Editors' Suggestion

Renormalization group flow as optimal transport

Jordan Cotler[1,2,3] and Semon Rezchikov[4]

[1]*Harvard Society of Fellows, Cambridge, Massachusetts 02138 USA*
[2]*Black Hole Initiative, Harvard University, Cambridge, Massachusetts 02138 USA*
[3]*Department of Physics, Harvard University, Cambridge, Massachusetts 02138 USA*
[4]*Department of Mathematics, Harvard University, Cambridge, Massachusetts 02138 USA*

(Received 7 November 2022; accepted 5 June 2023; published 5 July 2023)

We establish that Polchinski's equation for exact renormalization group (RG) flow is equivalent to the optimal transport gradient flow of a field-theoretic relative entropy. This provides a compelling information-theoretic formulation of the exact renormalization group, expressed in the language of optimal transport. A striking consequence is that a regularization of the relative entropy is in fact an RG monotone. We compute this monotone in several examples. Our results apply more broadly to other exact renormalization group flow equations, including widely used specializations of Wegner-Morris flow. Moreover, our optimal transport framework for RG allows us to reformulate RG flow as a variational problem. This enables new numerical techniques and establishes a systematic connection between neural network methods and RG flows of conventional field theories.

DOI: 10.1103/PhysRevD.108.025003

I. INTRODUCTION

The renormalization group is one of the central ideas in quantum field theory and statistical field theory, enabling us to understand how the effective description of a physical system changes as we tune the precision of our measurement apparatus. There are many ways of mathematically formulating the renormalization group (RG), although a particularly illuminating way is via so-called exact renormalization group (ERG) equations, pioneered by Wilson [1] and refined by Polchinski [2] and many others [3–5]. ERG equations are intrinsically nonperturbative and have been used extensively in analytical and numerical investigations of RG flow over the past 40 years [3–5].

A widely used ERG equation is Polchinski's [2], which is a functional differential equation for RG flow in a natural RG scheme. We show that Polchinski's equation can be recast as a gradient flow of a relative entropy. The gradient here is with respect to a functional generalization of the optimal transport metric (specifically, a version of the Wasserstein-2 metric). The theory of optimal transport [6] is presently less known to physicists, but it is a rich subject which has had a profound impact on partial differential equations and probability theory in mathematics, and optimization as well as machine learning in computer science. We provide a review of the subject for physicists. Our results show that optimal transport is deeply ingrained in the theory of RG, enabling us to bring powerful tools from optimal transport to bear on nonperturbatively analyzing RG flows. For instance, we precisely explain the manner in which RG flows generate entropy and clarify how this interplays with scheme dependence; we discover a new (nonperturbative) RG monotone; and we develop a novel variational formula for RG flow which can be applied in the design of numerical methods for the renormalization group. Our methods work for a more general class of ERG equations beyond Polchinski's, and moreover our framework provides an elegant explanation of otherwise unintuitive features of popular ERG schemes [7–9].

Let us provide a brief sketch of our results in slightly more detail. To illustrate the basic setup of ERG equations, we consider a Euclidean scalar field theory on $\mathbb{R}^d$. This means that we have a probability functional $P[\phi(x)] \propto e^{-S[\phi]}$, where $\phi\colon \mathbb{R}^d \to \mathbb{R}$ and $S[\phi]$ is the Euclidean action. Suppose that our measurement apparatus can only probe the system down to some small distance scale ℓ, corresponding to a UV cutoff $\Lambda \sim 1/\ell$ on the largest momenta we can access. Now let $P_\Lambda[\phi] \propto e^{-S_\Lambda[\phi]}$ denote the probability functional corresponding to an effective description of our system given that we can only probe momentum scales less than Λ. We are interested in how this effective description changes as we tune the value of Λ, i.e. change the precision of our measurement apparatus. An ERG will address this in the form of a functional differential equation

$$-\Lambda\frac{d}{d\Lambda}P_\Lambda[\phi]=\mathcal{F}\left[P_\Lambda[\phi],\frac{\delta P_\Lambda[\phi]}{\delta\phi},\frac{\delta^2 P_\Lambda[\phi]}{\delta\phi\delta\phi},\ldots\right]. \tag{1.1}$$

The minus sign on the left-hand side indicates that we are *coarse graining* $P_\Lambda[\phi]$ in momentum space (which is done on a log scale on account of the $\Lambda\frac{d}{d\Lambda}$). Also, the precise form of the function $\mathcal{F}$ on the right-hand side is contingent on the details of our RG scheme, or equivalently the manner in which we coarse grain our description of the physical system in order to provide an effective description commensurate with the capabilities of our measurement apparatus. Later on, we will precisely specify $\mathcal{F}$ for common RG schemes.

One of our main results is that Polchinski's equation can be written as

$$-\Lambda\frac{d}{d\Lambda}P_\Lambda[\phi]=-\nabla_{\mathcal{W}_2}S(P_\Lambda[\phi]\|Q_\Lambda[\phi]), \tag{1.2}$$

where $\nabla_{\mathcal{W}_2}$ is a gradient with respect to a functional generalization of the Wasserstein-2 metric, $S(P\|Q)\coloneqq\int[d\phi]P[\phi]\log(P[\phi]/Q[\phi])$ is a functional version of the relative entropy, and $Q_\Lambda[\phi]$ is a background probability functional which essentially defines our RG scheme. We emphasize that our formula has the flexibility of capturing an enormous class of RG schemes. The ingredients of our formula require further explanation, which we will provide in detail later. Intuitively, (1.2) tells us that the coarse graining of our theory is generated by a decrease in a relative entropy. We will later see that the relative entropy is in fact an RG monotone; although this may seem clear from the form of (1.2), a more detailed analysis is required which involves unpacking the definition of the gradient.

The remainder of the paper is organized as follows. In Sec. II we review ERG with an emphasis on Polchinksi's equation, as well as the theory of optimal transport. In Sec. III we establish Eq. (1.2) and a generalization pertaining to a broader class of ERG equations. In Sec. IV we prove that the relative entropy appearing in our flow equations is in fact a nonperturbative RG monotone. In Sec. V we compute some explicit examples of the RG monotone for both a free and interacting scalar field. In Sec. VI we leverage dual formulations of optimal transport to develop a variational formula for RG flows, and then explain how it can be leveraged for new numerical methods. Finally in Sec. VII we conclude with a discussion.

II. REVIEW OF EXACT RG AND OPTIMAL TRANSPORT

Here we review pertinent tools and results about the exact renormalization group, as well as optimal transport theory.

A. Exact RG

The ERG is a nonperturbative framework for implementing the renormalization group in quantum and statistical field theory [5]. In standard treatments of field theory, RG is usually implemented perturbatively via an expansion in small couplings. By contrast, ERG provides a means to perform RG for all couplings including large couplings; in practice this is often implemented by numerical approximation schemes, but sometimes analytic methods are possible. We begin by reviewing one of the simplest ERG equations due to Polchinski [2] which will be our jumping off point for generalizations.

1. Polchinki's equation

In the spirit of Polchinski's analysis, we restrict ourselves to scalar field theory for simplicity. We note that Polchinski's equation can be generalized to fermionic theories [4,10,11] and gauge theories [10,12,13].

Let us recapitulate a version of Polchinski's derivation from [2]. Consider a Euclidean scalar field theory with a source J. We will set $\hbar=1$ throughout. The partition function is

$$Z_\Lambda[J]\coloneqq\int[d\phi]e^{-\frac{1}{2}\int\frac{d^dp}{(2\pi)^d}(\phi(p)\phi(-p)(p^2+m^2)K_\Lambda^{-1}(p^2)+J(p)\phi(-p))-S_{\mathrm{int},\Lambda}[\phi]}, \tag{2.1}$$

where $S_{\mathrm{int},\Lambda}[\phi]$ includes interaction terms (possibly including quadratic terms which contribute to the explicit kinetic term) and where $K_\Lambda(p^2)$ is a soft cutoff function, i.e. it is 1 for $p^2\lesssim\Lambda^2$ and ≈ 0 for $p^2\gtrsim\Lambda^2$, and $K_\Lambda^{-1}(p^2)$ denotes $1/K_\Lambda(p^2)$. This soft cutoff function ensures that correlation functions are regulated at high momentum. For our purposes, it will be convenient for $K_\Lambda(p^2)$ to never equal zero, even if it is extremely close to zero; this way $K_\Lambda^{-1}(p^2)$ is never strictly infinite. An example of a soft cutoff function is shown in Fig. 1. Also note that the mass m appearing above in (2.1) is the bare mass, and the couplings implicit in $S_{\mathrm{int},\Lambda}[\phi]$ are bare couplings.

We desire to consider some smaller scale $\Lambda_R<\Lambda$, and integrate out all modes down to Λ_R. As such, we are only interested in computing correlation functions below the scale Λ_R, and so let us assume that our source satisfies $J(p)=0$ for $p^2>\Lambda_R^2-\varepsilon$ for some small $\varepsilon>0$. It is convenient to restrict $|m^2|\ll\Lambda_R$, i.e. we are not integrating out the mass scale.

Suppose that Λ_R is infinitesimally smaller than Λ. Then we would like for

$$-\Lambda\frac{d}{d\Lambda}Z_\Lambda[J]=C_\Lambda Z_\Lambda[J] \tag{2.2}$$

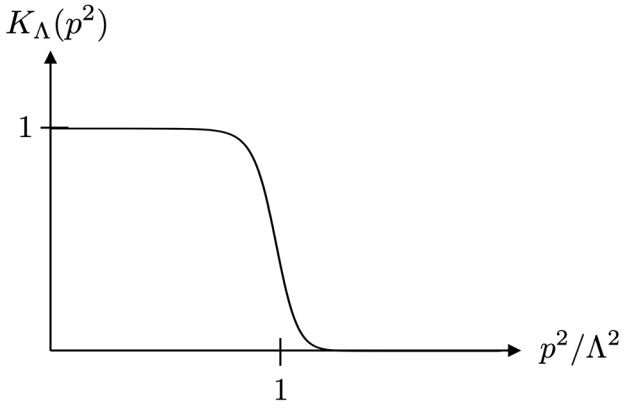

FIG. 1. Depiction of a smooth cutoff function $K_\Lambda(p^2)$.

for some constant C_Λ only depending on Λ. This would mean that as we change the cutoff scale Λ, which both affects the kinetic term in the action in an explicit way and the interaction terms in a way to be determined, any correlation functions below the changed scale (i.e. generated by taking functional J derivatives) stay the same. Expanding out the left-hand side we find

$$-\Lambda\frac{d}{d\Lambda}Z_\Lambda[J]=\int[d\phi]\left(\frac{1}{2}\int\frac{d^dp}{(2\pi)^d}\phi(p)\phi(-p)(p^2+m^2)\times\Lambda\frac{\partial K_\Lambda^{-1}(p^2)}{\partial\Lambda}+\Lambda\frac{\partial S_{\mathrm{int},\Lambda}[\phi]}{\partial\Lambda}\right)e^{-S_\Lambda[\phi,J]}. \quad (2.3)$$

If we want (2.2) to hold, then $\Lambda\frac{\partial S_{\mathrm{int}}[\phi,\Lambda]}{\partial\Lambda}$ must have an appropriate form to facilitate this. Remarkably, Polchinski found such a sufficient form which corresponds to a spatially local coarse graining of $S_{\mathrm{int},\Lambda}[\phi]$ upon Fourier transforming to position space. In particular, we will demand that $S_{\mathrm{int},\Lambda}[\phi]$ changes with respect to Λ via

$$-\Lambda\frac{\partial S_{\mathrm{int},\Lambda}[\phi]}{\partial\Lambda}=\frac{1}{2}\int d^dp(2\pi)^d(p^2+m^2)^{-1}\Lambda\frac{\partial K_\Lambda(p^2)}{\partial\Lambda}\times\left\{\frac{\delta^2S_{\mathrm{int},\Lambda}}{\delta\phi(p)\delta\phi(-p)}-\frac{\delta S_{\mathrm{int},\Lambda}}{\delta\phi(p)}\frac{\delta S_{\mathrm{int},\Lambda}}{\delta\phi(-p)}\right\}. \quad (2.4)$$

This is what is known as Polchinski's equation, and it is sometimes written as

$$-\Lambda\frac{\partial}{\partial\Lambda}e^{-S_{\mathrm{int},\Lambda}[\phi]}=\frac{1}{2}\int d^dp(2\pi)^d(p^2+m^2)^{-1}\times\Lambda\frac{\partial K_\Lambda(p^2)}{\partial\Lambda}\frac{\delta^2}{\delta\phi(p)\delta\phi(-p)}e^{-S_{\mathrm{int},\Lambda}[\phi]} \quad (2.5)$$

in order to resemble a functional version of the heat equation. Note the appearance of $\Lambda\frac{\partial K_\Lambda(p^2)}{\partial\Lambda}$ in both (2.4) and (2.5); this is localized in momentum space around $p^2=\Lambda^2$, corresponding to a smearing kernel with scale $\sim 1/\Lambda$ in position space. See Fig. 2 for a depiction in momentum space. Plugging (2.4) into (2.3) and simplifying, we find

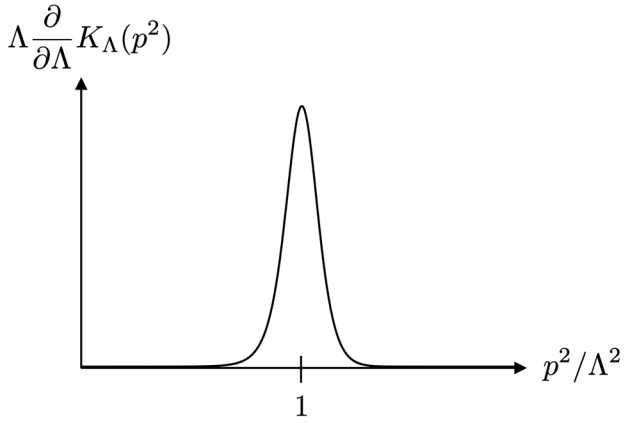

FIG. 2. The derivative $\Lambda\frac{\partial}{\partial\Lambda}K_\Lambda(p^2)$ of the smooth cutoff function.

$$-\Lambda\frac{d}{d\Lambda}Z_\Lambda[J]=\left(-\frac{1}{2}\int d^dp\Lambda\frac{\partial\log K_\Lambda(p^2)}{\partial\Lambda}\delta^d(0)\right)Z_\Lambda[J] \quad (2.6)$$

which has the form of the desired transformation from (2.2).

While Polchinski's equation (2.4) is formulated in terms of a functional equation for $S_{\mathrm{int},\Lambda}[\phi]$, it will be convenient for us to recast it in terms of a functional equation for the probability functional $P_\Lambda[\phi]=e^{-S_\Lambda[\phi]}/Z_\Lambda$. Reprocessing the above derivation we arrive at

$$-\Lambda\frac{d}{d\Lambda}P_\Lambda[\phi]=\frac{1}{2}\int d^dp(2\pi)^d(p^2+m^2)^{-1}\times\Lambda\frac{\partial K_\Lambda(p^2)}{\partial\Lambda}\frac{\delta^2}{\delta\phi(p)\delta\phi(-p)}P_\Lambda[\phi]+\int d^dp\Lambda\frac{\partial\log K_\Lambda(p^2)}{\partial\Lambda}\frac{\delta}{\delta\phi(p)}(\phi(p)P_\Lambda[\phi]), \quad (2.7)$$

which has the form of a functional convection-diffusion equation. To see the connection more clearly, we rewrite the above as

$$-\Lambda\frac{d}{d\Lambda}P_\Lambda[\phi]=\frac{1}{2}\int d^dp(2\pi)^d(p^2+m^2)^{-1}\times\Lambda\frac{\partial K_\Lambda(p^2)}{\partial\Lambda}\frac{\delta^2}{\delta\phi(p)\delta\phi(-p)}P_\Lambda[\phi]+\frac{1}{2}\int d^dp(2\pi)^d(p^2+m^2)^{-1}\times\Lambda\frac{\partial K_\Lambda(p^2)}{\partial\Lambda}\frac{\delta}{\delta\phi(p)}\times\left(\frac{2(p^2+m^2)}{(2\pi)^dK_\Lambda(p^2)}\phi(p)P_\Lambda[\phi]\right), \quad (2.8)$$

which formally takes the same form as the finite-dimensional convection-diffusion equation

$$\frac{d}{dt}p_t(x) = \partial_i\partial^i p_t(x) + \partial_i(v^i(x)p_t(x)), \tag{2.9}$$

where we identify $-\log\Lambda$ with t. An example of a solution to (2.7) [or equivalently (2.8)] is the free theory itself; that is, $P_\Lambda[\phi] = \frac{1}{Z_\Lambda}\exp(-\frac{1}{2}\int\frac{d^dp}{(2\pi)^d}\phi(p)\phi(-p)(p^2+m^2)K_\Lambda^{-1}(p^2))$ solves Polchinski's equation.

Let us summarize the logic of Polchinski's derivation. We explicitly differentiated $Z_\Lambda[J]$ by $\Lambda\frac{d}{d\Lambda}$, and then found a choice of $\Lambda\frac{\partial S_{\mathrm{int},\Lambda}[\phi]}{\partial\Lambda}$ such that $-\Lambda\frac{d}{d\Lambda}Z_\Lambda[J] = C_\Lambda Z_\Lambda$ is satisfied for a constant C_Λ. The suitable choice of $\Lambda\frac{\partial S_{\mathrm{int}}[\phi]}{\partial\Lambda}$, given by the functional differential equation in (2.4), corresponds to changing $S_{\mathrm{int},\Lambda}$ in a manner which is localized in momentum space at scale Λ, and hence local in position space at scale $\sim 1/\Lambda$. While Polchinski's inspired *Ansatz* (2.4) does the job, there are in fact an infinitude of other choices which have similar properties and also render $-\Lambda\frac{d}{d\Lambda}Z_\Lambda[J] = C_\Lambda Z_\Lambda$. These other choices correspond to alternative RG schemes than the one proposed by Polchinski. We explore a large family of them via our discussion of the Wegner-Morris flow equation below.

2. Wegner-Morris flow equation

Polchinski's equation is a special case of the Wegner-Morris flow equation [8,14–16]. The latter provides insights into the structure of RG flows which are obscured by Polchinski's formulation. The Wegner-Morris equation is[1]

$$-\Lambda\frac{d}{d\Lambda}P_\Lambda[\phi] = \int d^dx\frac{\delta}{\delta\phi(x)}(\Psi_\Lambda[\phi,x]P_\Lambda[\phi]) \tag{2.10}$$

and implements ERG for a scheme determined by $\Psi_\Lambda[\phi,x]$. We note that $\Psi_\Lambda[\phi,x]$ will depend on $P_\Lambda[\phi]$ in a nontrivial way, which we explain below. At first glance (2.10) does not appear to readily connect to RG flow, but its meaning will be clear shortly.

To gain some intuition for (2.10), it is useful to compare with a finite-dimensional analog. This would be the equation for p_t given by

$$\frac{d}{dt}p_t(x) + \partial_i(V^i(p_t,x)p_t) = 0. \tag{2.11}$$

In this equation the vector field V^i, the analog of Ψ_Λ in the Wegner-Morris flow, is chosen to depend not just on the coordinate position x but also on the entire probability distribution p_t. It is natural for V^i to satisfy $V^i(p_t,x) = \partial^i W(p_t,x)$, namely for V^i to have a potential W. This gives us

$$\frac{d}{dt}p_t(x) + \partial_i(\partial^i W(p_t,x)p_t) = 0. \tag{2.12}$$

We will find an analog of this potential in the Wegner-Morris flow equation for many cases of interest.

Equation (2.10) has several features which illuminate its meaning. First, performing the functional integral of both sides of (2.10) with respect to $\phi(x)$ and noting that $\Psi_\Lambda[\phi,x]P_\Lambda[\phi]$ goes to zero for large $\phi(x)$, we immediately see that $-\Lambda\frac{d}{d\Lambda}\int[d\phi]P_\Lambda[\phi] = 0$ and so the flow equation preserves probability. More generally, the meaning of (2.10) is that as the scale Λ changes the flow induces the field reparametrization

$$\phi'(x) = \phi(x) + \frac{\delta\Lambda}{\Lambda}\Psi_\Lambda[\phi,x]. \tag{2.13}$$

This means that the probability functional is simply reparametrized by the flow, and so probability is clearly conserved and positivity of the probability density is maintained. As explained in [16], essentially all RG schemes (with a soft cutoff) can be cast into the form of the Wegner-Morris flow equation above. In all schemes Ψ_Λ instantiates field redefinitions which are localized in momentum space near scale Λ, i.e. we are reparametrizing the field at or near the cutoff scale. We will henceforth refer to Ψ_Λ as the reparametrization kernel.

A common form of $\Psi_\Lambda[\phi,x]$ is given by [5,7–9,17,18]

$$\Psi_\Lambda[\phi,x] = -\int d^dy\frac{1}{2}\dot{C}_\Lambda(x-y)\frac{\delta\Sigma_\Lambda[\phi]}{\delta\phi(y)}, \tag{2.14}$$

where $\dot{C}_\Lambda(x-y)$ is called the ERG kernel[2] which satisfies $\dot{C}_\Lambda(x-y) \geq 0$, and

$$\Sigma_\Lambda[\phi] \coloneqq S_\Lambda[\phi] - 2\hat{S}_\Lambda[\phi], \tag{2.15}$$

where $S_\Lambda[\phi]$ is the action appearing in $P_\Lambda[\phi] = e^{-S_\Lambda[\phi]}/Z_\Lambda$ and $\hat{S}_\Lambda[\phi]$ is another action called the "seed action." The multiplicative factor of 2 in front of the seed action is conventional. In its present form, the meaning of the seed action is physically obscure. Fortunately, our optimal transport analysis later on will elucidate its meaning. Notice that $\Psi_\Lambda[\phi,x]$ is a gradient of $\Sigma_\Lambda[\phi]$, where $\frac{1}{2}\dot{C}_\Lambda(x-y)$ plays the role of an inverse metric, and so in this setting the Wegner-Morris flow equation (2.10) takes the form of the finite-dimensional equation (2.12).

[1]We are in fact writing down a slight modification of the usual Wegner-Morris equation; the original equation only implies $-\Lambda\frac{d}{d\Lambda}Z_\Lambda = 0$, whereas we have modified the equation to allow for $-\Lambda\frac{d}{d\Lambda}Z_\Lambda = C_\Lambda Z_\Lambda$.

[2]We have chosen a different sign convention than the usual literature, namely $\dot{C}_{\Lambda,\mathrm{us}} = -\dot{C}_{\Lambda,\mathrm{them}}$. This extra minus sign will make some of our later formulas more intuitive.

Importantly, we can reproduce the Polchinski's equation with the choices

$$\dot{C}_\Lambda(p^2) = (2\pi)^d (p^2+m^2)^{-1}\Lambda \frac{\partial K_\Lambda(p^2)}{\partial \Lambda}, \tag{2.16}$$

$$\hat{S}_\Lambda = \frac{1}{2}\int \frac{d^d p}{(2\pi)^d}(p^2+m^2)K_\Lambda^{-1}(p^2)\phi(p)\phi(-p), \tag{2.17}$$

here expressed in momentum space.[3] Notice that $\hat{S}$ is just an action for a free massive scalar field with the same initial bare mass as our scalar field theory of interest.

An initially puzzling feature of Wegner-Morris flow is that (2.13) can be inverted if $\Psi_\Lambda[\phi, x]$ is well enough behaved. This would mean that the exact RG flow is invertible. However, we often think of RG as being non-invertible, perhaps the most famous example being Kadanoff's block spin decimation for spin systems (see e.g. [19,20]). For continuum field theories, exact RG flows are typically invertible, although the inversion is ill-conditioned. As an example close in spirit to Kadanoff's block spin methods, suppose that our RG flow is prescribed by the coarse graining[4] $P_\Lambda[\phi] = \int [d\psi]\delta[\phi - b_\Lambda[\psi]]P_{\Lambda_0}[\psi]$, where Λ_0 is the initial RG scale and $\Lambda \leq \Lambda_0$. Here $b_\Lambda[\psi](x) := \int d^d y f_\Lambda(x-y)\psi(y)$, where $f_\Lambda(x-y)$ is a smearing kernel with width $\sim 1/\Lambda$ in position space. Perhaps $f_\Lambda(x-y)$ is a Gaussian distribution, or a d-dimensional unit box function (which has compact support). So we are performing a continuum version of Kadanoff's procedure. However, a key difference is that the smearing $\int d^d y f_\Lambda(x-y)\psi(y)$ is invertible in the continuum as can be seen by transforming to Fourier space to get $f_\Lambda(p)\psi(p)$ and dividing by $f_\Lambda(p)$. Indeed, if $f_\Lambda(x-y)$ is a Gaussian, then so is its Fourier transform; dividing by a Gaussian is well-defined, albeit ill-conditioned since we are dividing by very small numbers in the tail regions. Likewise the Fourier transform of a box function is the product of sinc functions, and division by them is likewise ill-conditioned.[5]

More broadly, even when we perform exact versions of the more standard Wilsonian RG, the flow is only in general invertible if we keep the infinitely many irrelevant terms in the action generated by the flow.

[3]In the equation for $\dot{C}_\Lambda(p^2)$, the right-hand side is greater than or equal to zero. Since $\dot{C}_\Lambda(p^2)$ is continuous, Bochner's theorem implies that its Fourier transform $\dot{C}_\Lambda(x-y)$ is likewise greater than or equal to zero.

[4]This can be written in the form of the Wegner-Morris flow equation (2.10), albeit with Ψ_Λ taking a form different from the *Ansatz* class in (2.14), (2.15). See Ref. [5] for a discussion.

[5]Here we also need to be careful about dividing by zero at isolated points, but this can be dealt with if the fields ψ belong to a sufficiently nice function class. Relatedly, the invertibility of block spin renormalization fails to apply to the discretized lattice setting due the function class corresponding to latticized fields.

B. Optimal transport

As discussed above, Polchinski's equation for P_Λ is an infinite-dimensional convection-diffusion equation, which can be thought of as a generalized form of heat flow. The RG monotones we present later on will be analogs of the entropy of a distribution. The fact that the entropy of a distribution is monotone along heat flows was already known to Gibbs. However, the understanding that the entropy functional generates heat flow under the Wasserstein metric required a synthesis [21] of ideas about *optimal transport*. This synthesis occurred relatively recently in the 1990s, in the work of Otto, Benamou-Brenier, and many others. We will review some of these developments here.

At a high level, the problem of optimal transportation is to determine an optimal method for moving and rearranging a given mass distribution into a desired mass distribution, given a cost for moving mass across a specified distance. In the next three subsections, we will review the basic mathematical formalization, discuss fundamental results about this problem, and explain how it connects with heat flow. Beyond this connection, there is a rich theory connecting optimal transport with probability theory and mathematical physics, and we will provide a short guide to relevant literature for interested readers.

1. Monge and Kantorovich formulations

Given a space X and a pair of probability or mass distributions p, q on X, the *Monge formulation* of the optimal transport problem asks to find a (measurable) *transport function* $T: X \to X$ such that:

(1) The pushforward of p under T is q, i.e. $T_* p = q$; equivalently $\int_{T^{-1}(S)} dx p(x) = \int_S dx q(x)$ for every measurable set S; and
(2) The transport function minimizes the total cost

$$\mathsf{M}[T] = \int_X dx p(x) c(x, T(x)) \tag{2.18}$$

for some cost function $c: X \times X \to \mathbb{R}$.

A natural choice for the cost function is $c(x,y) = d(x,y)^2$, where d is a distance function on X. A depiction of the mapping $T_* p = q$ can be seen in Fig. 3.

The constraint $T_* p = q$ is highly nonlinear, making the existence of a solution nonobvious. For concreteness,

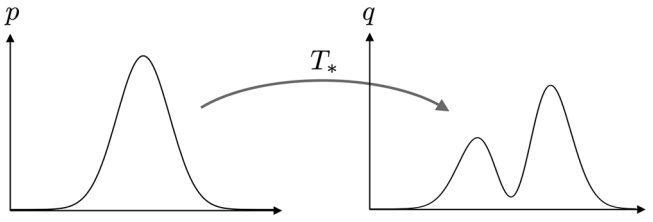

FIG. 3. Schematic of the probability mass distributions p and q and the map T_* between them.

suppose that $X = \mathbb{R}^n$ and T is a smooth function. We let T_j denote the jth coordinate output of T. Then the constraint can be written as

$$q(T(x))|\det(\partial_i T_j(x))| = p(x). \tag{2.19}$$

This nonlinear constraint above makes it difficult to establish the existence of solutions to the Monge problem via methods from the calculus of variations. Worse, solutions to the Monge problem no longer exist once the distributions are not smooth: if the distributions p, q are sums of delta functions, i.e. $p(x) = \sum_i p_i \delta(x - a_i)$ while $q(x) = \sum_i q_i \delta(x - b_i)$, then for generic choices of supports $\{a_i\}, \{b_j\}$, it is clear that no transport map T exists. For instance, if p is supported on one point and q is supported on two points, then there is no transport map T such that $T_* p = q$.

To better understand the Monge problem, it is convenient to first solve a relaxation known as the Kantorovich problem. In the Kantorovich problem, one searches for a positive measure π on $X \times X$ such that

(1) The pushforward of π to X is p, and the pushforward of π to Y is q [i.e. $\int_X dy \pi(x, y) = p(x)$ and $\int_X dx \pi(x, y) = q(y)$]; and
(2) The measure π minimizes

$$\mathsf{K}(\pi) = \int_{X \times X} dx dy \pi(x, y) c(x, y). \tag{2.20}$$

The interpretation of $dxdy\pi(x, y)$ is that it is the infinitesimal amount of mass at x which is transported to y. If we set $\pi_{x,y} = p(x)\delta(y - T(x))$ then it is clear that $\mathsf{K}(\pi) = \mathsf{M}(T)$. Thus, candidate solutions to the Monge problem give candidate solutions to the Kantorovich problem. However, the Kantorovich problem is much easier, as it is a problem in infinite dimensional *convex* optimization. Indeed, the function $\mathsf{K}(\pi)$ is a linear function on the convex cone of positive measures on $X \times X$ and the constraints arising from p and q are also linear. Discretizing this optimization problem yields a familiar finite-dimensional linear program: if $p(x) = \sum_i p_i \delta(x - a_i)$, $q(y) = \sum_j q_j \delta(y - b_j)$, and $\pi(x, y) = \sum_{i,j} \pi_{ij} \delta(x - a_i)\delta(y - b_j)$, then the Kantorovich problem immediately reduces to

$$\begin{aligned} &\text{Minimize} \sum_{i,j} \pi_{ij} c(a_i, b_j) \\ &\text{subject to} \quad \pi_{ij} \geq 0, \quad \sum_j \pi_{ij} = p_i, \quad \sum_i \pi_{ij} = q_j. \end{aligned} \tag{2.21}$$

Despite that fact that the Kantorovich problem is a relaxation of the Monge problem, in a large class of cases solutions to the Kantorovich problem actually arise from solutions to the Monge problem:

Theorem 2.1. If $p(x)$ and $q(x)$ are smooth functions having support on all of $\mathbb{R}^n$, then the Monge problem with $c(x, y) = |x - y|^2$ has a smooth solution; indeed, we have

$$T_i(x) = \partial_i f(x) \tag{2.22}$$

for some smooth convex function $f : \mathbb{R}^n \to \mathbb{R}$.

This result requires the development of a significant amount of mathematics: it follows from a combination of duality for the Kantorovich problem, Brenier's theorem [22], and Cafarelli's regularity theory [23] for solutions to the Monge-Ampère equation. To explain the proof of this theorem would take us too far afield, although [6, Chapters 2–4] gives a good introduction. We only note that once the existence of a function satisfying (2.22) is established by the duality theory, one concludes by the constraint (2.19) that f satisfies the equation

$$\det(\text{Hess } f(x)) = \frac{p(x)}{q(\nabla f(x))}, \tag{2.23}$$

which is the form of the Monge-Ampère equation that appears in this setting.

2. Wasserstein distance

Since we will be primarily interested in cases for which the space X is a metric space (and often a Riemannian metric space), we will henceforth denote the space by M. For the quadratic cost $c(x, y) = |x - y|^2$ where x, y are Cartesian coordinates on Euclidean space $M = \mathbb{R}^n$, the optimum value of $\mathsf{K}(\pi)$ in the Kantorovich problem is called the *Wasserstein-2 distance* $\mathcal{W}_2(p_1, p_2)$. (This is alternatively called the L^2-Wasserstein distance.) The distance can be written as

$$\mathcal{W}_2(p_1, p_2) := \left(\inf_{\pi \in \Gamma(p_1, p_2)} \int_{M \times M} dx dy \pi(x, y) |x - y|^2 \right)^{1/2} \tag{2.24}$$

where $\Gamma(p_1, p_2)$ is the space of probability distributions $\pi(x, y)$ on $M \times M$ such that $\int_M dy \pi(x, y) = p(x)$ and $\int_M dx \pi(x, y) = q(y)$. This metric distance on the space of probability distributions, and in particular various path integral generalizations of it, will play a central role in our analyses.

3. Otto calculus

To explain the connection between heat flow and optimal transport, we first recall how to view heat flow as a gradient flow with respect to the usual L^2 metric.

Heat flow as gradient flow of Dirichlet energy.—For a function $F : \mathcal{M} \to \mathbb{R}$ on a Riemannian manifold $\mathcal{M}$ with metric $\langle \cdot, \cdot \rangle$, the gradient of F at $x_0 \in \mathcal{M}$ is the vector $\nabla F(x_0)$ such that

$$\frac{d}{dt}F(x(t))\bigg|_{t=0} = \left\langle \nabla F(x_0), \frac{\partial}{\partial t}x(t)\right\rangle\bigg|_{t=0} \tag{2.25}$$

for every curve $x(t) \in \mathcal{M}$ with $x(0) = x_0$.

For our purposes, we let $\mathcal{M} = \mathrm{dens}(M)$ be the space probability densities on a manifold M, where we suppose M is equipped with a volume form dV. That is, $\mathrm{dens}(M)$ is an infinite-dimensional manifold defined by

$$\mathrm{dens}(M) := \left\{ p \in C^\infty(M) \,\middle|\, p \geq 0, \int dV p = 1 \right\}. \tag{2.26}$$

The tangent space at $p \in \mathrm{dens}(M)$ is

$$T_p\mathrm{dens}(M) = \left\{ \eta \in C^\infty(M) \,\middle|\, \int dV \eta = 0 \right\}. \tag{2.27}$$

We can equip each tangent space $T_p\mathrm{dens}(M)$ with a Riemannian metric

$$\langle \eta_1, \eta_2 \rangle_{L^2} = \int dV \eta_1 \eta_2. \tag{2.28}$$

This corresponds to the L^2 inner product on functions on M. Defining the Dirichlet energy functional as

$$\mathcal{E}[p] := \frac{1}{2}\int dV |\nabla p|^2, \tag{2.29}$$

we can compute its gradient with respect to the infinite-dimensional L^2 metric in (2.28) using (2.25). In particular, let $\rho(t)$ be a differentiable path through $\mathrm{dens}(M)$ such that $\rho(0) = p$. Then

$$\begin{aligned}\frac{d}{dt}\mathcal{E}[\rho(t)] &= \int dV \nabla\rho \cdot \nabla \frac{\partial}{\partial t}\rho, \\ &= -\int dV \Delta\rho \frac{\partial}{\partial t}\rho, \\ &= \left\langle -\Delta\rho, \frac{\partial}{\partial t}\rho \right\rangle_{L^2}.\end{aligned} \tag{2.30}$$

Evaluating the above at $t = 0$ and comparing with (2.25), we read off that

$$\nabla_{L^2}\mathcal{E}[p] = -\Delta p. \tag{2.31}$$

It follows that the heat equation $\frac{\partial}{\partial t}p = \Delta p$ is the negative gradient flow of the Dirichlet energy functional $\mathcal{E}$, namely

$$\frac{\partial}{\partial t}p(x,t) = -\nabla_{L^2}\mathcal{E}[p(x,t)]. \tag{2.32}$$

This in fact implies that the Dirichlet energy monotonically decreases along the heat flow.

Wasserstein distance and the gradient flow of entropy.—Another monotone for the heat flow is given by the differential entropy

$$S[p] := -\int dV p \log(p). \tag{2.33}$$

We will have more to say about this quantity in Sec. III B. By analogy with (2.32) above, we might ask if there is any Riemannian metric g on $\mathrm{dens}(M)$ such that the heat equation can be written as $\frac{\partial}{\partial t}p = \nabla_g S[p]$? In other words, is there some (natural) metric on the space of probability distributions for which the heat equation is the gradient flow of the differential entropy?

Remarkably, the answer yes—this was discovered by Otto [21] and widely exploited by subsequent researchers in partial differential equations and probability theory. In fact, there are a large collection of entropylike monotones $\tilde{S}$ which have associated metrics $\tilde{g}$ on $\mathrm{dens}(M)$ such that the heat equation can be written as $\frac{\partial}{\partial t}p = \nabla_{\tilde{g}}\tilde{S}[p]$. All of these metrics have deep connections to optimal transport. Since we will be interested in the particular case of the differential entropy, we will not discuss these related entropic gradient flow formulations here.

We now turn to constructing the metric g on $\mathrm{dens}(M)$ such that $\frac{\partial}{\partial t}p = \nabla_g S[p]$. To write the metric in the most transparent way, an isomorphism of the tangent space $T_p\mathrm{dens}(M)$ is required. Given a tangent vector $\eta \in T_p\mathrm{dens}(M)$, we can solve for a $\bar{\eta}$ satisfying

$$\nabla \cdot (p\nabla\bar{\eta}) = \eta. \tag{2.34}$$

The solution is unique up to addition of a constant, and so we get an identification $\eta \leftrightarrow \bar{\eta}$ which we notate by the isomorphism

$$\begin{aligned}T_p\mathrm{dens}(M) &\simeq \overline{T_p\mathrm{dens}(M)} \\ &:= \{\bar{\eta} \in C^\infty(M)\}/\{\text{constants}\}.\end{aligned} \tag{2.35}$$

Using this identification we define the Riemannian metric

$$\begin{aligned}\langle \eta_1, \eta_2 \rangle_{\mathcal{W}_2} &:= \int dV p \nabla\bar{\eta}_1 \cdot \nabla\bar{\eta}_2 = -\int dV \eta_1 \bar{\eta}_2 \\ &= -\int dV \bar{\eta}_1 \eta_2,\end{aligned} \tag{2.36}$$

where the last two equalities can be checked via integration by parts. This metric is in fact the infinitesimal form of the Wasserstein-2 distance $\mathcal{W}_2$. A rigorous argument establishing this fact is given in [24] [Lemma 4.3]; we will explain the heuristic connection in Appendix A.

Now let us show that $\nabla_{\mathcal{W}_2}S[p] = \Delta p$. Let $\rho(t)$ be a path through $\mathrm{dens}(M)$ with $\rho(0) = p$, and define $\eta := \frac{d}{dt}\rho(t)|_{t=0}$

which is definitionally an element of $T_p\mathrm{dens}(M)$. Let $\bar{\eta}$ be the corresponding solution to (2.34). Then we compute

$$\begin{aligned}\frac{d}{dt}S[\rho(t)]|_{t=0} &= -\int dV\eta(\log p+1),\\ &= -\int dV\nabla\cdot(p\nabla\bar{\eta})(\log p+1),\\ &= \int dV\nabla\bar{\eta}\cdot\nabla p,\\ &= -\int dV\bar{\eta}\cdot\Delta p,\\ &= \langle\Delta p,\eta\rangle_{\mathcal{W}_2},\\ &= \left\langle\Delta p,\frac{\partial}{\partial t}\rho\right\rangle_{\mathcal{W}_2}\Big|_{t=0},\end{aligned} \tag{2.37}$$

and so comparing with (2.25) we indeed find

$$\nabla_{\mathcal{W}_2}S[p]=\Delta p. \tag{2.38}$$

Then the heat equation can be written as

$$\frac{\partial}{\partial t}p(x,t)=\nabla_{\mathcal{W}_2}S[p]. \tag{2.39}$$

Thus, the heat flow is the gradient flow of the differential entropy (2.33) with respect to the Wasserstein-2 metric. While it was known to Gibbs that entropy is a heat flow monotone, the above equation clarifies that in fact heat flow is *completely governed* by the entropy, with optimal transport playing a central role in this formulation.

4. A guide to further literature

In the rest of this paper, we will exploit formal, infinite-dimensional analogs of the optimal-transport formulation of heat flow to study the renormalization group. We expect that that there are further profitable connections to be made between the rich mathematics of optimal transport and the structure of the renormalization group, and we view the present work as an initial study.

The lecture notes [6] are a very readable mathematical introduction to the subject of optimal transport, and the original paper [21] remains full of geometric insight. The papers [22,23,25] mentioned above are all fundamental. The book [6] covers connections to Ricci curvature, while the review article [26] summarizes applications in partial differential equation (PDE) and applied mathematics. The logarithmic Sobolev inequalities proven in [27] were reproven using optimal transport in [28] and had a dramatic impact on probability theory; they were originally motivated by problems in constructive quantum field theory and so it is not surprising that the ideas should come full circle. Some very recent applications of Polchinski's equation to constructive quantum field theory can be found in [29,30].

We note also that ideas around the logarithmic Sobolev inequalities together with the fact that Ricci flow is renormalization group flow for a σ-model was a stated motivation for Perelman's work on the Poincarè conjecture [31]; following this idea, McCann and Topping [32] began an ongoing research program founding Ricci flow in ideas based on optimal transport.

III. RG FLOW AS AN OPTIMAL TRANSPORT GRADIENT FLOW

A. Deriving the optimal transport gradient flow equation for RG

Since Polchinski's equation (2.7) is a special case of the Wegner-Morris flow equation (2.10), we find it prudent to derive our optimal transport equation for the latter. Suppose we intend to flow a Euclidean field theory with probability functional $P_\Lambda[\phi]=e^{-S_\Lambda[\phi]}/Z_{P,\Lambda}$. Recall the Wegner-Morris flow equation

$$-\Lambda\frac{d}{d\Lambda}P_\Lambda[\phi]=\int d^dx\frac{\delta}{\delta\phi(x)}(\Psi_\Lambda[\phi,x]P_\Lambda[\phi]),$$

where will adopt the functional forms in (2.14) and (2.15) for the reparametrization kernel Ψ_Λ, namely

$$\begin{aligned}\Psi_\Lambda[\phi,x] &= -\int d^dy\frac{1}{2}\dot{C}_\Lambda(x-y)\frac{\delta\Sigma_\Lambda[\phi]}{\delta\phi(y)},\\ \Sigma_\Lambda[\phi] &= S_\Lambda[\phi]-2\hat{S}_\Lambda[\phi],\end{aligned}$$

where $\dot{C}_\Lambda(x-y)\geq 0$. Now we define a Riemannian metric on the tangent space to the space of probability functionals which we will later explain is the infinitesimal version of a functional $\mathcal{W}_2$ metric. We let

$$\begin{aligned}\langle\delta P_1[\phi],\delta P_2[\phi]\rangle_{\mathcal{W}_2} = \frac{1}{2}\int[d\phi]P[\phi]\int d^dxd^dy\dot{C}_\Lambda(x-y)\\ \times\frac{\delta\Phi_1[\phi]}{\delta\phi(x)}\frac{\delta\Phi_2[\phi]}{\delta\phi(y)},\end{aligned} \tag{3.1}$$

where we define $\Phi_i[\phi]$ for $i=1,2$ via the functional differential equations

$$\delta P_i[\phi]-\frac{1}{2}\int d^dxd^dy\dot{C}_\Lambda(x-y)\frac{\delta}{\delta\phi(x)}\left(P[\phi]\frac{\delta\Phi_i[\phi]}{\delta\phi(y)}\right)=0. \tag{3.2}$$

Analogous to the finite-dimensional heat flow setting, the Φ_is are only specified by the above equation up to additive functions not depending on ϕ. Note that since $\dot{C}_\Lambda(x-y)\geq 0$, the norm induced by the metric is automatically greater than or equal to zero. Similar to Otto's calculation we can perform an integration by parts in (3.1) to obtain the more compact expressions

$$\langle \delta P_1[\phi], \delta P_2[\phi] \rangle_{\mathcal{W}_2} = -\int [d\phi] \delta P_1[\phi] \Phi_2[\phi] = -\int [d\phi] \Phi_1[\phi] \delta P_2[\phi]. \tag{3.3}$$

Coopting the results of Otto [24] and generalizing them appropriately to our setting, we have that our metric is the infinitesimal form of the distance

$$\mathcal{W}_2(P_1, P_2) := \left(\inf_{\Pi \in \Gamma(P_1, P_2)} 2 \int [d\phi_1][d\phi_2] \Pi[\phi_1, \phi_2] \int d^d x d^d y \dot{C}_\Lambda^{-1}(x,y)(\phi_1(x) - \phi_2(x))(\phi_1(y) - \phi_2(y)) \right)^{1/2}, \tag{3.4}$$

where $\Gamma(P_1, P_2)$ is the space of probability functionals $\Pi[\phi_1, \phi_2]$ such that $\int [d\phi_2] \Pi[\phi_1, \phi_2] = P[\phi_1]$ and $\int [d\phi_1] \Pi[\phi_1, \phi_2] = P_2[\phi_2]$. Above $\dot{C}_\Lambda^{-1}(x,y)$ is the inverse of the kernel $\dot{C}_\Lambda(x,y)$ in the sense that $\int d^d z \dot{C}_\Lambda^{-1}(x,z) \dot{C}_\Lambda(z,y) = \delta^d(x-y)$; since in our setting $\dot{C}_\Lambda(x,y) = \dot{C}_\Lambda(x-y)$, in momentum space the kernel $\dot{C}_\Lambda(p^2)$ has as its inverse $\dot{C}_\Lambda^{-1}(p^2) = 1/\dot{C}_\Lambda(p^2)$. The distance $\mathcal{W}_2(P_1, P_2)$ represents the minimum cost of "transporting" P_1 into P_2 (or vice versa), where the cost is given by an L^2 penalty on rearranging field degrees of freedom away from the spatial scale $\ell \sim 1/\Lambda$.

We are now almost ready to state our main result and then subsequently derive it. Define the probability functional

$$Q_\Lambda[\phi] := \frac{e^{-2\hat{S}_\Lambda[\phi]}}{Z_{Q,\Lambda}}, \tag{3.5}$$

where $Z_{Q,\Lambda} = \int [d\phi] e^{-2\hat{S}_\Lambda[\phi]}$ and let the functional relative entropy be

$$S(P[\phi] \| Q[\phi]) := \int [d\phi] P[\phi] \log\left(\frac{P[\phi]}{Q[\phi]}\right). \tag{3.6}$$

Then we have the remarkable formula

$$-\Lambda \frac{d}{d\Lambda} P_\Lambda[\phi] = -\nabla_{\mathcal{W}_2} S(P_\Lambda[\phi] \| Q_\Lambda[\phi]), \tag{3.7}$$

which is equivalent to the Wegner-Morris flow equation (2.10). To establish this connection, we need to show that $-\nabla_{\mathcal{W}_2} S(P_\Lambda[\phi] \| Q_\Lambda[\phi])$ equals $\int d^d x \frac{\delta}{\delta\phi(x)} (\Psi_\Lambda[\phi, x] P_\Lambda[\phi])$.

For any $\mathcal{F}[P[\phi]]$ which takes probability functionals to the real numbers, the differential-geometric definition of the gradient $\nabla_{\mathcal{W}_2} \mathcal{F}[P]$ is given by

$$\langle \nabla_{\mathcal{W}_2} \mathcal{F}[P], \delta P \rangle_{\mathcal{W}_2} = \int [d\phi] \frac{\delta \mathcal{F}[P]}{\delta P} \delta P[\phi]. \tag{3.8}$$

A slightly unusual feature of the right-hand side is that $\frac{\delta \mathcal{F}[P]}{\delta P}$ is not an ordinary functional derivative but rather a functional-of-a-functional derivative, i.e. a derivative with respect to the functional $P[\phi]$. In our case, we choose $\mathcal{F}[P_\Lambda] := S(P_\Lambda \| Q_\Lambda)$; then computing the right-hand side of (3.8) we obtain

$$\begin{aligned}
&\int [d\phi] (\log P_\Lambda[\phi] + 1 - \log Q_\Lambda[\phi]) \delta P[\phi] \\
&= \int [d\phi] (-S_\Lambda[\phi] - \log Z_P + 1 + 2\hat{S}_\Lambda[\phi] + \log Z_Q) \delta P[\phi], \\
&= \int [d\phi] (-S_\Lambda[\phi] + 2\hat{S}_\Lambda[\phi]) \delta P[\phi], \\
&= -\int [d\phi] \Sigma_\Lambda[\phi] \delta P[\phi].
\end{aligned} \tag{3.9}$$

In going from the first to second line we used $\int [d\phi] \delta P[\phi] = 0$ since this is a property of elements of the tangent space to probability functionals so that $\int [d\phi](P[\phi] + \delta P[\phi]) = 1$. Next we use (3.2) to rewrite δP in terms of a Φ field, giving us

$$-\frac{1}{2} \int [d\phi] \Sigma_\Lambda[\phi] \int d^d x d^d y \dot{C}_\Lambda(x-y) \frac{\delta}{\delta\phi(x)} \left(P_\Lambda[\phi] \frac{\delta \Phi[\phi]}{\delta\phi(y)} \right). \tag{3.10}$$

Integrating by parts twice in the functional ϕ derivatives, we obtain

$$\begin{aligned}
&-\int [d\phi] \int d^d x \frac{\delta}{\delta\phi(x)} \\
&\times \left(\int d^d y \frac{1}{2} \dot{C}_\Lambda(x-y) \frac{\delta \Sigma_\Lambda[\phi]}{\delta\phi(y)} P_\Lambda[\phi] \right) \Phi[\phi] \\
&= \int [d\phi] \int d^d x \frac{\delta}{\delta\phi(x)} (\Psi_\Lambda[\phi, x] P_\Lambda[\phi]) \Phi[\phi], \\
&= \left\langle -\int d^d x \frac{\delta}{\delta\phi(x)} (\Psi_\Lambda[\phi, x] P_\Lambda[\phi]), \delta P \right\rangle_{\mathcal{W}_2},
\end{aligned} \tag{3.11}$$

where in the last line we have used (3.3). Comparing with (3.8) this establishes

$$-\nabla_{\mathcal{W}_2} S(P_\Lambda[\phi]\|Q_\Lambda[\phi]) = \int d^d x \frac{\delta}{\delta\phi(x)}(\Psi_\Lambda[\phi,x]P_\Lambda[\phi]), \tag{3.12}$$

which implies our main result (3.7).

B. Comments and interpretation

Our result (3.7) provides a new way of thinking about the renormalization group, and elucidates some key technical aspects of Polchinski's equation and the Wegner-Morris flow equation more broadly. First let us discuss (3.7) itself.

The relative entropy $S(P\|Q)$, also called the Kullback-Leibler divergence, is a core object in information theory which provides a measure of similarity between two probability distributions P, Q [33]. While it is not a metric distance (for instance, it is not symmetric between P and Q and does not satisfy the triangle inequality), it is positive and enjoys a host of other properties; a useful discussion aimed for physicists is [34]. Heuristically, the relative entropy tells us how good Q is as a proxy for P. For instance, the relative entropy quantifies how much additional memory is required to compress a list of samples from P if we are only given just enough memory to optimally compress a list of as many samples from Q. There have been other works on RG flow which have leveraged the relative entropy [35–45], albeit in a manner which does not involve optimal transport.

A particular conceptual feature of the relative entropy is worth commenting on. If we have a discrete probability distribution p_i, then its entropy is simply $-\sum_i p_i \log(p_i)$. Passing to the continuum via $p_i \to dx p(x)$, the entropy becomes $-\int dx p(x)\log(dx p(x))$. The dx inside the logarithm is somewhat pathological, and reflects that the strict continuum limit of the entropy is ill-defined. Relatedly, if we give dx units of length so that $p(x)$ has units of inverse length, then the quantity inside the logarithm must be dimensionless, which is achieved by $\log(dx p(x))$. To cure the issue of a dx inside the logarithm, the continuum entropy is obliged to have an alternative defining formula which is partially divorced from its discrete version. A common option is $S[p] = -\int dx p(x)\log(p(x))$, which is called the differential entropy. In the differential entropy, the $\log(p(x))$ should be thought of as $\log(a p(x))$ for $a = 1$, where a has "units" of length. Notably, the relative entropy is free of the aforementioned issue. For suppose we consider $-\sum_i p_i \log(q_i)$ for some second discrete probability distribution q_i and pass to the continuum limit in the same way to get $-\int dx p(x)\log(dx q(x))$. Subtracting this from $-\int dx p(x)\log(dx p(x))$, we obtain minus the relative entropy

$$-S(p\|q) = -\int dx p(x)\log\left(\frac{p(x)}{q(x)}\right), \tag{3.13}$$

where in effect the unwanted dxs in the log have canceled out. As such, we can think of minus the relative entropy as a well-defined and meaningful replacement for the continuum entropy.

One interpretation of our result (3.7) is that the RG flow of the probability functional P_Λ seeks to minimize the relative entropy between P_Λ and Q_Λ according to the appropriate $\mathcal{W}_2$ gradient. Minimizing the relative entropy can be viewed as a proxy for maximizing the entropy of P_Λ, in light of the discussion in the preceding paragraph. This makes intuitive sense: as we coarse grain due to RG flow, there is a form of entropy production. But what is more striking from (3.7) is that the entropy production is *precisely* what determines the flow itself. An alternative formulation of this statement is provided in Sec. VI where we develop a variational formula for RG flow.

An interesting special case of (3.7) is Polchinski's equation for a free scalar field, corresponding to

$$-\Lambda\frac{d}{d\Lambda}P_\Lambda^{\text{free}}[\phi] = -\nabla_{\mathcal{W}_2} S(P_\Lambda^{\text{free}}[\phi]\|Q_\Lambda^{\text{free}}[\phi]), \tag{3.14}$$

with $P_\Lambda^{\text{free}}[\phi] = e^{-S_{\text{free},\Lambda}[\phi]}/Z_{P,\Lambda}$ and $Q_\Lambda^{\text{free}}[\phi] = e^{-2S_{\text{free},\Lambda}[\phi]}/Z_{Q,\Lambda}$, where we note the factor of 2 in the exponent. Using the identities

$$\begin{aligned}-\nabla_{\mathcal{W}_2} S(P_\Lambda^{\text{free}}[\phi]\|Q_\Lambda^{\text{free}}[\phi]) &= -\nabla_{\mathcal{W}_2} S(P_\Lambda^{\text{free}}[\phi]\|(P_\Lambda^{\text{free}}[\phi])^2)\\ &= \nabla_{\mathcal{W}_2} S(P_\Lambda^{\text{free}}[\phi]),\end{aligned} \tag{3.15}$$

where $S(P) = -\int [d\phi]P[\phi]\log P[\phi]$ is a functional analog of the differential entropy, we find

$$-\Lambda\frac{d}{d\Lambda}P_\Lambda^{\text{free}}[\phi] = \nabla_{\mathcal{W}_2} S(P_\Lambda^{\text{free}}[\phi]). \tag{3.16}$$

Thus the free scalar field flows exactly according to its differential entropy.

Another feature of (3.7) is that it explains the role of "seed action" $\hat{S}_\Lambda[\phi]$ in (2.15). In particular, the seed action (and its conventional prefactor of 2) provides us with $Q_\Lambda[\phi] = e^{-2\hat{S}_\Lambda[\phi]}/Z_{Q,\Lambda}$ as per (3.5), which is the baseline distribution in the relative entropy $S(P_\Lambda\|Q_\Lambda)$ appearing in the gradient flow. Indeed, the definition of the functional $\mathcal{W}_2$ distance together with Q_Λ define our choice of RG scheme. We emphasize that the seed action $\hat{S}_\Lambda[\phi]$ has a prescribed Λ dependence, and does not itself need to satisfy a flow equation.

Finally, we comment on the meaning of $\Sigma_\Lambda[\phi]$. Suggestively rewriting it as

$$\begin{aligned}\Sigma_\Lambda[\phi] = &-\log(P_\Lambda[\phi]) + \log(Q_\Lambda[\phi]) - \log(Z_{P,\Lambda})\\ &+ \log(Z_{Q,\Lambda}),\end{aligned} \tag{3.17}$$

we observe that $\Sigma_\Lambda[\phi]$ ultimately enters into our formulas only through its functional derivative $\frac{\delta\Sigma_\Lambda[\phi]}{\delta\phi}$. As such, we are free to redefine $\Sigma_\Lambda[\phi]$ by adding ϕ-independent terms. Thus, subtracting the constant terms off of (3.17) and combining the residual logarithms, we can replace $\Sigma_\Lambda[\phi]$ in (2.10), (2.14) with

$$\tilde{\Sigma}_\Lambda[\phi] = -\log\left(\frac{P_\Lambda[\phi]}{Q_\Lambda[\phi]}\right), \tag{3.18}$$

where the tilde reminds us that we have made a modification (albeit an innocuous one) to the original definition without changing the resulting Wegner-Morris flow equation. Notice that this new quantity $\tilde{\Sigma}_\Lambda[\phi]$ is information-theoretically natural: it is minus the log likelihood ratio between P_Λ and Q_Λ, and so we can write

$$S(P_\Lambda\|Q_\Lambda) = -\int [d\phi] P_\Lambda[\phi]\tilde{\Sigma}_\Lambda[\phi]. \tag{3.19}$$

Accordingly, we have repackaged the major ingredients in the Wegner-Morris flow equation (and Polchinski's equation as a special case) in terms of information-theoretic quantities.

IV. RG MONOTONES

In this section we derive a nonperturbative RG monotone using our optimal transport flow equation in (3.7). There have been previous attempts at formulating RG monotones using the ERG framework but this has only been successful in the local potential approximation, essentially where we ignore higher-derivative contributions to the action [46–49]. By contrast, our RG monotone holds without any approximations.

Our proposed monotone for a P_Λ solving (3.7) is formally given by

$$M_\Lambda(P_\Lambda) \coloneqq S(P_\Lambda\|Q_\Lambda) - \log(Z_{Q,\Lambda}), \tag{4.1}$$

under the assumption that $Q_\Lambda[\phi] = e^{-S_{Q,\Lambda}[\phi]}/Z_{Q,\Lambda}$ for

$$S_{Q,\Lambda}[\phi] = C\int \frac{d^dp}{(2\pi)^d}\hat{K}_\Lambda^{-1}(p^2)G^{-1}(p^2)\phi(p)\phi(-p). \tag{4.2}$$

Here $\hat{K}_\Lambda(p^2)$ is a smooth cutoff function which need not equal $K_\Lambda(p^2)$, and $G(p^2)$ is the Green's function of some positive semidefinite elliptic differential operator [e.g. $G(p^2) = 1/(p^2+m^2)$]. Accordingly, our monotone pertains to Polchinski's equation, as well as more generally the Wegner-Morris flow equation with a quadratic seed action.

Due to interesting subtleties with orders of limits and divergences, in Sec. IV B [see in particular (4.24)] we will introduce a regulated version of $M_\Lambda(P_\Lambda)$. We will show below that the quantity (4.1) is formally divergent, but it can be regularized in a way that is independent of the renormalization scheme. There is an extensive discussion in Sec. IV B which provides appropriate context. The proof of monotonicity below is unaffected.

A. Proof of monotonicity

Let us establish the monotonicity of the monotone. We have

$$-\Lambda\frac{d}{d\Lambda}M_\Lambda(P_\Lambda) = -\int [d\phi]\Lambda\frac{\partial P_\Lambda}{\partial\Lambda}(\log(P_\Lambda) - \log(e^{-S_{Q,\Lambda}})) - \int [d\phi]\left(\Lambda\frac{\partial P_\Lambda}{\partial\Lambda} + P_\Lambda\frac{\partial S_{Q,\Lambda}}{\partial\Lambda}\right). \tag{4.3}$$

Here we are differentiating under the integral sign by bringing $\Lambda\frac{d}{d\Lambda}$ into integrand of the functional integral. This has some subtleties related to regularization of $M_\Lambda(P_\Lambda)$ which we treat in detail in Sec. IV B, but indeed we will find that integrating under the integral sign is a good prescription. Using (3.7) and dropping total derivative terms, we find

$$\int [d\phi]\int d^dx\frac{\delta(\Psi_\Lambda P_\Lambda)}{\delta\phi(x)}(\log(P_\Lambda) - \log(e^{-S_{Q,\Lambda}})) - \int [d\phi]P_\Lambda\Lambda\frac{\partial S_{Q,\Lambda}}{\partial\Lambda}. \tag{4.4}$$

Integrating by parts on the first term, we obtain

$$\frac{1}{2}\int [d\phi]P_\Lambda[\phi]\int d^dxd^dy\dot{C}_\Lambda(x-y)\frac{\delta\Sigma_\Lambda}{\delta\phi(x)}\frac{\delta\Sigma_\Lambda}{\delta\phi(y)} - \int [d\phi]P_\Lambda\Lambda\frac{\partial S_{Q,\Lambda}}{\partial\Lambda}. \tag{4.5}$$

The first term[6] is manifestly positive semidefinite since $\dot{C}_\Lambda(x-y) \geq 0$. For the second term, we have

$$-\int [d\phi]P_\Lambda\Lambda\frac{\partial S_{Q,\Lambda}}{\partial\Lambda} = -\int\frac{d^dp}{(2\pi)^d}\Lambda\frac{\partial\hat{K}_\Lambda^{-1}(p^2)}{\partial\Lambda}G^{-1}(p^2)\langle\phi(p)\phi(-p)\rangle_{P_\Lambda}. \tag{4.6}$$

Since $-\Lambda\frac{\partial\hat{K}^{-1}(p^2/\Lambda^2)}{\partial\Lambda} \geq 0$ and $\langle\phi(p)\phi(-p)\rangle_P \geq 0$, the entire quantity is greater than or equal to zero. Accordingly, we have established that

[6]It can also be written as functional generalization of the relative Fisher information between P_Λ and Q_Λ with background metric $\langle F[\phi(x)], G[\phi(y)]\rangle = \frac{1}{2}\int d^dxd^dy\dot{C}(x-y)F[\phi(x)]G[\phi(y)]$.

$$-\Lambda\frac{d}{d\Lambda}M_\Lambda(P_\Lambda)\geq 0 \tag{4.7}$$

and so $M_\Lambda(P_\Lambda)$ is an RG monotone.

A slight surprise about the definition of the monotone (4.1) is the presence of the $-\log(Z_{Q,\Lambda})$. The necessity of this term can be understood as follows. Suppose we did not include $-\log(Z_{Q,\Lambda})$ in the monotone, so that $-\Lambda\frac{d}{d\Lambda}M_\Lambda(P_\Lambda)$ is just the relative entropy. This would affect the left-hand side of (4.7) by adding the term

$$-\Lambda\frac{\partial\log(Z_{Q,\Lambda})}{\partial\Lambda}=-\Lambda\frac{1}{Z_{Q,\Lambda}}\frac{\partial Z_{Q,\Lambda}}{\partial\Lambda}. \tag{4.8}$$

Unfortunately this term is less than zero, and so can interfere with the bound (4.7) if we include it. In particular, as we raise the cutoff Λ, more modes are introduced in the $S_{Q,\Lambda}$ action, with variances $\sigma^2\simeq 1/(p^2+m^2)$ for $p\sim\Lambda$; the variances of these modes were formerly extremely small before we raised the cutoff. Accordingly, the partition function $Z_{Q,\Lambda}$ will increase when we raise the cutoff, and so $\Lambda\frac{\partial Z_{Q,\Lambda}}{\partial\Lambda}\geq 0$. This is why we have elected to define our monotone to avoid this issue.

Note, however, that if we had a hard cutoff instead of a soft cutoff, the story would be different. In the hard cutoff setting, introducing more modes by raising the cutoff would cause $\Lambda\frac{\partial Z_Q}{\partial\Lambda}\leq 0$, and render (4.8) to be positive. However, other subtleties with ERG in the hard cutoff setting pertaining to changing the domain of path integration dissuade us from pursuing this direction at present.

Having defined a nonperturbative RG monotone $M_\Lambda(P_\Lambda)$ for quantum field theories, it is natural to inquire about the finiteness of $M_\Lambda(P_\Lambda)$. This will become clearer when we compute some examples below, but here we overview some general structure. In our examples we will find that

$$-\Lambda\frac{d}{d\Lambda}M_\Lambda(P_\Lambda)=\delta^d(0)c_1(\Lambda), \tag{4.9}$$

where $c_1(\Lambda)$ is a finite quantity. Here the $\delta^d(0)$ divergence comes from momentum space contact terms. If we considered a field theory on, say, a torus where the momenta range over a lattice, then the $\delta^d(0)$ would be rendered finite. Since the $\delta^d(0)$ is multiplicative on the right-hand side of (4.9) it is essentially innocuous: the positivity of $-\Lambda\frac{d}{d\Lambda}M_\Lambda(P_\Lambda)$ implies

$$c_1(\Lambda)\geq 0. \tag{4.10}$$

Then an appropriate antiderivative of $c_1(\Lambda)$, namely a $C_1(\Lambda)$ satisfying $-\Lambda\frac{d}{d\Lambda}C_1(\Lambda)=c_1(\Lambda)$, is evidently a finite RG monotone since

$$-\Lambda\frac{d}{d\Lambda}C_1(\Lambda)\geq 0. \tag{4.11}$$

Our discussion above pertained to $-\Lambda\frac{d}{d\Lambda}M_\Lambda(P_\Lambda)$ instead of $M_\Lambda(P_\Lambda)$ itself. As we will see in in Sec. IV B below, there are some subtleties in computing $M_\Lambda(P_\Lambda)$ directly. It can be done, however, with sufficient care. Nonetheless, the derivative $-\Lambda\frac{d}{d\Lambda}M_\Lambda(P_\Lambda)$ can be computed rather directly using the formula (4.5) which automatically accounts for subtleties in the definition of $M_\Lambda(P_\Lambda)$.

B. Differentiating under the functional integral and regularization

In our derivation of the monotonicity of $M_\Lambda(P_\Lambda)$ above, we differentiated under the integral sign in (4.3). This interchange of limits is particularly subtle in our path integral setting as we will now show. Let us start with an illuminating example before turning to generalities.

1. Order of limits in the setting of free field theory

Consider Polchinski's equation in (2.7); a solution to this is the free probability distribution, given by $P_\Lambda[\phi]=\frac{1}{Z_\Lambda}\exp\left(-\frac{1}{2}\int\frac{d^dp}{(2\pi)^d}\phi(p)\phi(-p)(p^2+m^2)K_\Lambda^{-1}(p^2)\right)$. Plugging this into our monotone $M_\Lambda(P_\Lambda)$ defined in (4.1), there are terms proportional to

$$\begin{aligned}&\int[d\phi]P_\Lambda^{\text{free}}[\phi]S_{\text{free},\Lambda}[\phi]\\&=\frac{1}{2Z_{P,\Lambda}}\int[d\phi]e^{-\frac{1}{2}\int\frac{d^dp}{(2\pi)^d}\phi(p)\phi(-p)(p^2+m^2)K_\Lambda^{-1}(p^2)}\\&\quad\times\int\frac{d^dp}{(2\pi)^d}\phi(p)\phi(-p)(p^2+m^2)K_\Lambda^{-1}(p^2),\\&=\frac{1}{2}\int d^dp\,\delta^d(0).\end{aligned} \tag{4.12}$$

This is related to the fact that in the more ordinary n-dimensional integral setting we have $\int d^nx\frac{\det(A)^{1/2}}{(2\pi)^{n/2}}\times e^{-\frac{1}{2}x\cdot A\cdot x}(\frac{1}{2}x\cdot A\cdot x)=\frac{1}{2}$, except that in the functional setting we have a (momentum space) contact term $\delta^d(0)$ and a residual $\int d^dp$ integral at the end. From (4.12) we infer that

$$-\Lambda\frac{d}{d\Lambda}\int[d\phi]P_\Lambda[\phi]S_{\text{free},\Lambda}[\phi]=0. \tag{4.13}$$

But now let us perform the computation of (4.13) another way by differentiating under the integral sign. In this setting we have

$$-\int [d\phi]\Lambda\frac{\partial}{\partial\Lambda}(P^{\text{free}}_{\Lambda}[\phi]S_{\text{free},\Lambda}[\phi]) = -\int [d\phi]\Lambda\frac{\partial P^{\text{free}}_{\Lambda}[\phi]}{\partial\Lambda}S_{\text{free},\Lambda}[\phi] - \int [d\phi]P^{\text{free}}_{\Lambda}[\phi]\Lambda\frac{\partial S_{\text{free},\Lambda}[\phi]}{\partial\Lambda}. \tag{4.14}$$

The first term on the right-hand side is

$$\begin{aligned}-\int [d\phi]\Lambda\frac{\partial P^{\text{free}}_{\Lambda}[\phi]}{\partial\Lambda}S_{\text{free},\Lambda}[\phi] &= \frac{1}{4Z_{P,\Lambda}}\int [d\phi]e^{-\frac{1}{2}\int\frac{d^dp}{(2\pi)^d}\phi(p)\phi(-p)(p^2+m^2)K^{-1}_{\Lambda}(p^2)}\int\frac{d^dp}{(2\pi)^d}\phi(p)\phi(-p)(p^2+m^2)K^{-1}_{\Lambda}(p^2)\\ &\quad\times\int\frac{d^dq}{(2\pi)^d}\phi(q)\phi(-q)(q^2+m^2)\Lambda\frac{\partial K^{-1}_{\Lambda}(q^2)}{\partial\Lambda} + \Lambda\frac{\partial\log Z_{P,\Lambda}}{\partial\Lambda}\int [d\phi]P^{\text{free}}_{\Lambda}[\phi]S_{\text{free},\Lambda}[\phi],\\ &= \left(\frac{1}{2}\int d^dp\delta^d(0)\right)\left(\Lambda\frac{\partial\log Z_{P,\Lambda}}{\partial\Lambda} - \frac{1}{2}\delta^d(0)\int d^dp\Lambda\frac{\partial\log K_{\Lambda}(p^2)}{\partial\Lambda}\right)\\ &\quad-\frac{1}{2}\delta^d(0)\int d^dp\Lambda\frac{\partial\log K_{\Lambda}(p^2)}{\partial\Lambda}.\end{aligned} \tag{4.15}$$

To further simplify, we observe that for an infinitesimal perturbation change of scale any $P_\Lambda[\phi]$ changes as $P_{\Lambda-\delta\Lambda}[\phi] = P_\Lambda[\phi] - \delta\Lambda\frac{\partial P_\Lambda[\phi]}{\partial\Lambda}$. Then integrating both sides with respect to ϕ and using the normalization of the probability functional we find

$$\int [d\phi]\Lambda\frac{\partial P_{\Lambda}[\phi]}{\partial\Lambda} = 0. \tag{4.16}$$

In the free setting,

$$\begin{aligned}&\int [d\phi]\Lambda\frac{\partial P^{\text{free}}_{\Lambda}[\phi]}{\partial\Lambda}\\ &=\frac{1}{2}\delta^d(0)\int d^dp\Lambda\frac{\partial\log K_{\Lambda}(p^2)}{\partial\Lambda} - \Lambda\frac{\partial\log Z_{P,\Lambda}}{\partial\Lambda} = 0\end{aligned} \tag{4.17}$$

which implies

$$\Lambda\frac{\partial\log Z_{P,\Lambda}}{\partial\Lambda} = \frac{1}{2}\delta^d(0)\int d^dp\Lambda\frac{\partial\log K_{\Lambda}(p^2)}{\partial\Lambda}. \tag{4.18}$$

Plugging this into (4.15) we see that there is a helpful cancellation which leaves us with

$$\begin{aligned}&-\int [d\phi]\Lambda\frac{\partial P^{\text{free}}_{\Lambda}[\phi]}{\partial\Lambda}S_{\text{free},\Lambda}[\phi]\\ &\quad= -\frac{1}{2}\delta^d(0)\int d^dp\Lambda\frac{\partial\log K_{\Lambda}(p^2)}{\partial\Lambda}.\end{aligned} \tag{4.19}$$

Turning to the second term on the right-hand side of (4.14), a less elaborate computation yields

$$-\int [d\phi]P^{\text{free}}_{\Lambda}[\phi]\Lambda\frac{\partial S_{\text{free},\Lambda}[\phi]}{\partial\Lambda} = \delta^d(0)\int d^dp\Lambda\frac{\partial\log K_{\Lambda}(p^2)}{\partial\Lambda}. \tag{4.20}$$

Plugging (4.19) and (4.20) into (4.14), we finally find

$$\begin{aligned}&-\int [d\phi]\Lambda\frac{\partial}{\partial\Lambda}(P_{\Lambda}[\phi]S_{\text{free},\Lambda}[\phi])\\ &\quad=\frac{1}{2}\delta^d(0)\int d^dp\Lambda\frac{\partial\log K_{\Lambda}(p^2)}{\partial\Lambda}.\end{aligned} \tag{4.21}$$

Comparing (4.21) with (4.13), we see that surprisingly

$$\begin{aligned}&-\Lambda\frac{d}{d\Lambda}\int [d\phi]P_{\Lambda}[\phi]S_{\text{free},\Lambda}[\phi]\\ &\quad\neq -\int [d\phi]\Lambda\frac{\partial}{\partial\Lambda}(P_{\Lambda}[\phi]S_{\text{free},\Lambda}[\phi]),\end{aligned} \tag{4.22}$$

and so evidently the order of limits does not commute. This also holds more generally for RG flows of interacting theories. In our proof of the monotonicity of $M_\Lambda(P_\Lambda)$, we differentiated under the integral sign, and so apparently our proof is contingent on a certain order of limits. To resolve the ambiguity, let us *define*

$$\begin{aligned}M_{\Lambda}(P^{\text{free}}_{\Lambda}[\phi]) \coloneqq -\int_{\log\Lambda}^{\log\Lambda_0} d\log\Lambda'\Bigg[\int [d\phi]\Lambda'\frac{\partial}{\partial\Lambda'}\Bigg(P^{\text{free}}_{\Lambda'}[\phi]\\ \times\log\left(\frac{P^{\text{free}}_{\Lambda'}[\phi]}{Q^{\text{free}}_{\Lambda'}[\phi]}\right)\Bigg) - \Lambda'\frac{\partial\log(Z_{Q,\Lambda'})}{\partial\Lambda'}\Bigg],\end{aligned} \tag{4.23}$$

where we assume $\Lambda \le \Lambda_0$. In words, we are defining $M_\Lambda(P_{\text{free},\Lambda}[\phi])$ as an antiderivative of the differentiated-under-the-integral-sign quantity. This more fully specifies what we mean by $M_\Lambda(P_{\text{free},\Lambda}[\phi])$, and in particular the manner in which its divergent terms depend on Λ.

2. Order of limits in more general RG flows

In the general setting, by analogy to (4.23) we define

$$M_\Lambda(P_\Lambda[\phi]) \coloneqq -\int_{\log\Lambda}^{\log\Lambda_0} d\log\Lambda' \times\left[\int[d\phi]\Lambda'\frac{\partial}{\partial\Lambda'}\left(P_{\Lambda'}[\phi]\log\left(\frac{P_{\Lambda'}[\phi]}{Q_{\Lambda'}[\phi]}\right)\right) -\Lambda'\frac{\partial\log(Z_{Q,\Lambda'})}{\partial\Lambda'}\right], \tag{4.24}$$

where again we assume $\Lambda\le\Lambda_0$. This is the true definition of the monotone $M_\Lambda(P_\Lambda)$. Indeed, the proof of monotonicity in Sec. IV A in fact implicitly uses this prescription.

We conclude this section by reiterating a useful formula we used in our free analysis above. Equation (4.16) is $\int[d\phi]\Lambda\frac{\partial P_\Lambda[\phi]}{\partial\Lambda}=0$, which holds for general $P_\Lambda[\phi]$, and so

$$\Lambda\frac{\partial\log Z_{P,\Lambda}}{\partial\Lambda}=-\int[d\phi]P_\Lambda[\phi]\Lambda\frac{\partial S_{P,\Lambda}[\phi]}{\partial\Lambda}, \tag{4.25}$$

which is just a generalization of (4.18). This identity (4.25) will be useful for us in the section which follows.

V. EXAMPLES WITH SCALAR FIELD THEORIES

Below we exhibit computations of our RG monotone in some examples. First we consider free scalar field theory which has an exactly soluble ERG flow; hence we can compute our RG monotone exactly in this case. Next we turn to scalar ϕ^4 theory for which we can compute the RG monotone perturbatively.

A. Free scalar field

Consider a free massive scalar field which evolves via Polchinski's equation in (2.7). Using the definition of $M_\Lambda(P^{\text{free}}_\Lambda[\phi])$ in (4.23), we find

$$M_\Lambda(P^{\text{free}}_\Lambda[\phi])=-\frac{3}{2}\delta^d(0)\int d^dp\log\left(\frac{K_\Lambda(p^2)}{K_{\Lambda_0}(p^2)}\right), \tag{5.1}$$

where we suppose $\Lambda\le\Lambda_0$. The integral $\int d^dp\log(\frac{K_\Lambda(p^2)}{K_{\Lambda_0}(p^2)})$ above is divergent, but its Λ derivatives can be finite. Upon taking a Λ derivative, we find

$$-\Lambda\frac{d}{d\Lambda}M_\Lambda(P^{\text{free}}_\Lambda[\phi])=\frac{3}{2}\delta^d(0)\int d^dp\frac{\partial\log K_\Lambda(p^2)}{\partial\Lambda}\ge 0. \tag{5.2}$$

Although the above is positive for any $K_\Lambda(p^2)$ that is a smooth, monotonically decreasing cutoff function, it can be infinite. There is a nice class of $K_\Lambda(p^2)$ for which the above is finite; this is explained in Appendix B.

B. Interacting scalar field

Now we perform some explicit perturbative computations of the derivative of the RG monotone $M_\Lambda(P_\Lambda)$ in (4.24) for massive scalar ϕ^4 theory, namely where the action is

$$S_{\Lambda=\Lambda_0}[\phi]=\frac{1}{2}\int\frac{d^dp}{(2\pi)^d}(p^2+m^2)K^{-1}(p^2/\Lambda_0^2)\phi(p)\phi(-p) +\frac{\lambda}{4!}\int\frac{d^dp_1d^dp_2d^dp_3d^dp_4}{(2\pi)^{3d}}\phi(p_1)\phi(p_2) \times\phi(p_3)\phi(p_4)\delta^d(p_1+p_2+p_3+p_4). \tag{5.3}$$

We have given the action at the initial value of the cutoff $\Lambda=\Lambda_0$, where the RG flow is to be initiated. In other words, we desire to study the flow equation for $P_\Lambda[\phi]$ given its initial condition at $\Lambda=\Lambda_0$. Recall that in the context of Polchinski's equation we have

$$\dot C_\Lambda(p^2)=(2\pi)^d(p^2+m^2)^{-1}\Lambda\frac{\partial K_\Lambda(p^2)}{\partial\Lambda}, \qquad \hat S_\Lambda=\frac{1}{2}\int\frac{d^dp}{(2\pi)^d}(p^2+m^2)K_\Lambda^{-1}(p^2)\phi(p)\phi(-p). \tag{5.4}$$

Equation (4.5) provides a nice expression for the derivative of our RG monotone at $\Lambda=\Lambda_0$, namely

$$-\Lambda\frac{d}{d\Lambda}M_\Lambda(P_\Lambda)\Big|_{\Lambda=\Lambda_0} =\frac{1}{2}\int[d\phi]P_{\Lambda_0}[\phi]\int d^dxd^dy\dot C_{\Lambda_0}(x-y)\frac{\delta\Sigma_{\Lambda_0}}{\delta\phi(x)}\frac{\delta\Sigma_{\Lambda_0}}{\delta\phi(y)} -2\int[d\phi]P_{\Lambda_0}[\phi]\Lambda_0\frac{\partial\hat S_{\Lambda_0}}{\partial\Lambda_0}. \tag{5.5}$$

Working in momentum space and plugging in (5.4) we find

$$\frac{1}{2}\int[d\phi]P_{\Lambda_0}[\phi]\int d^dp(2\pi)^d(p^2+m^2)^{-1} \times\Lambda_0\frac{\partial K(p^2/\Lambda_0^2)}{\partial\Lambda_0}\frac{\delta(S_{\Lambda_0}-2\hat S_{\Lambda_0})}{\delta\phi(p)}\frac{\delta(S_{\Lambda_0}-2\hat S_{\Lambda_0})}{\delta\phi(-p)} -2\int[d\phi]P_{\Lambda_0}[\phi]\Lambda_0\frac{\partial\hat S_{\Lambda_0}}{\partial\Lambda_0}. \tag{5.6}$$

We will compute this perturbatively to second order in the quartic coupling λ.

To compress the form of our formulas, it is convenient to define

$$\tilde K_\Lambda(p)\coloneqq K_\Lambda(p^2)\frac{1}{p^2+m^2} \tag{5.7}$$

and also

$$f_0(\Lambda) = \int d^d p \Lambda \frac{\partial \log(\tilde{K}_\Lambda(p))}{\partial \Lambda}, \tag{5.8}$$

$$f_1(\Lambda) = \int d^d p \Lambda \frac{\partial \log(\tilde{K}_\Lambda(p))}{\partial \Lambda} \tilde{K}_\Lambda(p) \int \frac{d^d q}{(2\pi)^d} \tilde{K}_\Lambda(q), \tag{5.9}$$

$$f_2(\Lambda) = \int d^d p \Lambda \frac{\partial \log(\tilde{K}_\Lambda(p))}{\partial \Lambda} \tilde{K}_\Lambda(p) \int \frac{d^d q_1}{(2\pi)^d} \frac{d^d q_2}{(2\pi)^d} \times \tilde{K}_\Lambda(p - q_1 - q_2) \tilde{K}_\Lambda(q_1) \tilde{K}_\Lambda(q_2), \tag{5.10}$$

which are all finite for appropriate choices of $K_\Lambda(p^2)$ (see Appendix B). Above, we have written the $\frac{\partial \log(\tilde{K}_\Lambda(p))}{\partial \Lambda}$ term to make all the $f_i(\Lambda)$ s have the same form, but it can also be clarifying to simplify $f_1(\Lambda)$ and $f_2(\Lambda)$ using the identity $\frac{\partial \log(\tilde{K}_\Lambda(p))}{\partial \Lambda} \tilde{K}_\Lambda(p) = \frac{\partial \tilde{K}_\Lambda(p)}{\partial \Lambda}$. With our notation at hand, the main quantities in (5.6) are

$$\left\langle -2\Lambda_0 \frac{\partial \hat{S}_{\Lambda_0}}{\partial \Lambda_0} \right\rangle_{P_{\Lambda_0}} = \delta^d(0) \left(f_0(\Lambda_0) + \frac{\lambda^2}{6} f_2(\Lambda_0) + O(\lambda^3) \right), \tag{5.11}$$

$$\left\langle 2 \int d^d p (2\pi)^d \Lambda_0 \frac{\partial \tilde{K}_{\Lambda_0}}{\partial \Lambda_0} \frac{\delta \hat{S}_{\Lambda_0}}{\delta \phi(p)} \frac{\delta \hat{S}_{\Lambda_0}}{\delta \phi(-p)} \right\rangle_{P_{\Lambda_0}} = \delta^d(0) \left(2 f_0(\Lambda_0) + \frac{\lambda^2}{3} f_2(\Lambda_0) + O(\lambda^3) \right), \tag{5.12}$$

$$\left\langle -2 \int d^d p (2\pi)^d \Lambda_0 \frac{\partial \tilde{K}_{\Lambda_0}}{\partial \Lambda_0} \frac{\delta S_{\Lambda_0}}{\delta \phi(p)} \frac{\delta \hat{S}_{\Lambda_0}}{\delta \phi(-p)} \right\rangle_{P_{\Lambda_0}} = \delta^d(0) \left(-2 f_0(\Lambda_0) - \lambda f_1(\Lambda_0) - \frac{2\lambda^2}{3} f_2(\Lambda_0) + O(\lambda^3) \right), \tag{5.13}$$

$$\left\langle \frac{1}{2} \int d^d p (2\pi)^d \Lambda_0 \frac{\partial \tilde{K}_{\Lambda_0}}{\partial \Lambda_0} \frac{\delta S_{\Lambda_0}}{\delta \phi(p)} \frac{\delta S_{\Lambda_0}}{\delta \phi(-p)} \right\rangle_{P_{\Lambda_0}} = \delta^d(0) \left(\frac{1}{2} f_0(\Lambda_0) + \frac{\lambda}{2} f_1(\Lambda_0) + \frac{11\lambda^2}{24} f_2(\Lambda_0) + O(\lambda^3) \right). \tag{5.14}$$

Plugging these into (5.6) and simplifying (i.e. we just add up the above four equations), we find

$$-\Lambda \frac{d}{d\Lambda} M_\Lambda(P_\Lambda)\Big|_{\Lambda=\Lambda_0} = \delta^d(0) \left(\frac{3}{2} f_0(\Lambda_0) - \frac{1}{2} \lambda f_1(\Lambda_0) + \frac{7}{24} \lambda^2 f_2(\Lambda_0) + O(\lambda^3) \right). \tag{5.15}$$

For perturbatively small λ, the above is greater than or equal to zero as it ought to be.

Our above computation shows how the RG monotone changes along an infinitesimal step of the flow, $\Lambda_0 \to \Lambda_0 - \delta\Lambda$. We could continue with the next perturbative step along the flow, corresponding to computing $-\Lambda \frac{d}{d\Lambda} M_\Lambda(P_\Lambda)|_{\Lambda=\Lambda_0-\delta\Lambda}$. Thereafter we could continue on from there to successively smaller cutoff scales, but we will not pursue this here.

VI. VARIATIONAL FORMULATION OF RG FLOWS

We have established that Wegner-Morris flow is equivalent to the gradient flow of relative entropy with respect to a Wasserstein-2 distance on the space of fields. In this section, we show that this connection allows us to construct a variational formulation of RG flow, which may be amenable to numerical methods. Moreover, our analysis here establishes a new and precise connection between RG flows of conventional quantum field theories and numerical methods based on neural networks, which has previously only been established on a heuristic level.

A. Variational discretization of the renormalization group flow

Consider $\mathbb{R}^n$ as a Riemannian manifold with the Euclidean metric, and let F be a differentiable function $F\colon \mathbb{R}^n \to \mathbb{R}$. Then a solution to the gradient flow equation

$$\frac{d}{dt} X(t) = -\nabla F(X(t)), \tag{6.1}$$

with $X(0) = X_0$ can be approximated by a sequence of elements $X_0, X_\tau, X_{2\tau}, \ldots$ solving

$$\frac{X_{(n+1)\tau} - X_{n\tau}}{\tau} = -\nabla F(X_{(n+1)\tau}). \tag{6.2}$$

For smaller τ, the approximation becomes better. We can equivalently recast (6.2) as an optimization problem,

$$X_{(n+1)\tau} = \operatorname*{argmin}_X \left(\frac{1}{2\tau} |X - X_{n\tau}|^2 + F(X) \right). \tag{6.3}$$

These considerations also apply to a more general Riemannian manifold $\mathcal{M}$ and a function $F\colon \mathcal{M} \to \mathbb{R}$. In this setting a candidate approximate solution to (6.1) with $X(0) = X_0$ is given by a sequence of elements $X_0, X_\tau, X_{2\tau}, \ldots$ solving

$$X_{(n+1)\tau} = \operatorname*{argmin}_X \left(\frac{1}{2\tau} d(X, X_{n\tau})^2 + F(X) \right), \tag{6.4}$$

where $d(x, y)$ is the distance function on $\mathcal{M}$. This *implicit Euler discretization scheme* can be proven in many

cases [50] to give approximate solutions that converge to the solution to the gradient flow equation (6.1) in the following sense. For any fixed time $T > 0$, if we let $X(t)$ be the unique solution to the gradient flow equation (6.1) with $X(0) = X_0$, then as $\tau \to 0$ we have

$$\text{Error}(\tau) = \sup_{n=0,\ldots,\lfloor T/\tau \rfloor} d(X_{n\tau}, X(n\tau)) \to 0. \tag{6.5}$$

In this manner, we can often approximate a gradient flow on a Riemannian manifold by a solution to a sequence of optimization problems. The discussion in [51] [Chapter 8.4] sketches the general scheme of proofs for such results; a careful general discussion of several cases of this convergence result, e.g. for geodesically convex functionals F on nonpositively curved manifolds $\mathcal{M}$, can be found in [50].

Since we have recasted RG flow as a gradient flow of relative entropy with respect to a Wasserstein-2 metric, it is natural to ask if there is an approximation to RG flow along the lines of (6.4). We will find that indeed there is such an approximation, and that it is amenable to numerical optimization methods. In the finite-dimensional context, implicit gradient numerical methods, now called JKO schemes, which simulate partial differential equations arising from gradient flows of entropylike functionals, were first proposed by the pioneering [52]. In particular, [52] proves that the implicit Euler scheme (6.4) in the setting of the gradient flow of the entropy on Wasserstein space converges to the heat equation, and thus establishes the validity of this scheme in the finite-dimensional analog of the setting of statistical field theory. For large gradient steps τ, these methods require an efficient algorithmic approximation of the Wasserstein distance, which is available via the Sinkhorn algorithm [53]. Numerical methods based on the JKO scheme are a topic of current interest in the applied mathematics community [54–59], and in particular a number of recent proposals are based on approximating the Wasserstein distance by a neural-network-based method [60–62], analogous to the methodology that we propose here. Below, we explain the basic variational equations arising from the discretization of RG flows, and propose a novel numerical algorithm to compute the flow.

Recall from (2.13) that Wegner-Morris flow is in fact a field reparametrization. Suppose our initial probability functional is $P_{\Lambda_0}[\phi]$ at some scale Λ_0, and that we want to flow it to $P_{\Lambda_0 - t}[\phi]$. Then the Wegner-Morris equation says that this can be expressed as

$$P_{\Lambda_0 - t}[\phi] = \left| \frac{\delta \mathcal{R}_t[\phi]}{\delta \phi} \right| P_{\Lambda_0}[\mathcal{R}_t[\phi]] \tag{6.6}$$

for some reparametrization $\mathcal{R}_t$ which takes fields to fields. Note that $|\frac{\delta \mathcal{R}_t[\phi]}{\delta \phi}| P_{\Lambda_0}[\mathcal{R}_t[\phi]] = \mathcal{R}_{t*}^{-1} P_{\Lambda_0}$ where $\mathcal{R}_{t*}^{-1} P_{\Lambda_0}$ is the pushforward of P_{Λ_0} by the compositional inverse $\mathcal{R}_t^{-1}$ of $\mathcal{R}_t$.

For ease of notation, let us define

$$P^{\mathcal{R}}_{\Lambda_0}[\phi] := \left| \frac{\delta \mathcal{R}[\phi]}{\delta \phi} \right| P_{\Lambda_0}[\mathcal{R}[\phi]]$$

for an arbitrary reparametrization map $\mathcal{R}$. Then we claim that a solution $P_\Lambda[\phi]$ to

$$-\Lambda \frac{d}{d\Lambda} P_\Lambda[\phi] = -\nabla_{\mathcal{W}_2} S(P_\Lambda[\phi] \| Q_\Lambda[\phi]) \tag{6.7}$$

with initial probability functional $P_{\Lambda_0}[\phi]$ satisfies

$$P_{\Lambda_0 - \tau}[\phi] \approx P^{\mathcal{R}_\tau}_{\Lambda_0}[\phi] \tag{6.8}$$

for small τ, where

$$\mathcal{R}_\tau = \underset{\mathcal{R}}{\text{argmin}} \left(\frac{1}{2\tau} \mathcal{W}_2(P^{\mathcal{R}}_{\Lambda_0}, P_{\Lambda_0})^2 + S(P^{\mathcal{R}}_{\Lambda_0} \| Q_{\Lambda_0}) \right) \tag{6.9}$$

More generally, consider a sequence of reparametrizations $\mathcal{R}_\tau, \mathcal{R}_{2\tau}, \mathcal{R}_{3\tau}, \ldots$ and define

$$\tilde{\mathcal{R}}_{n\tau} := \mathcal{R}_{n\tau} \circ \cdots \circ \mathcal{R}_{2\tau} \circ \mathcal{R}_\tau. \tag{6.10}$$

Then we have

$$P_{\Lambda_0 - n\tau}[\phi] \approx P^{\tilde{\mathcal{R}}_{n\tau}}_{\Lambda_0}[\phi] \tag{6.11}$$

if the $\mathcal{R}_n$s satisfy

$$\mathcal{R}_{(n+1)\tau} = \underset{\mathcal{R}}{\text{argmin}} \left(\frac{1}{2\tau} \mathcal{W}_2(P^{\mathcal{R}}_{\Lambda_0}, P^{\tilde{\mathcal{R}}_{n\tau}}_{\Lambda_0})^2 + S(P^{\mathcal{R}}_{\Lambda_0} \| Q_{\Lambda_0 - n\tau}) \right). \tag{6.12}$$

Moreover, the approximation (6.11) should become exact as $\tau \to 0$, by analogy to the finite dimensional JKO scheme [52]. We emphasize that (6.12) is striking since it provides a variational formulation of RG flow.

A natural question is if (6.12) defines the $\mathcal{R}_{n\tau}$s uniquely. For simplicity, let us consider the $n = 1$ case, given in (6.9). A solution to (6.9) is supposed to provide us with an $\mathcal{R}_\tau$ such that $P_{\Lambda_0 - \tau}[\phi] \approx P^{\mathcal{R}_\tau}_{\Lambda_0}[\phi]$, becoming exact in the $\tau \to 0$ limit. However, there exist many reparametrizations $\mathcal{R}'_\tau$ such that

$$P^{\mathcal{R}'_\tau}_{\Lambda_0}[\phi] = P^{\mathcal{R}_\tau}_{\Lambda_0}[\phi]. \tag{6.13}$$

The fact at play here is that given a fixed probability distribution, there are many reparametrizations which transform that distribution in the same way. Accordingly, a solution to (6.9) is not unique, nor are solutions to (6.12).

While this nonuniqueness may seem bothersome, we will see shortly that the flexibility it provides is a virtue.

Examining the variational formulation of RG flow in (6.12), an undesirable aspect in practice is that the $\mathcal{W}_2(P^{\mathcal{R}}_{\Lambda_0}, P^{\tilde{\mathcal{R}}_{n\tau}}_{\Lambda_0})^2$ term itself requires an optimization to compute, on account of the infimum in (3.4). However, we can get rid of this infimum in the following, interesting way.

We begin by considering (6.9) as the $n = 1$ case of (6.12). Recasting $\mathcal{W}_2(P^{\mathcal{R}}_{\Lambda_0}, P_{\Lambda_0})^2$ in the Monge formulation (i.e. with the plausible assumption that our Kantorovich solutions are also Monge solutions), we find

$$\begin{aligned}\mathcal{W}_2(P^{\mathcal{R}}_{\Lambda_0}, P_{\Lambda_0})^2 &= \inf_{\{\mathcal{F}:\mathcal{F}_* P_{\Lambda_0} = P^{\mathcal{R}}_{\Lambda_0}\}} 2\int [d\phi] P_{\Lambda_0}[\phi] \int d^d x d^d y \dot{C}^{-1}_{\Lambda}(x,y)(\phi(x) - \mathcal{F}[\phi(x)])(\phi(y) - \mathcal{F}[\phi(y)]), \\ &= \inf_{\{\mathcal{F}:\mathcal{F}_* P_{\Lambda_0} = P^{\mathcal{R}}_{\Lambda_0}\}} \mathsf{M}_{P_{\Lambda_0}}[\mathcal{F}],\end{aligned} \tag{6.14}$$

where $\mathcal{F}$ is a reparametrization from fields to fields and $\mathsf{M}_{P_{\Lambda_0}}$ is the analog of the Monge functional in our setting. Thus we can rewrite (6.9) as

$$\mathcal{R}_\tau = \operatorname*{argmin}_{\mathcal{R}} \left(\frac{1}{2\tau} \inf_{\{\mathcal{F}:\mathcal{F}_* P_{\Lambda_0} = P^{\mathcal{R}}_{\Lambda_0}\}} \mathsf{M}_{P_{\Lambda_0}}[\mathcal{F}] + S(P^{\mathcal{R}}_{\Lambda_0} \| Q_{\Lambda_0}) \right). \tag{6.15}$$

Next we observe that that for a fixed $\mathcal{R}$, the infimum inside the *argmin* on the right-hand side will pick out an $\mathcal{F}$ such that $\mathcal{F}_* P_{\Lambda_0} = P^{\mathcal{R}}_{\Lambda_0}$. As such, we can rewrite the above equation as

$$\mathcal{R}_\tau = \operatorname*{argmin}_{\mathcal{R}} \inf_{\{\mathcal{F}:\mathcal{F}_* P_{\Lambda_0} = P^{\mathcal{R}}_{\Lambda_0}\}} \left(\frac{1}{2\tau} \mathsf{M}_{P_{\Lambda_0}}[\mathcal{F}] + S(\mathcal{F}_* P_{\Lambda_0} \| Q_{\Lambda_0}) \right). \tag{6.16}$$

Since we are ultimately interested in having access to the RG-flowed distribution $P^{\mathcal{R}_\tau}_{\Lambda_0}$ and not necessarily $\mathcal{R}_\tau$ itself, the above equation suggests the following convenient reformulation: we have

$$P_{\Lambda_0 - \tau}[\phi] \approx \mathcal{F}_{\tau *} P_{\Lambda_0}[\phi], \tag{6.17}$$

where $\mathcal{F}_\tau$ satisfies

$$\mathcal{F}_\tau = \operatorname*{argmin}_{\mathcal{F}} \left(\frac{1}{2\tau} \mathsf{M}_{P_{\Lambda_0}}[\mathcal{F}] + S(\mathcal{F}_* P_{\Lambda_0} \| Q_{\Lambda_0}) \right). \tag{6.18}$$

Moreover this should become exact as $\tau \to 0$. This equation is more convenient than (6.9) since it only has a single minimization, i.e. we have successfully accommodated for the reparametrization minimization and the Wasserstein-2 minimization in one fell swoop.

Our reformulation of (6.9) into (6.18) can similarly be applied to (6.12). In particular, let $\mathcal{F}_\tau, \mathcal{F}_{2\tau}, \mathcal{F}_{3\tau}, \ldots$ be a sequence of reparametrizations and define

$$\tilde{\mathcal{F}}_{n\tau} := \mathcal{F}_\tau \circ \mathcal{F}_{2\tau} \circ \cdots \circ \mathcal{F}_{n\tau}. \tag{6.19}$$

Note the ordering of the composition relative to (6.10), since we can think of the $\mathcal{F}$s as acting inversely as the $\mathcal{R}$s. Then we have

$$\begin{aligned}P_{\Lambda_0 - n\tau}[\phi] &\approx \tilde{\mathcal{F}}_{n\tau *} P_{\Lambda_0}[\phi], \\ &= (\mathcal{F}_{n\tau *} \circ \cdots \circ \mathcal{F}_{2\tau *} \circ \mathcal{F}_{\tau *}) P_{\Lambda_0}[\phi],\end{aligned} \tag{6.20}$$

where the $\mathcal{F}_{n\tau}$s satisfy

$$\begin{aligned}&\mathcal{F}_{(n+1)\tau} \\ &= \operatorname*{argmin}_{\mathcal{F}} \left(\frac{1}{2\tau} \mathsf{M}_{\tilde{\mathcal{F}}_{n\tau *} P_{\Lambda_0}}[\mathcal{F}] + S((\mathcal{F}_* \circ \tilde{\mathcal{F}}_{n\tau *}) P_{\Lambda_0} \| Q_{\Lambda_0 - n\tau}) \right).\end{aligned} \tag{6.21}$$

This is the desired generalization of (6.12) which only has a single optimization.

In the next subsection, we will explore strategies for solving (6.21) via numerical optimization. For the moment, let us unpack (6.21) slightly, and write it in a more convenient form. First, by iteratively changing integration variables using the diffeomorphisms $\mathcal{F}_{n\tau}$, we can rewrite $\mathsf{M}_{\tilde{\mathcal{F}}_{n\tau *} P_{\Lambda_0}}[\mathcal{F}]$ as

$$\begin{aligned}\mathsf{M}_{\tilde{\mathcal{F}}_{n\tau *} P_{\Lambda_0}}[\mathcal{F}] &= 2\int [d\phi] P_{\Lambda_0}[\phi] \int d^d x d^d y \dot{C}^{-1}_{\Lambda}(x,y)(\tilde{\mathcal{F}}_{n\tau}[\phi(x)] - (\mathcal{F} \circ \tilde{\mathcal{F}}_{n\tau})[\phi(x)])(\tilde{\mathcal{F}}_{n\tau}[\phi(y)] - (\mathcal{F} \circ \tilde{\mathcal{F}}_{n\tau})[\phi(y)]), \\ &= \mathop{\mathbb{E}}_{P_{\Lambda_0}}[2 \int d^d x d^d y \dot{C}^{-1}_{\Lambda}(x,y)(\tilde{\mathcal{F}}_{n\tau}[\phi(x)] - (\mathcal{F} \circ \tilde{\mathcal{F}}_{n\tau})[\phi(x)])(\tilde{\mathcal{F}}_{n\tau}[\phi(y)] - (\mathcal{F} \circ \tilde{\mathcal{F}}_{n\tau})[\phi(y)])].\end{aligned} \tag{6.22}$$

In a similar fashion, we can write

$$\begin{aligned} S((\mathcal{F}_* \circ \tilde{\mathcal{F}}_{n\tau *})P_{\Lambda_0} \| Q_{\Lambda_0 - n\tau}) &= \int [d\phi_0] P_{\Lambda_0} \log (P_{\Lambda_0} / Q_{\Lambda_0 - n\tau}^{\tilde{\mathcal{F}}_{n\tau} \circ \mathcal{F}}), \\ &= \mathop{\mathbb{E}}_{P_{\Lambda_0}} [\log (P_{\Lambda_0} / Q_{\Lambda_0 - n\tau}^{\tilde{\mathcal{F}}_{n\tau} \circ \mathcal{F}})]. \end{aligned} \tag{6.23}$$

This change of variables follows from iteratively utilizing the infinite-dimensional analog of the reparametrization-invariance of the relative entropy, which in finite dimensions is the statement that $S(f_* p \| q) = S(p, \| f_*^{-1} q)$ for probability distributions p, q and a diffeomorphism f.

Combining (6.22) and (6.23), we can write (6.21) in the form

$$\mathcal{F}_{(n+1)\tau} = \operatorname*{argmin}_{\mathcal{F}} \mathop{\mathbb{E}}_{P_{\Lambda_0}} \mathsf{Loss}[\mathcal{F}, \tilde{\mathcal{F}}_{n\tau}, P_{\Lambda_0}, Q_{\Lambda_0 - n\tau}], \tag{6.24}$$

where $\mathsf{Loss}[\mathcal{F}, \tilde{\mathcal{F}}_{n\tau}, P_{\Lambda_0}, Q_{\Lambda_0 - n\tau}]$ is the function to be minimized.[7]

B. Numerical applications of variational formulas

The variational characterization of RG flows discussed above suggests new and interesting numerical methods for (approximately) computing such flows. In particular, suppose we have sample access to $P_{\Lambda_0}[\phi]$. For the purposes of this section, we will take our fields to be lattice discretized on a finite volume domain; then we can sample from $P_{\Lambda_0}[\phi]$ by employing standard Monte Carlo methods. Equation (6.24) tells us that such sampling access is sufficient in principle to solve for $\mathcal{F}_\tau, \mathcal{F}_{2\tau}, \ldots, \mathcal{F}_{n\tau}$. We will return shortly to the problem of how the requisite minimizations can be implemented in practice. For the moment, let us say we have $\mathcal{F}_\tau, \mathcal{F}_{2\tau}, \ldots, \mathcal{F}_{n\tau}$ at hand, in which case we would like to be able to sample from the RG-flowed distribution $P_{\Lambda_0 - n\tau}[\phi] \approx (\mathcal{F}_{n\tau *} \circ \cdots \circ \mathcal{F}_{2\tau *} \circ \mathcal{F}_{\tau *}) P_{\Lambda_0}[\phi]$. How can we sample from such a distribution? Fortunately, sampling is readily compatible with the pushforward operation, as Algorithm 1 demonstrates.

Now we turn to the more interesting problem of numerically solving (6.24) for the reparametrizations $\mathcal{F}_\tau, \mathcal{F}_{2\tau}, \ldots$. A natural way to proceed is to let our reparametrizations $\mathcal{F}_{n\tau}$ have a particular form that only depends on a finite-dimensional vector of real parameters θ_n; we write this dependence as $\mathcal{F}_{n\tau} = \mathcal{F}_{\theta_n}$. Moreover, we define

$$\tilde{\mathcal{F}}_{\theta_1, \ldots, \theta_n} \coloneqq \mathcal{F}_{\theta_1} \circ \mathcal{F}_{\theta_2} \circ \cdots \circ \mathcal{F}_{\theta_n}. \tag{6.25}$$

In this setting, (6.24) becomes

[7]Calling this the "loss" function is standard in the computer science literature.

Algorithm 1. Sampling from $(\mathcal{F}_{n\tau *} \circ \cdots \circ \mathcal{F}_{2\tau *} \circ \mathcal{F}_{\tau *}) P_{\Lambda_0}$.

Input: Reparametrizations $\mathcal{F}_\tau, \mathcal{F}_{2\tau}, \ldots, \mathcal{F}_{n\tau}$, sample access to P_{Λ_0}
Output: A sample $\hat{\phi}$ from $(\mathcal{F}_{n\tau *} \circ \cdots \circ \mathcal{F}_{2\tau *} \circ \mathcal{F}_{\tau *}) P_{\Lambda_0}$
Sample $\phi \leftarrow P_{\Lambda_0}$
Compute $\hat{\phi} = (\mathcal{F}_\tau \circ \mathcal{F}_{2\tau} \circ \cdots \circ \mathcal{F}_{n\tau})[\phi]$
return $\hat{\phi}$

$$\theta_{n+1} = \operatorname*{argmin}_{\theta} \mathop{\mathbb{E}}_{P_{\Lambda_0}} \mathsf{Loss}[\mathcal{F}_\theta, \tilde{\mathcal{F}}_{\theta_1, \ldots, \theta_n}, P_{\Lambda_0}, Q_{\Lambda_0 - n\tau}]. \tag{6.26}$$

If $\mathcal{F}_\theta$ is a differentiable function of θ, then we can bring to bear techniques from machine learning to perform the optimization in (6.26).

First we ask this: What is a good family of functionals $\mathcal{F}_\theta$ to choose? There are certain requirements that the family should have. For instance, if we work with a translationally invariant field theory, then we would like for

$$\mathcal{F}_\theta[\phi(y + a)](x) = \mathcal{F}_\theta[\phi(y)](x + a) \tag{6.27}$$

for all θ. This is an equivariance condition on $\mathcal{F}_\theta$ with respect to translations. Moreover, in spatially local RG schemes such as those we studied in the context of Polchinski's equation and the Wegner-Morris flow equation, we would also like $\mathcal{F}_\theta[\phi(y)](x)$ to only depend strongly on values of $\phi(y)$ near $y \approx x$.

These properties imply that a good *Ansatz* for $\mathcal{F}_\theta$ is to let it be a convolutional neural network [63] with weights θ. Indeed, in the context of numerical algorithms, convolutional neural networks are known [64] to give good *Ansätze* for general, translationally invariant functionals of fields such that the functional is spatially local in the sense we discussed above. For concreteness, we remind the reader that if ϕ is, for example, a lattice discretization $\phi(i, j)$ of a two-dimensional field, then a convolutional neural network of depth D would render $\mathcal{F}_\theta[\phi]$ as having the form

$$\mathcal{F}_\theta^0[\phi](i, j) = \phi(i, j), \tag{6.28}$$

$$\mathcal{F}_\theta^\ell[\phi](i, j) = \sigma\left(\sum_{m,n=-k}^{k} A_{m,n}^\ell \mathcal{F}^{\ell - 1}[\phi](i + m, j + n) \right), \tag{6.29}$$

$$\mathcal{F}_\theta^D[\phi] = \mathcal{F}_\theta[\phi] \tag{6.30}$$

for $\ell = 1, \ldots, D$ where the A_{mn}^ℓ are $2k \times 2k$ matrices representing discrete convolutional kernels, while $\sigma(x)$ is a nonlinear function such as $\sigma(x) = \frac{1}{1 + e^{-x}}$. The parameters θ of the neural network are the entries of the matrix kernels $A^1, \ldots, A^D$. We note also that k can depend on the "layer" ℓ, and that often one inserts intermediate "max-pool" or

"average-pool" layers that downsample the intermediate fields $\mathcal{F}^{\ell}_{\theta}[\phi]$. For example, in average pooling, one replaces $\mathcal{F}^{\ell}_{\theta}[\phi](i,j)$ by a new function

$$\widetilde{\mathcal{F}^{\ell}_{\theta}[\phi]}(i',j') = \frac{1}{|B(i',j')|}\sum_{(i,j)\in B(i',j')} \mathcal{F}^{\ell}_{\lambda}[\phi](i,j), \quad (6.31)$$

where (i',j') runs over a lattice with fewer sites, and $B(i',j')$ is a subset of the indices of the (i,j) lattice, exactly as in block spin renormalization.

With a particular neural network architecture for $\mathcal{F}_{\theta}$ in mind, we examine how to solve (6.26). Observe that the quantity to be minimized in that equation is an expectation value over P_{Λ_0}. Therefore, the gradient of this quantity with respect to the variational parameter θ is also an explicit expectation value over P_{Λ_0}. Accordingly, if we have sampling access to P_{Λ_0}, then as is standard in stochastic gradient descent, we can approximate $\mathbb{E}_{P_{\Lambda_0}}\nabla_{\theta}\mathsf{Loss}$ via a Monte Carlo approximation, replacing the expectation value with an evaluation of the gradient on a single sample from P_{Λ_0}. The nontrivial fact that the quantity to be optimized is an expectation value over P_{Λ_0} [which holds due to the reparametrizations in (6.22), (6.23)] is needed to even conceive of a plausible numerical algorithm in this setting, since general integrals over the space of fields are completely intractable. Indeed, the problem of turning this variational formulation of RG flow into a tractable numerical method is rather interesting, and must involve the use of several approximation techniques from neural networks beyond the most basic formulation given above.[8] We leave to future work a comprehensive exploration of neural network numerical methods based on the variational formulation of RG flow.

In more detail, the stochastic gradient descent algorithm for computing θ_{n+1} is found in Algorithm 2.

We note that the relative entropy contained in $\mathsf{Loss}[\theta,\phi]$ is in fact finite due to our (hard) lattice cutoff and finite volume domain. This is in contrast to our continuum analysis of the relative entropy in Sec. IV, in which we had to contend with divergences and associated subtleties with orders of limits.

Algorithm 2. Stochastic gradient descent for computing θ_{n+1}.

Input: Sampling access to P_{Λ_0}, loss function
$\mathsf{Loss}[\mathcal{F}_{\theta}, \tilde{\mathcal{F}}_{\theta_1,\ldots,\theta_n}, P_{\Lambda_0}, Q_{\Lambda_0-n\tau}] =: \mathsf{Loss}[\theta,\phi]$
parametrized as a neural network in a differentiable parameter θ, initial θ value θ_{init},
maximum step number $\tau_{\max}$, rate parameter r
Output: Approximation to θ_{n+1}
Initialize $\theta = \theta_{\text{init}}$
for $\tau = 1,\ldots,\tau_{\max}$ **do**
  Sample $\phi \leftarrow P_{\Lambda_0}$
  Compute $\nabla_{\theta}\mathsf{Loss}[\theta,\phi]$ via backpropagation
  Replace $\theta \to \theta - r\nabla_{\theta}\mathsf{Loss}[\theta,\phi]$
end for
return θ

There is a wealth of ideas that have circulated about connections between the renormalization group and the hierarchical structure of convolutional neural networks [67–73]. These connections have at times informed the theory and methodology around neural network training [74,75]. However, these considerations were largely heuristic, and did not connect convolutional neural networks with any explicit renormalization group flow for any particular theory. The formulation provided in this section seems to be the first precise connection between an optimization problem based on convolutional neural networks and an explicit instantiation of the renormalization group in the standard setting of field theory.

VII. DISCUSSION

In this paper we have provided a new approach to the exact renormalization group using the tools of optimal transport theory. In so doing, we defined new, nonperturbative RG monotones, developed a novel variational formula for RG flows, and suggested new numerical algorithms.

Going forward, it would be interesting to apply the techniques in this paper to a richer class of field theories, such as gauge theories (see e.g. [10,12,13]). Moreover, it would be desirable to compute more examples of our RG monotone in any setting. In the realm of scalar field theories, a natural target would be to consider flows in the neighborhood of the Wilson-Fisher fixed point.

Since we have defined RG monotones for a large class of RG flows, it seems possible that for a judicious choice of RG flow (for instance, a judicious choice of the seed action $\hat{S}_{\Lambda}[\phi]$ in the Wegner-Morris formulation) one could use the positivity of our monotone to constrain the signs of couplings in effective field theory.

[8] In particular, the efficient discretization of the functional determinant $|\frac{\delta\mathcal{F}_{\theta}[\phi]}{\delta\phi}|$ and its θ derivatives is tricky; the natural lattice analog of the functional determinant is $|\det\nabla_{\phi_{ij}}\mathcal{F}_{\theta}|$, which does not make sense for RG flows which decrease the number of effective lattice sites. The simplest solution is to let $\mathcal{F}_{\theta}$ be a convolution that does not decrease the number of lattice sites, and to utilize the RESNET technique [65]. This is a standard technique in neural networks which involves the replacement $\mathcal{F}_{\theta} \to \mathrm{Id} + \mathcal{F}_{\theta}$, which for small θ forces the Jacobian to be an invertible matrix, making the log-determinant nondegenerate. This neural network architecture is also natural since $\mathcal{F}_{\theta}$ should approximate a small renormalization group transformation, which should be thought of as a perturbation of the identity transformation. Finally, the log determinant and its gradient are challenging quantities to compute efficiently when the dimension is large; however, there are efficient tricks to analytically compute log determinants of convolutions using circulant matrices [66], which can also be utilized to force $\mathcal{F}_{\theta}$ to be a diffeomorphism.

It would be very interesting to empirically investigate numerical methods based on our proposal in Sec. VI B. The initial neural-network-based proposal that we describe connects nicely with recent advances in neural network approaches to Wasserstein gradient flows [60,61]. The use of specialized neural network algorithms [62], possibly coupled with connections to fast approximations of the Wasserstein distance [53], may allow one to robustly approximate RG flows with relatively few discrete steps. It is an important practical problem to make the initial numerical method proposed above significantly more numerically efficient using the wealth of ideas in the machine learning literature on generative models and normalizing flows; our proposal is only a first step towards a practical numerical method.

Since RG allows one to determine P_Λ for different scales Λ, the variational formulation of RG may allow for improved sampling algorithms via connections to recent advances in neural network generative models [76], which involve denoising procedures that are heuristically related to the inversion of the renormalization group flow. A related generalization would be to develop optimal transport algorithms for continuous MERA (cMERA) tensor networks [77–80], by leveraging and generalizing the known connection between cMERA and ERG [81]. A suggestive possibility is to implement a backwards gradient flow to go from an IR cMERA *Ansatz* to a UV state. In a similar vein, perhaps one could adapt the optimal transport technology to study the RG flow of quantum states using techniques from (see e.g. [82,83]).

More broadly, it seems that many more tools from optimal transport, possibly combined with information theory, can be brought to bear on the subject of ERG via our present formulation. For instance, it appears likely that our formulation of RG flow in this paper could be synthesized with the approaches of [35–45] from the physics community. Since the optimal transport community has enormous analytical and numerical traction in the PDE setting, it would be valuable to adapt these insights to the functional generalizations appropriate for ERG flows.

ACKNOWLEDGMENTS

We thank Kristan Jensen, Igor Klebanov, Nima Lashkari, Tim Morris, Yair Shenfeld, and Andrew Strominger for valuable discussions. We give a special thanks to Arthur Kosmala for identifying and correcting an error in our definition of the Wasserstein distance in the quantum field theory setting. J. C. is supported by a Junior Fellowship from the Harvard Society of Fellows, the Black Hole Initiative, as well as in part by the Department of Energy under Grant No. DE-SC0007870. S. R. is supported by the Simons Foundation Collaboration grant "Homological Mirror Symmetry and Applications" (Grant No. 385573).

APPENDIX A: COMMENTS ON THE INFINITESIMAL FORM OF THE $\mathcal{W}_2$ METRIC

This Appendix summarizes a derivation of the infinitesimal form of the Wasserstein metric from its finite-distance definition, clarifying the inversion of the Riemannian metric that occurs when passing from the finite-distance form to the infinitesimal form. An infinite-dimensional analog of this computation leads to the functional Wasserstein metric of (3.4).

For the purposes of this Appendix, it is useful to consider a modified version of the heat equation

$$\frac{\partial p}{\partial t} = A^{ij}(\partial_i \partial_j p), \tag{A1}$$

where A^{ij} is positive semidefinite as a matrix. Throughout this appendix, we will use Einstein index notation so that A^{ij} and A_{ij} are inverses, i.e. $A_{ij}A^{jk} = \delta_i^k$.

Let $\mathrm{dens}(M)$ be the space of probability distributions on $M = \mathbb{R}^d$ so that the tangent space is $T_p\mathrm{dens}(M) = \{\bar{\eta} \in C^\infty(\mathbb{R}^d) : \int dx \bar{\eta} = 0\}$ for any $p \in \mathrm{dens}(M)$. For any tangent vector η in $T_p\mathrm{dens}(M)$ we have an associated $\bar{\eta}$ obtained by solving

$$A^{ij}\partial_i(p\partial_j\bar{\eta}) = \eta. \tag{A2}$$

This solution $\bar{\eta}$ is unique up to an additive constant; this induces the identification $\eta \leftrightarrow \bar{\eta}$ via an isomorphism

$$\begin{aligned} T_p\mathrm{dens}(M) &\simeq \overline{T_p\mathrm{dens}(M)} \\ &:= \{\bar{\eta} \in C^\infty(M)\}/\{\text{constants}\}. \end{aligned} \tag{A3}$$

With this isomorphism in mind, we can write down the Riemannian metric

$$\begin{aligned} \langle \eta_1, \eta_2 \rangle_{\mathcal{W}_2} &:= \int dx p A^{ij}\partial_i\bar{\eta}_1\partial_j\bar{\eta}_2 = -\int dx \eta_1\bar{\eta}_2 \\ &= -\int dx \bar{\eta}_1\eta_2. \end{aligned} \tag{A4}$$

The last two equalities are obtained via integration by parts. Using this Riemannian metric we have that the modified heat equation in (A1) can be written as

$$\frac{\partial p}{\partial t} = \nabla_{\mathcal{W}_2} S[p], \tag{A5}$$

since here $\nabla_{\mathcal{W}_2} S[p] = A^{ij}(\partial_i\partial_j p)$. A rigorous argument given in [24] [Lemma 4.3] establishes that (A4) is the infinitesimal version of the Wasserstein-2 metric

$$\mathcal{W}_2(p_0,p_1) := \left(\inf_{\pi\in\Gamma(p_0,p_1)}\int dxdy\pi(x,y)A_{ij}(x^i-y^i)(x^j-y^j)\right)^{1/2},$$

and here we will explain some heuristics for key parts of the proof.

Recall that in Riemannian geometry, given a path $x(u)$ with $u\in[0,1]$ in a Riemannian manifold, the length of that path is given by

$$L[x(u)] = \int_0^1 du\sqrt{A_{ij}\partial_u x^i(u)\partial_u x^j(u)}. \tag{A6}$$

The minimizers of the length functional $L[x(u)]$ with fixed boundary conditions at $x(0)$ and $x(1)$ are geodesics. However, these minimizers are always nonunique because the length functional is invariant under reparametrizations. This high degree of nonuniqueness can be avoided by instead considering the energy functional

$$E[x(u)] = \frac{1}{2}\int_0^1 duA_{ij}\partial_u x^i(u)\partial_u x^j(u). \tag{A7}$$

Its minimizers with fixed boundary conditions at $x(0)$ and $x(1)$ are exactly geodesics with constant speed. [In essence, the energy functional picks out a preferred "reparametrization" of $x(u)$]. Now using the Cauchy-Schwarz inequality, we have

$$L[x(u)]^2 \le 2E[x(u)] \tag{A8}$$

with equality exactly when $|x'(u)|$ is constant in time, i.e. $x(u)$ is parametrized so it has constant speed. This implies that

$$\inf_{\{x(u):x(0)=a,x(u)=b\}} L[x(u)]^2 = \inf_{\{x(u):x(0)=a,x(1)=b\}} 2E[x(u)], \tag{A9}$$

namely that $L[x(u)]^2$ and $2E[x(u)]$ have the same minimizing values.

We will apply the above insights to study the Riemannian metric (A4) on $T_p\mathrm{dens}(M)$. Suppose we have a 1-parameter family of probability distributions $p(u)$ for $u\in[0,1]$ where we take $p(0)=p_0$ and $p(1)=p_1$. We emphasize that the u appearing in $p(u)$ [which we will also write as $p(x,u)$] is *different* from the time coordinate t appearing in (A1). The t there corresponds to time evolution, whereas the u here parametrizes a geodesic flow in the space of probability distributions. With this in mind, we define $\phi(u)$ as a 1-parameter family of solutions to the equations

$$\frac{\partial}{\partial u}p(u) = A^{ij}\partial_i(p(u)\partial_j\phi). \tag{A10}$$

The energy of the path $p(u)$ with respect to the Riemannian metric (A4) is given by

$$E_{\mathcal{W}_2}[p(u)] := \frac{1}{2}\int_0^1 du\int dxp(u)A^{ij}\partial_i\phi\partial_j\phi. \tag{A11}$$

Writing $V^i(x,u) = -A^{ij}\partial_j\phi(x,u)$, we can define a flow on $\mathrm{dens}(M)$, namely

$$\Phi_{u*}\colon \mathrm{dens}(M)\to\mathrm{dens}(M), \qquad u\in[0,1], \tag{A12}$$

via the differential equation

$$\frac{\partial}{\partial u}\Phi_u = V(\Phi_u,u), \qquad \Phi_0 = \mathrm{Id}. \tag{A13}$$

Let us check that $p(u)=\Phi_{u*}p_0$. It suffices to show that

$$p(x,u+du) = p(\Phi_{-du}(x),u)|\det\partial\Phi_{-du}|. \tag{A14}$$

We first note that

$$\begin{aligned}\Phi^i_{-du}(x) &= x^i - duV^i,\\ &= x^i + duA^{ik}\partial_k\phi(x,u)\end{aligned} \tag{A15}$$

and also

$$\partial_j\Phi^i_{-du}(x) = \delta^i_j + duA^{ik}\partial_j\partial_k\phi(x,u). \tag{A16}$$

It follows that

$$\begin{aligned}p(\Phi^i_{-du}(x),u) &= p(x^i+duA^{ij}\partial_j\phi(x,u),u),\\ &= p(x,u)+dt\partial_i p(x,u)A^{ij}\partial_j\phi(x,u),\end{aligned} \tag{A17}$$

and accordingly

$$|\det\partial\Phi_{-du}| = 1 + duA^{ij}\partial_i\partial_j\phi(x,u). \tag{A18}$$

Altogether we have

$$\begin{aligned}&p(\Phi_{-du}(x),u)|\det\partial\Phi_{-du}|\\ &\quad= p(x,u)+duA^{ij}\partial_i(p(x,u)\partial_j\phi(x,u)),\\ &\quad= p(x,u)+du\frac{\partial}{\partial u}p(x,u),\end{aligned} \tag{A19}$$

where we have used (A10) in going from the first line to the second line. This establishes (A14).

Having checked that $p(u)=\Phi_{u*}p_0$, we have the standard inequalities

$$\sqrt{A_{ij}(x-\Phi_1(x))^i(x-\Phi_1(x))^j}$$
$$= \sqrt{A_{ij}(\Phi_0(x)-\Phi_1(x))^i(\Phi_0(x)-\Phi_1(x))^j},$$
$$\leq \int_0^1 du \sqrt{A_{ij}\left(\frac{\partial}{\partial u}\Phi_u(x)\right)^i\left(\frac{\partial}{\partial u}\Phi_u(x)\right)^j}. \quad \text{(A20)}$$

In words, this equality holds because the metric distance between the end points of a curve is upper bounded by the length of that curve. Combining this with the inequality (A8), we obtain

$$A_{ij}(x-\Phi_1(x))^i(x-\Phi_1(x))^j \leq \int_0^1 du A_{ij}\left(\frac{\partial}{\partial u}\Phi_u(x)\right)^i\left(\frac{\partial}{\partial u}\Phi_u(x)\right)^j. \quad \text{(A21)}$$

This inequality (A21) will be useful to us below.

Let us show that $\frac{1}{2}\mathcal{W}_2(p_0,p_1)^2 \leq E_{\mathcal{W}_2}[p(u)]$. Using the definition of $\mathcal{W}_2(p_0,p_1)$, we have

$$\frac{1}{2}\mathcal{W}_2(p_0,p_1)^2 = \frac{1}{2}\inf_{\pi\in\Gamma(p_0,p_1)}\int dxdy\pi(x,y)A_{ij}(x^i-y^i)(x^j-y^j),$$
$$\leq \frac{1}{2}\int dxdy p_0(x)\delta(y-\Phi_1(x))A_{ij}(x^i-y^i)(x^j-y^j),$$
$$= \frac{1}{2}\int dx p_0(x)A_{ij}(x-\Phi_1(x))^i(x-\Phi_1(x))^j. \quad \text{(A22)}$$

This inequality comes from making the particular choice of $\pi(x,y) = p_0(x)\delta(y-\Phi_1(x))$, which may not be the minimizing choice of $\pi(x,y)$. Next, we use (A21) to upper bound the last line of (A22) as

$$\frac{1}{2}\int dx p_0 A_{ij}(x-\Phi_1(x))^i(x-\Phi_1(x))^j \leq \frac{1}{2}\int dx\int_0^1 du p_0 A_{ij}\left(\frac{\partial}{\partial u}\Phi_u(x)\right)^i\left(\frac{\partial}{\partial u}\Phi_u(x)\right)^j. \quad \text{(A23)}$$

Using the definition of (A13) and the fact that $V(x,u) = -A^{ij}\partial_j\phi(x,u)$, the right-hand side of the above inequality equals

$$\frac{1}{2}\int dx\int_0^1 du p_0 A^{ij}\partial_i\phi(\Phi_u(x),u)\partial_j\phi(\Phi_u(x),u)$$
$$= \frac{1}{2}\int dx\int_0^1 du\Phi_{u*}p_0A^{ij}\partial_i\phi(x,u)\partial_j\phi(x,u),$$
$$= \frac{1}{2}\int dx\int_0^1 du p(u)A^{ij}\partial_i\phi(x,u)\partial_j\phi(x,u),$$
$$= E_{\mathcal{W}_2}[p(u)]. \quad \text{(A24)}$$

Combining (A22), (A23), and (A24) we obtain the desired inequality

$$\frac{1}{2}\mathcal{W}_2(p_0,p_1)^2 \leq E_{\mathcal{W}_2}[p(u)]. \quad \text{(A25)}$$

One works harder to show that, in fact,

$$\frac{1}{2}\mathcal{W}_2(p_0,p_1)^2 = \inf_{\{p(u):p(0)=p_0,p(1)=p_1\}} E_{\mathcal{W}_2}[p(u)]. \quad \text{(A26)}$$

Defining the length

$$L_{\mathcal{W}_2}[p(u)] = \int_0^1 du\int dx\sqrt{pA^{ij}\partial_i\phi\partial_j\phi} \quad \text{(A27)}$$

and applying a similar logic as that which led to (A9), we find

$$\inf_{\{p(u):p(0)=p_0,p(1)=p_1\}} L_{\mathcal{W}_2}[p(u)]^2 := \inf_{\{p(u):p(0)=p_0,p(1)=p_1\}} 2E_{\mathcal{W}_2}[p(u)] \quad \text{(A28)}$$

and thus

$$\mathcal{W}_2(p_0,p_1) = \inf_{\{p(u):p(0)=p_0,p(1)=p_1\}} L_{\mathcal{W}_2}[p(u)]. \quad \text{(A29)}$$

This establishes that the metric (A4) is in fact the infinitesimal form of the Wasserstein-2 metric.

APPENDIX B: CONVENIENT CUTOFF FUNCTIONS

In order for the derivative $-\Lambda\frac{d}{d\Lambda}M_\Lambda$ of our RG monotone quantity to be perturbatively or nonperturbatively finite, we saw in Sec. V. 1 that we must have

$$\int d^dp\frac{\partial\log K_\Lambda(p^2)}{\partial\Lambda} < \infty. \quad \text{(B1)}$$

This finiteness property does not hold for all cutoff functions $K_\Lambda(p^2)$; for instance, it fails to hold for

$$K(p^2/\Lambda^2) = \frac{1+e^{-a}}{1+e^{a((p/\Lambda)^2-1)}}, \quad \text{(B2)}$$

where a is a constant, usually taken to be much greater than one.

The reason that cutoff functions like (B2) lead to divergent $\int d^dp\frac{\partial\log K_\Lambda(p^2)}{\partial\Lambda}$ is that for such cutoff functions, as one varies Λ, arbitrarily high scales are suppressed by a multiplicative factor which does not decay appreciably with p^2. In particular, for any Λ, the logarithmic derivative of the K_Λ in (B2) is greater than $1/|p|^d$ for all sufficiently large p^2, and so the integral $\int d^dp\frac{\partial\log K_\Lambda(p^2)}{\partial\Lambda}$ diverges.

The condition (B1) should thus be interpreted as requiring that the factor by which one suppresses large frequencies as one varies Λ infinitesimally should decay rapidly with p^2. This condition is readily satisfied in an infinite family of cutoff functions.

In order to construct this desirable class of cutoff functions, it is instructive to first understand what other properties a cutoff function should have. First, we would like $K_\Lambda(p^2)$ to decay faster than any polynomial as a function of p^2 for $p^2 \gtrsim \Lambda^2$ in order for perturbation theory of our field theory to be sensible when the cutoff scale is Λ. Second, $K_\Lambda(p^2)$ should be very close to one for $p^2 \leq \Lambda^2 - \varepsilon$ where $\varepsilon > 0$ is a small constant, and $K_\Lambda(p^2)$ should drop off to near zero for $p^2 \geq \Lambda^2 + \varepsilon$. For simplicity, we impose the following requirements:

$$\begin{aligned}
&1.\ K_\Lambda(p^2) = 1 \quad \text{for } p^2 \leq \Lambda^2,\\
&2.\ K_\Lambda(p^2) \text{ is monotonically decreasing for } p^2 \geq \Lambda^2,\\
&3.\ K_\Lambda(p^2) \leq \varepsilon \quad \text{for } p^2 \geq \Lambda^2 + \varepsilon,\\
&4.\ \int d^d p \, \frac{\partial \log K_\Lambda(p^2)}{\partial \Lambda} < \infty.
\end{aligned} \tag{B3}$$

We will show that cutoff functions $K_\Lambda(p^2)$ satisfying these requirements exist in abundance and can be chosen to take somewhat simple forms.

For illustrative purposes, we find such a $K_\Lambda(p^2)$ explicitly. Recall that

$$B(x) = \begin{cases} \exp\left(1 - \frac{1}{1-x^2}\right) & \text{if } |x| \leq 1 \\ 0 & \text{if } |x| > 1 \end{cases} \tag{B4}$$

is a bump function supported on $[-1, 1]$ such that $B(1) = B(-1) = 0$. Moreover, all higher derivatives of B at $x = \pm 1$ equal zero as well. Let us fix some auxiliary scale $\Lambda_{\max}$, which we take to be some scale larger than our initial value of the UV cutoff Λ_0. It will be convenient to specify $K_\Lambda(p^2)$ in three regimes: (i) $\Lambda = \Lambda_{\max}$, (ii) $\Lambda < \Lambda_{\max}$, and (iii) $\Lambda > \Lambda_{\max}$.

At $\Lambda = \Lambda_{\max}$, we fix $K_\Lambda(p^2)$ such that this function satisfies conditions 1, 2, and 3 of (B3):

$$K_{\Lambda=\Lambda_{\max}}(p^2) = \begin{cases} 1 & p^2 \leq \Lambda_{\max}^2 \\ (1-\varepsilon) B((p^2 - \Lambda_{\max}^2)/\varepsilon) + \varepsilon & \Lambda_{\max}^2 \leq p^2 \leq \Lambda_{\max}^2 + \varepsilon \\ \varepsilon e^{-(p^2 - (\Lambda_{\max}^2 + \varepsilon))} & p^2 > \Lambda_{\max}^2 + \varepsilon \end{cases}. \tag{B5}$$

We then define $K_\Lambda(p^2)$ for $\Lambda < \Lambda_{\max}$ by requiring that the function satisfies conditions 1, 2, and 3 of (B3), being identically ε from $\Lambda^2 + \varepsilon$ up to $\Lambda_{\max}^2 + \varepsilon$, and then agreeing with the exponentially decaying tail of $K_{\Lambda=\Lambda_{\max}}(p^2)$ for all larger values of p^2:

$$K_{\Lambda<\Lambda_{\max}}(p^2) = \begin{cases} 1 & p^2 \leq \Lambda^2 \\ (1-\varepsilon) B((p^2 - \Lambda^2)/\varepsilon) + \varepsilon & \Lambda^2 \leq p^2 \leq \Lambda^2 + \varepsilon \\ \varepsilon & \Lambda^2 + \varepsilon \leq p^2 \leq \Lambda_{\max}^2 + \varepsilon \\ \varepsilon e^{-(p^2 - (\Lambda_{\max}^2 + \varepsilon))} & p^2 > \Lambda_{\max}^2 + \varepsilon \end{cases}. \tag{B6}$$

Finally, for $\Lambda > \Lambda_{\max}$, we define $K_\Lambda(p^2)$ by requiring that this function satisfies conditions 1, 2, and 3 of (B3), but cuts off more sharply in the interval $\Lambda^2 \leq p^2 \leq \Lambda^2 + \varepsilon$ such that for $p^2 \geq \Lambda^2 + \varepsilon$ the function still agrees identically with the exponentially decaying tail of $K_{\Lambda=\Lambda_{\max}}(p^2)$:

$$K_{\Lambda>\Lambda_{\max}}(p^2) = \begin{cases} 1 & p^2 \leq \Lambda^2 \\ (1 - \varepsilon e^{-\Lambda^2 + \Lambda_{\max}^2}) B((p^2 - \Lambda^2)/\varepsilon e^{-\Lambda^2 + \Lambda_{\max}^2}) + \varepsilon e^{-\Lambda^2 + \Lambda_{\max}^2} & \Lambda^2 \leq p^2 \leq \Lambda^2 + \varepsilon \\ \varepsilon e^{-(p^2 - (\Lambda_{\max}^2 + \varepsilon))} & p^2 \geq \Lambda^2 + \varepsilon \end{cases}. \tag{B7}$$

The only condition of (B3) that remains to be verified is the fourth one. However, with our choice of $K_\Lambda(p^2)$ specified in (B5), (B6), and (B7) above, it is clear that for each fixed Λ, the derivative $\frac{\partial}{\partial \Lambda} K_\Lambda(p^2)$ is supported in $\Lambda^2 \leq p^2 \leq \Lambda^2 + \varepsilon$. By the chain rule, the same holds for $\frac{\partial}{\partial \Lambda} \log K_\Lambda(p^2)$, rendering this quantity integrable since it is continuous and supported on a compact set. Thus condition 4 of (B3) is also satisfied. A schematic of the cutoff

function specified by (B5), (B6), and (B7) is shown in Fig. 4.

Recall from our discussion above that the condition (B1) requires that the factor by which one suppresses large frequencies as one infinitesimally varies Λ should decay rapidly with p^2. Thus, in the extreme but simple case of the cutoff function $K_\Lambda(p^2)$ specified in (B5), (B6), and (B7), large frequencies are not suppressed beyond the fixed initial exponential suppression at frequencies above $\Lambda_{\max}$ (see the last case in each piecewise function definition). One can construct cutoff functions which offer additional suppression of large frequencies by using Λ-dependent functions more complicated than exponentials.

As an aside, we note that our explicit construction for $K_\Lambda(p^2)$ above is not a differentiable function of p^2 at $p^2 = \Lambda^2 + \varepsilon$. However, the RG flows we considered in this paper do not contain p derivatives of $K_\Lambda(p^2)$, and so the requirement of p^2 differentiability is not necessary. Nonetheless, by adding a small interpolating region before the last case of each of (B5), (B6), and (B7), we can modify $K_\Lambda(p^2)$ so that it is smooth everywhere at the cost of somewhat more complicated formulas.

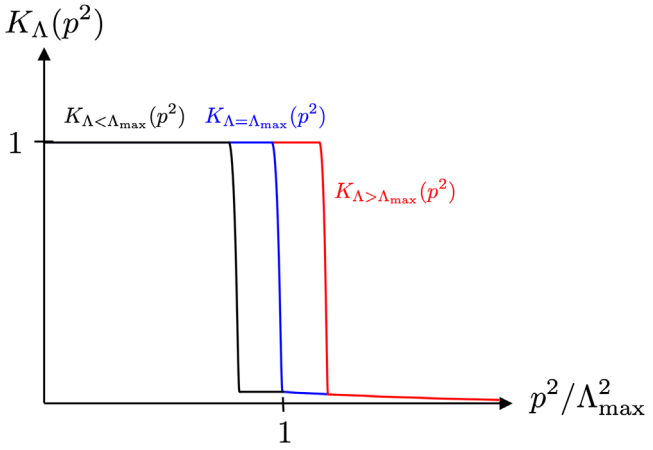

FIG. 4. Depiction of the cutoff function $K_\Lambda(p^2)$ given by (B5), (B6), and (B7).

[1] K. G. Wilson and J. Kogut, The renormalization group and the ϵ expansion, Phys. Rep. **12**, 75 (1974).
[2] J. Polchinski, Renormalization and effective Lagrangians, Nucl. Phys. **B231**, 269 (1984).
[3] C. Bagnuls and C. Bervillier, Exact renormalization group equations. An Introductory review, Phys. Rep. **348**, 91 (2001).
[4] J. Berges, N. Tetradis, and C. Wetterich, Nonperturbative renormalization flow in quantum field theory and statistical physics, Phys. Rep. **363**, 223 (2002).
[5] O. J. Rosten, Fundamentals of the exact renormalization group, Phys. Rep. **511**, 177 (2012).
[6] C. Villani, *Optimal Transport: Old and New* (Springer, New York, 2009), Vol. 338.
[7] S. Arnone, A. Gatti, T. R. Morris, and O. J. Rosten, Exact scheme independence at two loops, Phys. Rev. D **69**, 065009 (2004).
[8] T. R. Morris and O. J. Rosten, A Manifestly gauge invariant, continuum calculation of the SU(N) Yang-Mills two-loop beta function, Phys. Rev. D **73**, 065003 (2006).
[9] T. R. Morris and O. J. Rosten, Manifestly gauge invariant QCD, J. Phys. A **39**, 11657 (2006).
[10] J. M. Pawlowski, Aspects of the functional renormalisation group, Ann. Phys. (Amsterdam) **322**, 2831 (2007).
[11] M. Salmhofer and C. Honerkamp, Fermionic renormalization group flows: Technique and theory, Prog. Theor. Phys. **105**, 1 (2001).
[12] M. Reuter and C. Wetterich, Effective average action for gauge theories and exact evolution equations, Nucl. Phys. **B417**, 181 (1994).
[13] H. Gies, Introduction to the functional RG and applications to gauge theories, Lect. Notes Phys. **852**, 287 (2012).
[14] F. Wegner, Some invariance properties of the renormalization group, J. Phys. C **7**, 2098 (1974).
[15] T. R. Morris, A gauge invariant exact renormalization group. 1., Nucl. Phys. **B573**, 97 (2000).
[16] J. I. Latorre and T. R. Morris, Exact scheme independence, J. High Energy Phys. 11 (2000) 004.
[17] S. Arnone, A. Gatti, and T. R. Morris, A proposal for a manifestly gauge invariant and universal calculus in Yang-Mills theory, Phys. Rev. D **67**, 085003 (2003).
[18] S. Arnone, T. R. Morris, and O. J. Rosten, A generalised manifestly gauge invariant exact renormalisation group for SU(N) Yang-Mills, Eur. Phys. J. C **50**, 467 (2007).
[19] L. P. Kadanoff, Scaling laws for Ising models near T_c, Phys. Phys. Fiz. **2**, 263 (1966).
[20] L. P. Kadanoff, *Statistical Physics: Statics, Dynamics and Renormalization* (World Scientific, Singapore, 2000).
[21] F. Otto, The geometry of dissipative evolution equations: The porous medium equation, Commun. Partial Differ. Equations **26**, 101 (2001).
[22] Y. Brenier, Polar factorization and monotone rearrangement of vector-valued functions, Commun. Pure Appl. Math. **44**, 375 (1991).
[23] L. A. Caffarelli, The regularity of mappings with a convex potential, J. Am. Math. Soc. **5**, 99 (1992).
[24] F. Otto and M. Westdickenberg, Eulerian calculus for the contraction in the Wasserstein distance, SIAM J. Math. Anal. **37**, 1227 (2005).
[25] J.-D. Benamou and Y. Brenier, A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem, Numer. Math. **84**, 375 (2000).

[26] N. Guillen and R. McCann, Five lectures on optimal transportation: Geometry, regularity and applications, Analysis and geometry of metric measure spaces: lecture notes of the séminaire de Mathématiques Supérieure (SMS) Montréal, 145 (2010).
[27] L. Gross, Logarithmic Sobolev inequalities, Am. J. Math. **97**, 1061 (1975).
[28] F. Otto and C. Villani, Generalization of an inequality by talagrand and links with the logarithmic Sobolev inequality, J. Funct. Anal. **173**, 361 (2000).
[29] R. Bauerschmidt and T. Bodineau, Log-Sobolev inequality for the continuum Sine-Gordon model, Commun. Pure Appl. Math. **74**, 2064 (2021).
[30] Y. Shenfeld, Exact renormalization groups and transportation of measures, arXiv:2205.01642.
[31] G. Perelman, The entropy formula for the Ricci flow and its geometric applications, arXiv:math/0211159.
[32] R. J. McCann and P. M. Topping, Ricci flow, entropy and optimal transportation, Am. J. Math. **132**, 711 (2010).
[33] T. M. Cover, *Elements of Information Theory* (John Wiley & Sons, New York, 1999).
[34] E. Witten, A mini-introduction to information theory, Riv. Nuovo Cimento **43**, 187 (2020).
[35] J. Gaite and D. O'connor, Field theory entropy, the h theorem, and the renormalization group, Phys. Rev. D **54**, 5163 (1996).
[36] C. Bény and T. J. Osborne, Information-geometric approach to the renormalization group, Phys. Rev. A **92**, 022330 (2015).
[37] S. M. Apenko, Information theory and renormalization group flows, Physica (Amsterdam) **391A**, 62 (2012).
[38] C. Bény and T. J. Osborne, The renormalization group via statistical inference, New J. Phys. **17**, 083005 (2015).
[39] V. Balasubramanian, J. J. Heckman, and A. Maloney, Relative entropy and proximity of quantum field theories, J. High Energy Phys. 05 (2015) 104.
[40] P. Pessoa and A. Caticha, Exact renormalization groups as a form of entropic dynamics, Entropy, **20**, 1 (2018).
[41] N. Lashkari, Entanglement at a scale and renormalization monotones, J. High Energy Phys. 01 (2019) 219.
[42] K. Furuya, N. Lashkari, and S. Ouseph, Real-space RG, error correction and Petz map, J. High Energy Phys. 01 (2022) 170.
[43] R. Fowler and J. J. Heckman, Misanthropic entropy and renormalization as a communication channel, Int. J. Mod. Phys. A **37**, 2250109 (2022).
[44] A. Koenigstein, M. J. Steil, N. Wink, E. Grossi, and J. Braun, Numerical fluid dynamics for FRG flow equations: Zero-dimensional QFTs as numerical test cases—Part II: Entropy production and irreversibility of RG flows, Phys. Rev. D **106**, 065013 (2022).
[45] J. Erdmenger, K. T. Grosvenor, and R. Jefferson, Towards quantifying information flows: Relative entropy in deep neural networks and the renormalization group, SciPost Phys. **12**, 041 (2022).
[46] G. Zumbach, Almost Second Order Phase Transitions, Phys. Rev. Lett. **71**, 2421 (1993).
[47] G. Zumbach, The local potential approximation of the renormalization group and its applications, Phys. Lett. A **190**, 225 (1994).
[48] G. Zumbach, The renormalization group in the local potential approximation and its applications to the O(n) model, Nucl. Phys. **B413**, 754 (1994).
[49] J. Generowicz, C. Harvey-Fros, and T. R. Morris, C function representation of the local potential approximation, Phys. Lett. B **407**, 27 (1997).
[50] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient Flows: In Metric Spaces and in the Space of Probability Measures* (Springer Science & Business Media, New York, 2005).
[51] C. Villani, *Topics in Optimal Transportation* (American Mathematical Society, Providence, RI, 2003).
[52] R. Jordan, D. Kinderlehrer, and F. Otto, The variational formulation of the Fokker–Planck equation, SIAM J. Math. Anal. **29**, 1 (1998).
[53] M. Cuturi, Sinkhorn distances: Lightspeed computation of optimal transport, in *Proceedings of the 26th International Conference on Neural Information Processing Systems—Volume 2*, NIPS'13 (Curran Associates Inc., Red Hook, NY, USA, 2013), p. 2292-2300.
[54] G. Peyré, Entropic approximation of Wasserstein gradient flows, SIAM J. Imaging Sci. **8**, 2323 (2015).
[55] J.-D. Benamou, G. Carlier, Q. Mérigot, and É. Oudet, Discretization of functionals involving the Monge–Ampère operator, Numer. Math. **134**, 611 (2015).
[56] G. Carlier, V. Duval, G. Peyré, and B. Schmitzer, Convergence of entropic schemes for optimal transport and gradient flows, SIAM J. Math. Anal. **49**, 1385 (2017).
[57] W. Li, J. Lu, and L. Wang, Fisher information regularization schemes for Wasserstein gradient flows, J. Comput. Phys. **416**, 109449 (2020).
[58] J. A. Carrillo, K. Craig, L. Wang, and C. Wei, Primal dual methods for Wasserstein gradient flows, Found. Comput. Math. **22**, 389 (2021).
[59] M. Jacobs, W. Lee, and F. Léger, The back-and-forth method for Wasserstein gradient flows, ESAIM Control Optim. Calc. Var. **27**, 28 (2021).
[60] P. Mokrov, A. Korotin, L. Li, A. Genevay, J. Solomon, and E. Burnaev, Large-Scale Wasserstein Gradient Flows, Adv. Neural Inf. Process. Syst. **34**, 15243 (2021).
[61] C. Bonet, N. Courty, F. Septier, and L. Drumetz, Sliced-Wasserstein Gradient Flows, arXiv:2110.10972.
[62] D. Alvarez-Melis, Y. Schiff, and Y. Mroueh, Optimizing functionals on the space of probabilities with input convex neural networks, 2021.
[63] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, 2016).
[64] A. Krizhevsky, I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks, in *Advances in Neural Information Processing Systems*, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Association for Computing Machinery, New York, USA, 2012), Vol. 25.
[65] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016).
[66] M. Karami, D. Schuurmans, J. Sohl-Dickstein, L. Dinh, and D. Duckworth, Invertible convolutional flow, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc,

E. Fox, and R. Garnett (Curran Associates, Inc., 2019), Vol. 32.
[67] C. Bény, Deep learning and the renormalization group, arXiv:1301.3124.
[68] P. Mehta and D. J. Schwab, An exact mapping between the variational renormalization group and deep learning, arXiv:1410.3831.
[69] H. W. Lin, M. Tegmark, and D. Rolnick, Why does deep and cheap learning work so well?, J. Stat. Phys. **168**, 1223 (2017).
[70] M. Koch-Janusz and Z. Ringel, Mutual information, neural networks and the renormalization group, Nat. Phys. **14**, 578 (2018).
[71] S. S. Funai and D. Giataganas, Thermodynamics and feature extraction by machine learning, Phys. Rev. Res. **2**, 033415 (2020).
[72] S. Iso, S. Shiba, and S. Yokoo, Scale-invariant feature extraction of neural network and renormalization group flow, Phys. Rev. E **97**, 053304 (2018).
[73] E. de Mello Koch, R. de Mello Koch, and L. Cheng, *Is Deep Learning a Renormalization Group Flow?*, (IEEE, 2020).
[74] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein, Deep information propagation, arXiv:1611.01232.
[75] L. Xiao, Y. Bahri, J. Sohl-Dickstein, S. S. Schoenholz, and J. Pennington, Dynamical isometry and a mean field theory of cnns: How to train 10,000-layer vanilla convolutional neural networks, in *International Conference on Machine Learning* (2018).
[76] J. Ho, A. Jain, and P. Abbeel, Denoising diffusion probabilistic models, in *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Curran Associates, Inc., 2020), vol. 33, pp. 6840–6851.
[77] J. Haegeman, T. J. Osborne, H. Verschelde, and F. Verstraete, Entanglement Renormalization for Quantum Fields in Real Space, Phys. Rev. Lett. **110**, 100402 (2013).
[78] J. Cotler, M. R. M. Mozaffar, A. Mollabashi, and A. Naseh, Renormalization group circuits for weakly interacting continuum field theories, Fortschr. Phys. **67**, 1900038 (2019).
[79] J. S. Cotler, M. R. M. Mozaffar, A. Mollabashi, and A. Naseh, Entanglement renormalization for weakly interacting fields, Phys. Rev. D **99**, 085005 (2019).
[80] Q. Hu, A. Franco-Rubio, and G. Vidal, Continuous tensor network renormalization for quantum fields, arXiv:1809.05176.
[81] J. R. Fliss, R. G. Leigh, and O. Parrikar, Unitary networks from the exact renormalization of wave functionals, Phys. Rev. D **95**, 126001 (2017).
[82] E. A. Carlen and J. Maas, 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 (2014).
[83] E. A. Carlen and J. Maas, Gradient flow and entropy inequalities for quantum Markov semigroups with detailed balance, J. Funct. Anal. **273**, 1810 (2017).