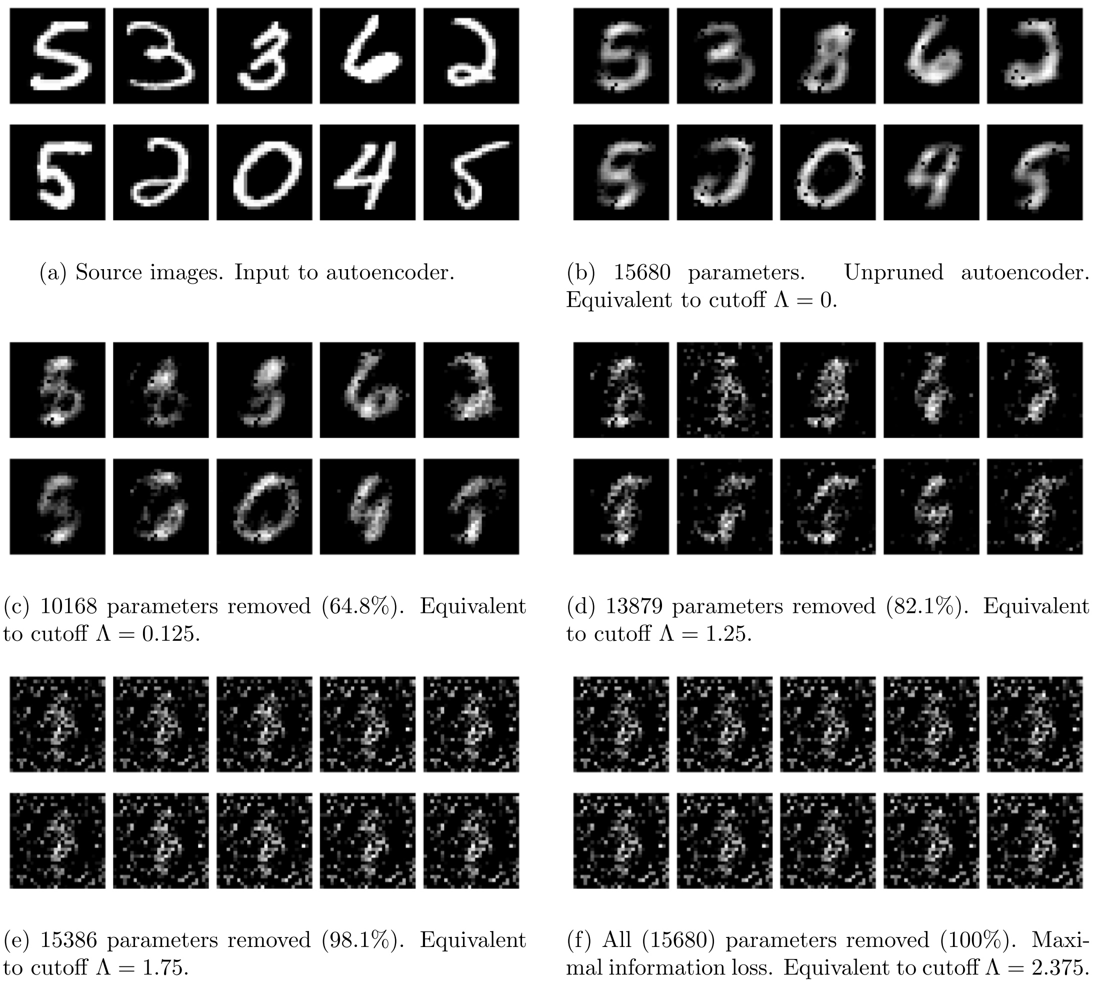
(f) Cutoff $\Lambda = 2.375$. All (15680) parameters removed (100%).

Figure 3. Fisher-motivated pruning with variable cutoffs Λ.

aforementioned cutoff at $\Lambda \sim 0.1$. Shortly thereafter, the losses of the Fisher motivated pruning scheme exceed those of the magnitude scheme. This is expected; for $\Lambda > 0.1$ the Fisher scheme begins to remove parameters that are relevant to the performance of the model. The loss responds strongly to the removal of these parameters. In other words, figure 2 demonstrates that the Fisher scheme effectively distinguishes both irrelevant and relevant parameters.

A different way of understanding the preceding discussion serves to contextualize the relationship between RG and statistical learning. One can imaging performing the pruning procedure in reverse, whereby one incorporates new parameters one at a time in order to improve the model's performance. From this perspective, figure 2 demonstrates that the Fisher scheme most rapidly reaches minimal losses among the three approaches here considered.

One can summarize the preceding discussion by observing the distinctive shape of the loss curve produced by the Fisher pruning scheme. Up to the aforementioned 'scale' of $\Lambda \sim 0.1$ the loss is essentially stable to the pruning of parameters. Thereafter, the removal of subsequent parameters results in noticeable decreases in the performance of the model. The result is a 'hockey-stick' shaped curve: for $\Lambda \lesssim 0.1$ the pruning removes irrelevant parameters that covary weakly with the output of the model, and thus only marginally affect the loss; likewise for $\Lambda > 0.1$ the pruning begins to remove 'relevant' parameters that covary heavily with the output of the model. In other words, the Bayesian RG scheme has identified a characteristic 'scale' in the model and has dictated that the inclusion of parameters which are irrelevant with respect to that scale may be excluded from the model without any change in performance—this is precisely the goal of an RG flow. It is instructive to recognize that the notion of scale which is relevant in this instantiation of RG has a direct relationship with the number of parameters included in the model.



(a) Source images. Input to autoencoder.

(b) 15680 parameters. Unpruned autoencoder. Equivalent to cutoff $\Lambda = 0$.

(c) 10168 parameters removed (64.8%). Equivalent to cutoff $\Lambda = 0.125$.

(d) 13879 parameters removed (82.1%). Equivalent to cutoff $\Lambda = 1.25$.

(e) 15386 parameters removed (98.1%). Equivalent to cutoff $\Lambda = 1.75$.

(f) All (15680) parameters removed (100%). Maximal information loss. Equivalent to cutoff $\Lambda = 2.375$.

Figure 4. Randomly pruned network with equivalent numbers of removed parameters to figure 3.

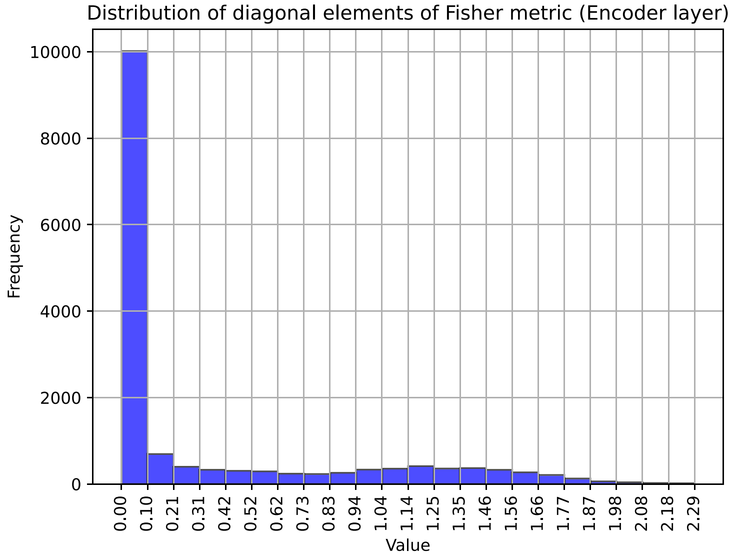


Figure 5. Histogram showing the distribution of diagonal Fisher matrix elements.

5. Discussion

In this note we have outlined a new perspective on renormalization that is fully information theoretic in nature. As a consequence, this approach to renormalization is amenable to arbitrary probability models, not just those which possess a physical interpretation. The major insight we have presented is that the Fisher metric should be interpreted as defining a correlation length in the space of models which defines an emergent scale through the distinguishability of probability distributions. From this perspective, a UV cutoff in a Bayesian renormalization scheme can be understood as fixing the maximum number of possible measurements that could be made about a system, thereby bounding the precision with which one can access information about the data generating model in an inference experiment. This perspective on renormalization is consistent with the more familiar physical picture where the amount of data that can be collected about a theory is bounded by the energy scales which can be probed experimentally. Hence, again, the UV cut-off dictates the set of possible independent measurements. More rigorously, this observation can be quantified through the observation that the KL-divergence is a ERG monotone [41], or the fact that even in RG schemes with a physical scale the ERG kernel can be identified with the Fisher metric on the space of theories [77].

The information theoretic approach to renormalization allows for a very satisfying connection to be made between renormalization and techniques from data science such as model selection, data compression, and data generation. Among the most important insights that arise from using the Fisher metric to define the emergent RG scale is that 'high energy modes' are naturally identified with 'sloppy' parameters which are systematically discarded to formulate 'low energy' effective theories that depend only on 'strict' parameters. In this respect, we regard Bayesian renormalization as an information geometrically inspired coarse-graining scheme. A related perspective on the relationship between renormalization and data compression has been studied through the so called *information bottleneck*, for a representative sample of papers on the topic see [19–21, 36]. The basic idea here is that the information bottleneck identifies a set of low dimensional effective degrees of freedom which efficiently encode the data contained in an otherwise high dimensional space of data realizations. Moving from the original degrees of freedom to the effective degrees of freedom involves a stochastic mapping (conditional probability distribution) which may be interpreted as a form of coarse graining based renormalization. In addition to the connection with data compression, the framing of Bayesian renormalization as a stochastic diffusion process allows one to interpret it as a refinement on the influential diffusion learning paradigm introduced in [42]. The duality between renormalization and statistical learning provides new context for the usefulness of score based generative algorithms for inverting diffusion processes.

At the moment the connections we have addressed are largely conceptual, but in future work we hope to demonstrate how the understanding of Bayesian renormalization can contribute to new automated techniques for data compression and data generation. As a first pass at this problem, we have explored a simple implementation of Bayesian renormalization to a neural network in section 4 to illustrate some of its most salient and useful features. The most crucial insight from this example is that Fisher inspired renormalization can systematically discard degrees of freedom in a hierarchy of importance relative to the model in question. The specific approach we used is a one-to-one adaptation of Wilson's momentum shell renormalization scheme in which momentum shells are replaced by regions of fixed radius in the space of models according to the measure of distance defined by the Fisher metric. No assumptions about the structure or symmetries of the data or model were required in order to perform the aforementioned renormalization. Rather, the structure of the system is learned and subsequently encoded in the hierarchy of scales communicated by the Fisher metric.

Although we have concentrated primarily on the value of the information theoretic character of Bayesian renormalization in data science contexts, let us close by noting that this approach to renormalization may also provide new insights in physics contexts as well. For starters, the information theoretic approach of Bayesian renormalization makes it an ideal tool for identifying and quantifying the precise information that is lost under an RG flow. In this way, one should be able to use Bayesian renormalization in order to interpret and construct RG monotonicity theorems [78–83]. On a different note, a modern perspective on renormalization would not be complete without including entanglement/holographic renormalization: [47, 84–90]. It is a challenge for future work to bring together Bayesian renormalization described in this paper with the holographic description of renormalization that has been developed in the above works. Of relevance to this is the relationship between canonical energy and Fisher metric in holographic context: [91–95]. Finally, the different ways that energy scales with entropy in a physical system tell us about the effectiveness of the usual momentum-based renormalization. An interesting question that one might ask is whether there is a different way of performing renormalization that more appropriately coarse grains information. In particular, large momentum shells may not be the right 'sloppy parameters' for a gravitation theory. Recall that the scaling of entropy with energy is different in gravity from that of local QFTs. This is a key ingredient that allows the holographic property of gravity. The perspective presented here is that the energy cut-off is really an information cut-off. As such this information theoretic perspective suggests that one might consider a different cut-off scheme for gravity than one uses for QFT[17]. This is the power of the Bayesian renormalization principle: it automatically encodes the appropriate designation of relevant and irrelevant degrees of freedom through the Fisher metric and thus ensures that the degrees of freedom that are 'integrated out of the model' correspond precisely with the sloppy parameters, whatever they may be.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/xand-stapleton/fisher_pruning.

Acknowledgments

We thank Jonathan Heckman for collaboration on dynamical Bayes which led to many of the ideas in this paper. We are also grateful for comments from Semon Rezchikov and Miranda Cheng on related work which was presented during Stringdata2022, and to Leenoy Meshulam, Adam Kline and Michael Abbott for helpful conversations about the intersection between renormalization and data science at the APS March Meeting 2023. Finally, we would like to thank Samuel Goldman and Robert Leigh for enlightening discussions on Exact Renormalization and its relationship with Entanglement Renormalization. D S B and A G S acknowledge support from Pierre Andurand over the course of this research. M S K is supported through the Physics department at the University of Illinois at Urbana-Champaign.

ORCID iDs

Marc S Klinger https://orcid.org/0009-0001-7666-1185
Alexander G Stapleton https://orcid.org/0009-0009-6784-7779

References

[1] Jaynes E T 1957 Information theory and statistical mechanics *Phys. Rev.* **106** 620
[2] Jaynes E T 2003 *Probability Theory: The Logic of Science* (Cambridge University Press)
[3] Gelman A and Shalizi C R 2013 Philosophy and the practice of Bayesian statistics *Br. J. Math. Stat. Psychol.* **66** 8–38
[4] Berman D S, Heckman J J and Klinger M 2022 On the dynamics of inference and learning (arXiv:2204.12939)
[5] Kadanoff L P 1966 Scaling laws for Ising models near T_c *Phys. Phys. Fiz.* **2** 263
[6] Wilson K G 1971 Renormalization group and critical phenomena. I. Renormalization group and the Kadanoff scaling picture *Phys. Rev.* B **4** 3174
[7] Wilson K G and Kogut J 1974 The renormalization group and the ε expansion *Phys. Rep.* **12** 75–199
[8] Wegner F J 1974 Some invariance properties of the renormalization group *J. Phys. C: Solid State Phys.* **7** 2098–108
[9] Wegner F J and Houghton A 1973 Renormalization group equation for critical phenomena *Phys. Rev.* A **8** 401
[10] Polchinski J 1984 Renormalization and effective Lagrangians *Nucl. Phys.* B **231** 269–95
[11] Morris T R 1994 Derivative expansion of the exact renormalization group *Phys. Lett.* B **329** 241–8
[12] Morris T R 1994 The exact renormalization group and approximate solutions *Int. J. Mod. Phys.* A **9** 2411–49
[13] Morris T R 1998 Elements of the continuous renormalization group *Prog. Theor. Phys. Suppl.* **131** 395–414
[14] Latorre J I and Morris T R 2000 Exact scheme independence *J. High Energy Phys.* JHEP11(2000)004
[15] Bagnuls C and Bervillier C 2001 Exact renormalization group equations: an introductory review *Phys. Rep.* **348** 91–157
[16] Morris T R and Rosten O J 2006 Manifestly gauge invariant QCD *J. Phys.* A **39** 11657–81
[17] Rosten O J 2012 Fundamentals of the exact renormalization group *Phys. Rep.* **511** 177–272
[18] Berman D S and Klinger M S 2022 The inverse of exact renormalization group flows as statistical inference (arXiv:2212.11379)
[19] Meshulam L, Gauthier J L, Brody C D, Tank D W and Bialek W 2018 Coarse graining, fixed points and scaling in a large population of neurons *Phys. Rev. Lett.* **123** 178103
[20] Meshulam L, Gauthier J L, Brody C D, Tank D W and Bialek W 2018 Coarse–graining and hints of scaling in a population of 1000+ neurons (arXiv:1812.11904)
[21] Kline A G and Palmer S E 2022 Gaussian information bottleneck and the non-perturbative renormalization group *New J. Phys.* **24** 033007
[22] Mehta P and Schwab D J 2014 An exact mapping between the variational renormalization group and deep learning (arXiv:1410.3831)
[23] Lin H W, Tegmark M and Rolnick D 2017 Why does deep and cheap learning work so well? *J. Stat. Phys.* **168** 1223–47
[24] Halverson J, Maiti A and Stoner K 2021 Neural networks and quantum field theory *Mach. Learn.: Sci. Technol.* **2** 035002
[25] Luo D and Halverson J 2023 Infinite neural network quantum states (arXiv:2112.00723)
[26] Halverson J 2021 Building quantum field theories out of neurons (arXiv:2112.04527)
[27] Brown A R, Freedman M H, Lin H W and Susskind L 2021 Effective geometry, complexity, and universality (arXiv:2111.12700)
[28] He Y-H, Heyes E and Hirst E 2023 Machine learning in physics and geometry (arXiv:2303.12626)

[17] A manifestation of this idea can be seen in [96, 97], where the authors identified a relationship between the vacuum energy and entropy, and subsequently proposed a new approach to the cosmological constant problem in which the regulator scheme constrains the number of gravitational degrees of freedom by evoking the holographic principle.

[29] Erdmenger J, Grosvenor K T and Jefferson R 2022 Towards quantifying information flows: relative entropy in deep neural networks and the renormalization group *SciPost Phys.* **12** 041

[30] Amari S-I and Nagaoka H 2000 *Methods of Information Geometry* vol 191 (American Mathematical Society)

[31] Amari S-I 2016 *Information Geometry and Its Applications* vol 194 (Springer)

[32] Nielsen F 2020 An elementary introduction to information geometry *Entropy* **22** 1100

[33] Strandkvist C, Chvykov P and Tikhonov M 2020 Beyond RG: from parameter flow to metric flow (arXiv:2011.12420)

[34] Quinn K N, Abbott M C, Transtrum M K, Machta B B and Sethna J P 2022 Information geometry for multiparameter models: new perspectives on the origin of simplicity (arXiv:2111.07176)

[35] Balasubramanian V, Heckman J J and Maloney A 2015 Relative entropy and proximity of quantum field theories *J. High Energy Phys.* JHEP05(2015)104

[36] Gordon A, Banerjee A, Koch-Janusz M and Ringel Z 2021 Relevance in the renormalization group and in information theory *Phys. Rev. Lett.* **126** 240601

[37] Tishby N, Pereira F C and Bialek W 2000 The information bottleneck method (arXiv:physics/0004057)

[38] Bény C and Osborne T J 2015 Information-geometric approach to the renormalization group *Phys. Rev.* A **92** 022330

[39] Bény C and Osborne T J 2015 The renormalization group via statistical inference *New J. Phys.* **17** 083005

[40] Raju A, Machta B B and Sethna J P 2018 Information loss under coarse graining: a geometric approach *Phys. Rev.* E **98** 052112

[41] Cotler J and Rezchikov S 2023 Renormalization group flow as optimal transport (arXiv:2202.11737)

[42] Sohl-Dickstein J, Weiss E A, Maheswaranathan N and Ganguli S 2015 Deep unsupervised learning using nonequilibrium thermodynamics (arXiv:1503.03585)

[43] Cardy J 2010 (available at: https://www-thphys.physics.ox.ac.uk/people/JohnCardy/qft/qftcomplete.pdf) (Accessed 6 September 2023)

[44] Peskin M E, Schroeder D V and Martinec E 1996 An introduction to quantum field theory *Phys. Today* **49** 69–72

[45] Matsumoto M, Tanaka G and Tsuchiya A 2020 The renormalization group and the diffusion equation *Prog. Theor. Exp. Phys.* **2021** 023B02

[46] Faulkner T 2020 The holographic map as a conditional expectation (arXiv:2008.04810)

[47] Furuya K, Lashkari N and Ouseph S 2022 Real-space RG, error correction and petz map *J. High Energy Phys.* JHEP01(2022)170

[48] Gesteau E 2023 Large *N* von Neumann algebras and the renormalization of Newton's constant (arXiv:2302.01938)

[49] Takesaki M 1972 Conditional expectations in von Neumann algebras *J. Funct. Anal.* **9** 306–21

[50] Carlen E A and Maas J 2014 An analog of the 2-Wasserstein metric in non-commutative probability under which the fermionic Fokker-Planck equation is gradient flow for the entropy *Commun. Math. Phys.* **331** 887–926

[51] Carlen E A and Maas J 2017 Gradient flow and entropy inequalities for quantum Markov semigroups with detailed balance *J. Funct. Anal.* **273** 1810–69

[52] Carlen E A and Maas J 2020 Non-commutative calculus, optimal transport and functional inequalities in dissipative quantum systems *J. Stat. Phys.* **178** 319–78

[53] Song Y, Sohl-Dickstein J, Kingma D P, Kumar A, Ermon S and Poole B 2020 Score-based generative modeling through stochastic differential equations (arXiv:2011.13456)

[54] Anderson B D 1982 Reverse-time diffusion equation models *Stoch. Process. Appl.* **12** 313–26

[55] Legeza O and Sólyom J 2004 Quantum data compression, quantum information generation and the density-matrix renormalization-group method *Phys. Rev.* B **70** 205118

[56] Kingma D P and Welling M 2022 Auto-encoding variational Bayes (arXiv:1312.6114)

[57] Machta B B, Chachra R, Transtrum M K and Sethna J P 2013 Parameter space compression underlies emergent theories and predictive models *Science* **342** 604–7

[58] Kingma D, Salimans T, Poole B and Ho J 2021 Variational diffusion models *Advances in Neural Information Processing Systems* vol 34 pp 21696–707

[59] Gui J, Sun Z, Wen Y, Tao D and Ye J 2021 A review on generative adversarial networks: algorithms, theory and applications *IEEE Trans. Knowl. Data Eng.* **35** 3313–32

[60] Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M and Sutskever I 2021 Zero-shot text-to-image generation *Int. Conf. on Machine Learning* (PMLR) pp 8821–31

[61] Ramesh A, Dhariwal P, Nichol A, Chu C and Chen M 2022 Hierarchical text-conditional image generation with CLIP latents (arXiv:2204.06125)

[62] Teng Y, Sachdev S and Scheurer M S 2023 Classifying topological neural network quantum states via diffusion maps (arXiv:2301.02683)

[63] Bansal A, Borgnia E, Chu H-M, Li J S, Kazemi H, Huang F, Goldblum M, Geiping J and Goldstein T 2022 Cold diffusion: inverting arbitrary image transforms without noise (arXiv:2208.09392)

[64] Fischer K, René A, Keup C, Layer M, Dahmen D and Helias M 2022 Decomposing neural networks as mappings of correlation functions *Phys. Rev. Res.* **4** 043143

[65] Fleig P and Nemenman I 2022 Statistical properties of large data sets with linear latent features *Phys. Rev.* E **106** 014102

[66] Rodriguez-Nieva J F and Scheurer M S 2019 Identifying topological order through unsupervised machine learning *Nat. Phys.* **15** 790–5

[67] Neal R M 2001 Annealed importance sampling *Stat. Comput.* **11** 125–39

[68] Bahri Y, Dyer E, Kaplan J, Lee J and Sharma U 2021 Explaining neural scaling laws (arXiv:2102.06701)

[69] Daniels B C, Chen Y-J, Sethna J P, Gutenkunst R N and Myers C R 2008 Sloppiness, robustness and evolvability in systems biology *Curr. Opin. Biotechnol.* **19** 389–95

[70] Transtrum M K, Machta B B, Brown K S, Daniels B C, Myers C R and Sethna J P 2015 Perspective: sloppiness and emergent theories in physics, biology and beyond *J. Chem. Phys.* **143** 07B201_1

[71] Abbott M C and Machta B B 2023 Far from asymptopia (arXiv:2205.03343)

[72] Dashti M and Stuart A M 2015 The Bayesian approach to inverse problems (arXiv:1302.6989)

[73] Wasserman L 2004 *All of Statistics: A Concise Course in Statistical Inference* vol 26 (Springer)

[74] Jospin L V, Laga H, Boussaid F, Buntine W and Bennamoun M 2022 Hands-on Bayesian neural networks—a tutorial for deep learning users *IEEE Comput. Intell. Mag.* **17** 29–48

[75] George T 2021 NNGeometry: easy and fast fisher information matrices and neural tangent kernels in PyTorch

[76] Kingma D P and Ba J 2017 Adam: a method for stochastic optimization (arXiv:1412.6980)

[77] Floerchinger S 2023 Exact flow equation for the divergence functional (arXiv:2303.04082)

[78] Zamolodchikov A B 1986 Irreversibility of the flux of the renormalization group in a 2D field theory *J. Exp. Theor. Phys. Lett.* **43** 730–2
[79] Alvarez E and Gomez C 1999 Geometric holography, the renormalization group and the c-theorem *Nucl. Phys.* B **541** 441–60
[80] Myers R C and Sinha A 2010 Seeing a c-theorem with holography *Phys. Rev.* D **82** 046006
[81] Casini H and Huerta M 2007 A c-theorem for entanglement entropy *J. Phys. A: Math. Theor.* **40** 7031
[82] Casini H, Huerta M, Myers R C and Yale A 2015 Mutual information and the f-theorem *J. High Energy Phys.* JHEP10(2015)003
[83] Casini H, Testé E and Torroba G 2017 Markov property of the conformal field theory vacuum and the a theorem *Phys. Rev. Lett.* **118** 261602
[84] Swingle B 2012 Entanglement renormalization and holography *Phys. Rev.* D **86** 065007
[85] Nozaki M, Ryu S and Takayanagi T 2012 Holographic geometry of entanglement renormalization in quantum field theories *J. High Energy Phys.* JHEP10(2012)193
[86] Mollabashi A, Naozaki M, Ryu S and Takayanagi T 2014 Holographic geometry of cMERA for quantum quenches and finite temperature *J. High Energy Phys.* JHEP03(2014)098
[87] Leigh R G, Parrikar O and Weiss A B 2014 Holographic geometry of the renormalization group and higher spin symmetries *Phys. Rev.* D **89** 106012
[88] Leigh R G, Parrikar O and Weiss A B 2015 Exact renormalization group and higher-spin holography *Phys. Rev.* D **91** 026002
[89] Evenbly G and Vidal G 2015 Tensor network renormalization *Phys. Rev. Lett.* **115** 180405
[90] Goldman S, Lashkari N, Leigh R G and Moosa M 2023 Exact renormalization of wave functionals yields continuous MERA (arXiv:2301.09669)
[91] Lashkari N 2016 Modular Hamiltonian for excited states in conformal field theory *Phys. Rev. Lett.* **117** 041601
[92] Lashkari N and Van Raamsdonk M 2016 Canonical energy is quantum fisher information *J. High Energy Phys.* JHEP04(2016)153
[93] Banerjee S, Erdmenger J and Sarkar D 2018 Connecting fisher information to bulk entanglement in holography *J. High Energy Phys.* JHEP08(2018)001
[94] Faulkner T, Haehl F M, Hijano E, Parrikar O, Rabideau C and Raamsdonk M V 2017 Nonlinear gravity from entanglement in conformal field theories *J. High Energy Phys.* JHEP08(2017)057
[95] Erdmenger J, Grosvenor K and Jefferson R 2020 Information geometry in quantum field theory: lessons from simple examples *SciPost Phys.* **8** 073
[96] Freidel L, Kowalski-Glikman J, Leigh R G and Minic D 2023 The vacuum energy density and gravitational entropy (arXiv:2212.00901)
[97] Freidel L, Kowalski-Glikman J, Leigh R G and Minic D 2023 On the inevitable lightness of vacuum (arXiv:2303.17495)