# The inverse of Exact Renormalization Group Flows as Statistical Inference

Based on 2212.11379 [hep-th] and 2204.12939 [cond-mat.dis-nn]

David S. Berman, Marc S. Klinger and Jon Heckman

# Motivation

This work was largely motivated by two questions:

1. Can we describe Renormalization Group flows in an information theoretic way?
2. Can we use ideas in theoretical physics to better understand machine learning?

**Plan**

- We will introduce the basics of ERG and then less familiarly the idea of dynamical Bayesian Inference which is the basis of statistical inference.
- Then we will give the construction of an explicit map between renormalization and statistical learning.
- At the end of the presentation I will discuss a few ways that this correspondence can be used to bring new understandings from data science into physics and visa-versa.

# Wilsonian RG

The Wilsonian picture of RG describes an action functional which is changing as a function of the scale $\Lambda$:

$$S_\Lambda[\Phi \mid \lambda] = \sum_i \int_M \mathrm{Vol}_M(x)\ \lambda^i_\Lambda \mathscr{L}_i[\Phi] \tag{1}$$

Here $\{\mathscr{L}_i[\Phi]\}$ are the set of terms that are allowed in the action by symmetry. $\lambda^i$ are the coupling constants which govern the strength of the contribution of each term to the overall action.

- We can regard $\lambda^i$ as coordinates in the space of possible actions, $\mathscr{S}$.
- A Wilsonian RG flow is then generated by a vector field, $\underline{\beta} \in T\mathscr{S}$ such that:

$$-\frac{d}{d\ln\Lambda}\lambda^i_\Lambda = \beta^i(\lambda_\Lambda) \tag{2}$$

- This generates a one parameter family of actions at each scale, $\Lambda$.

# A Probabilistic Interpretation and ERG

From the perspective of Euclidean QFT, we can interpret $S_\Lambda$ as the log of an unnormalized probability distribution:

$$\hat{P}_\Lambda[\Phi] = e^{-S_\Lambda[\Phi|\lambda]} \tag{3}$$

$\lambda^i$ coordinatize an exponential family of probability distributions.

- An ERG flow is a one parameter family of probability distributions, $P_\Lambda$, governed by a functional differential equation:

$$-\frac{d}{d\ln\Lambda} P_\Lambda = \mathcal{F}[P_\Lambda] \tag{4}$$

- In this presentation, we shall regard an ERG as a diffusive flow on the space of probability distributions.

# Bayesian Inference

Bayesian inference is an approach to deducing the data generating distribution for a random variable $Y$.

- To begin, we postulate a parametric form of the distribution of $Y$ depending upon a set of parameters, $U$. For example, an exponential family with parameters $u^i$:

$$p_{Y|U}(y \mid u) = e^{u^i q_i(y) - \psi(u)} \tag{5}$$

- We can then complete a joint probability distribution for $(Y, U)$ by specifying a prior probability distribution, $\pi_0(u)$, for the parameters $U$: $p_{Y,U}(y, u) = \pi_0(u) p_{Y|U}(y \mid u)$.
- By observing the data $\{Y_t\}_{t=1}^T$, the distribution over parameters can be inferred through the implementation of Bayes' theorem:

$$\pi_T(u) \propto \pi_0(u) \prod_{t=1}^{T} p_{Y|U}(y_t \mid u) \tag{6}$$

# Dynamical Bayesian Inference

In **cond-mat.dis-nn/2204.12939** we considered performing Bayesian inference continuously, and tracking how the posterior distribution changes with respect to the inflow of new data. The main result of that paper may be written:

$$\frac{\partial}{\partial T}\pi_T(u) = -\left(D(p_{Y|U}) - \mathbb{E}_{\pi_T}(D(p_{Y|U}))\right)\pi_T(u) \tag{7}$$

Here $D(p) = D_{KL}(p^*_Y \parallel p)$ is the $KL$-Divergence between a probability distribution $p$ for the random variable $Y$, and the data generating distribution $p^*_Y$.

- Thus, much like ERG, Dynamical Bayesian Inference results in a one parameter family of posterior predictive distributions for the random variable $Y$:

$$p_T(y) = \int_{S_Y} \mathrm{Vol}_U(u)\pi_T(u)p_{Y|U}(y \mid u) \tag{8}$$

- Unlike an RG flow, however, this flow *gains* information over time.

# So, what's the idea?

- Renormalization describes how we can arrive at an effective description of a physical system by starting with a more complete model and systematically eliminating the information which is beyond our ability to observe experimentally.
- *In renormalization, we throw information away by acknowledging our ignorance*.
- Statistical Inference describes how, beginning from a state of ignorance about a system, we can arrive at a more complete model by observing the system at increasing levels of accuracy.
- *In statistical inference, we learn new information and thereby dissolve our ignorance.*
- The upshot is that Bayesian statistical inference can be regarded as an inverse process to an ERG flow.

# Overview

# ERG in the Wegner-Morris Form

Let us briefly review ERG in the Wegner-Morris formulation **Wegner 1973, Morris 1994, Cotler, Rezchikov 2022**:

$$-\frac{d}{d\ln\Lambda}P_\Lambda[\phi] = \int \mathrm{Vol}_M(x)\frac{\delta}{\delta\phi}\left(\Psi_\Lambda[\phi,x]P_\Lambda[\phi]\right) \tag{9}$$

- At each scale the field degree of freedom is merely reparameterized:

$$\phi \mapsto \phi' = \phi + (\delta\ln\Lambda)\ \Psi_\Lambda[\phi,x] \tag{10}$$

- The necessary ingredients in defining the "Wegner-Morris Scheme" are:
  1. A regulated two-point function $\dot{C}_\Lambda(x,y)$
  2. A functional $\Sigma_\Lambda[\phi] = S_\Lambda[\phi] - 2\hat{S}_\Lambda[\phi]$

  through which the *reparameterization kernel* is defined:

$$\Psi_\Lambda[\phi,x] = -\frac{1}{2}\int_M \mathrm{Vol}_M(y)\ \dot{C}_\Lambda(x,y)\frac{\delta\Sigma_\Lambda[\phi]}{\delta\phi(y)} \tag{11}$$

# The Space of Probability Distributions

- Let $(S, g)$ be a Riemannian manifold endowed with a reference measure $\mathrm{Vol}_S$. We shall refer to $S$ as the sample space.
- Let $\mathcal{M}$ denote the space of probability distributions on $S$. That is, $p \in \mathcal{M}$ is a map $p : S \to \mathbb{R}$ which is normalized

$$\int_S \star p = \int_S \mathrm{Vol}_S(x)\, p(x) = 1 \tag{12}$$

- The space tangent to $\mathcal{M}$, $T\mathcal{M}$, consists of deformations to probability distributions leaving the normalization unchanged. In other words, $\eta \in T\mathcal{M}$ is a map $\eta : S \to \mathbb{R}$ which integrates to *zero*

$$\int_S \star \eta = 0 \tag{13}$$

# One Parameter Flows on $\mathcal{M}$

- The "exterior derivative" on $\mathcal{M}$ can be identified with functional differentiation. Given a metric $G : T\mathcal{M} \times T\mathcal{M} \to \mathbb{R}$, we define the *gradient* of a map $F : \mathcal{M} \to \mathbb{R}$ in the direction of the vector field $\eta$ through the equation:

$$\delta F(\eta) = G(\text{grad}_G F, \eta) \tag{14}$$

- Example: The Dirichlet energy functional is defined by $\mathcal{E}[p] = \frac{1}{2}\int_S dp \wedge \star dp$. The $\ell^2$ metric on $\mathcal{M}$ is given by $G_{\ell^2}(\eta_1, \eta_2) = \int_S \eta_1 \wedge \star\eta_2$. Then, we can compute:

$$\delta\mathcal{E} = G_{\ell^2}(-\Delta p, \delta p) \tag{15}$$

  from which we recognize the gradient of $\mathcal{E}$ with respect to the $\ell^2$ metric on $\mathcal{M}$ as the Laplacian of $p$ viewed as a function on $S$.

- This suggests the following form for the heat equation:

$$\frac{\partial}{\partial t} p_t = -\text{grad}_{\ell^2}\mathcal{E}[p_t] \tag{16}$$

# One Parameter Flow on $\mathcal{M}$ continued

We can reconcile the heat equation as a gradient flow on $\mathcal{M}$ with respect to a wide array of different functions by choosing an appropriate metric.

- a good choice of metric is the Wasserstein two metric, inspired by Optimal Transport theory **[Villani2009]**:

$$G_{\mathcal{W}_2}(\eta_1, \eta_2) = \int_S p \, d\hat{\eta}_1 \wedge \star d\hat{\eta}_2 = -\int_S \eta_1 \wedge \star \hat{\eta}_2 = -\int_S \hat{\eta}_1 \wedge \star \eta_2 \tag{17}$$

- Given $\eta \in T\mathcal{M}$, we define $\hat{\eta}$ by the equation

$$\eta = \star d i_{p(d\hat{\eta})_\sharp} \mathrm{Vol}_S; \quad \eta = \mathrm{div}(p \, \mathrm{grad}\hat{\eta}) \tag{18}$$

- With respect to the Wasserstein metric, the gradient of the differential entropy, $S[p] = -\int_S p \wedge \star \ln(p)$, is equal to the Laplacian of $p$:

$$\mathrm{grad}_{\mathcal{W}_2} S[p] = \Delta p \tag{19}$$

# One Parameter Flows on $\mathcal{M}$ completed

With respect to a generic functional $\mathcal{F} : \mathcal{M} \to \mathbb{R}$, one can compute:

$$\text{grad}_{\mathcal{W}_2}\mathcal{F}[p] = \star di_{pd(\frac{\delta \mathcal{F}}{\delta p}|_p)}\text{Vol}_S \tag{20}$$

- A very general set of functionals sourcing convection diffusion flows can be written in the form:

$$\mathcal{F}[p] = S[p] + \mathcal{V}[p] = -\int_S p \wedge \star \ln(p) + \int_S p \wedge \star V \tag{21}$$

  where $V : S \to \mathbb{R}$ is a potential function that is independent of $p$.
- The gradient of such a functional gives rise to a pair of terms, and its associated one parameter flow equation is equivalent to the Fokker-Planck equation:

$$\frac{\partial}{\partial t}p_t = \text{grad}_{\mathcal{W}_2}\mathcal{F}[p_t] = \Delta p_t + \text{div}(p_t \, \text{grad} V) \tag{22}$$

# Fokker-Planck and Optimal Transport

Provided the normalization $Z = \int_S \star e^{-V}$ is finite, $q = \frac{1}{Z} e^{-V}$ defines a probability distribution on $S$, which is a stationary state of the Fokker-Planck equation with potential $V$.

- In this case, the gradient flow with respect to $\mathcal{F}$ can moreover be understood as the gradient flow with respect to the KL-Divergence between $p$ and $q$:

$$D_{KL}(p \parallel q) = \mathcal{F}[p] + \ln(Z) \tag{23}$$

  Hence, $\frac{\delta \mathcal{F}}{\delta p} = \frac{\delta D_{KL}(p \| q)}{\delta p}$.

- The Fokker-Planck equation can therefore be understood as the gradient flow:

$$\frac{\partial}{\partial t} p_t = \mathrm{grad}_{\mathcal{W}_2} D_{KL}(p_t \parallel q) \tag{24}$$

# Wegner-Morris and Fokker-Planck

We can now realize the Wegner-Morris equation governing ERG as a gradient flow of a relative entropy.

- Let $p = \frac{e^{-S_p(x)}}{Z_p}$. Similarly, we can define the fixed point distribution $q = \frac{e^{-S_q(x)}}{Z_q}$.
- Specifying $q$ along with the metric $g$ on $S$ is equivalent to the data of $\Sigma_\Lambda$ and $\dot{C}_\Lambda$ in the functional case. In particular we take $\Sigma = S_q - S_p$, and thus define the reparameterization kernel:

$$\Psi[\phi, x] \propto \int_M \mathrm{Vol}_M(y)\dot{C}_\Lambda(x,y)\frac{\delta\Sigma[\phi]}{\delta\phi(y)} \to \Psi = \mathrm{grad}_g\Sigma = g^{ij}\frac{\partial}{\partial x^i}\Sigma \tag{25}$$

- It is then a straightforward task to compute:

$$\frac{\partial}{\partial t}p_t = -\mathrm{grad}_{\mathcal{W}_2}D_{KL}(p_t \parallel q_t) = - \star d(\star p_t d\Sigma) = -\frac{\partial}{\partial x^i}\left(\Psi^i_t p_t\right) \tag{26}$$

  which we recognize as the finite dimensional Wegner-Morris equation.

# ERG = Convection-Diffusion

The upshot of this first section is that an exact renormalization group flow can be understood as a *Convection-Diffusion* equation.

- The two pieces of data that quantify the ERG scheme now have a clear interpretation.
- The regulated two point function $\dot{C}_\Lambda$ plays the role of an inverse metric on the space of field configurations. Its purpose is to define a notion of the changing width of the distribution as a function of time.
- The functional $\Sigma$ tracks the difference between the scale dependent action and a fixed point action. Its purpose is to describe the mean tendency of the probability distribution as it flows in the space of models.
- In the next sections we will observe that these pieces of data respectively define the drift and diffusion aspects of a stochastic differential equation.

# Overview

# Stochastic Processes and Stochastic Differential Equations

A stochastic process is a one parameter family of random variables, $\{X_t\}_{t\geq 0}$. We will be interested in stochastic processes governed by stochastic differential equations of the Ito form [**ito1944109**]:

$$dX_t^i = m^i(X_t, t)dt + \sigma_j^i(X_t, t)dW_t^i \tag{27}$$

- This equation can be compared to the *Langevin Equation*, $\frac{dX^i}{dt} = m^i(X, t) + \eta_t^i$, which describes a variable whose mean path is determined by the law $\frac{dX}{dt} = m$, but is also subject to random fluctuations around the mean path.
- A stochastic process is called a *Martingale* if it satisfies the equation

$$\mathbb{E}(X_t \mid \{X_s\}_{s\leq\tau}) = X_\tau; \quad \forall\, \tau \leq t \tag{28}$$

- It isn't hard to show that a stochastic process governed by an Ito SDE is a Martigale if it has vanishing drift.

# From Stochastic Differential Equations to Partial Differential Equations

Given a map $f(x,t)$, we can define a stochastic process: $\{f(X_t,t)\}_{t\geq 0}$. Its governed SDE can be realized by implementing the principal of *quadratic variation*:

$$df(X_t,t) = \left(\frac{\partial f}{\partial t} + m^i \frac{\partial f}{\partial x^i} + \frac{1}{2}\delta^{kl}\sigma^i_k\sigma^j_l \frac{\partial^2 f}{\partial x^i \partial x^j}\right) dt + \frac{\partial f}{\partial x^i} dW^i_t \tag{29}$$

- The stochastic process $\{f(X_t,t)\}_{t\geq 0}$ will be a martingale if and only if $f$ satisfies the partial differential equation:

$$\frac{\partial f}{\partial t} + m^i \frac{\partial f}{\partial x^i} + \frac{1}{2}\delta^{kl}\sigma^i_k\sigma^j_l \frac{\partial^2 f}{\partial x^i \partial x^j} = 0 \tag{30}$$

- In the case where the drift of $X_t$ is governed by the gradient of a potential function,

$$dX_t = -(\mathrm{grad}_g V)^i dt + \sqrt{2}\delta^i_j dW^j_t \tag{31}$$

this partial differential equation can be directly related to the Fokker-Planck equation.

# From Partial Differential Equations to Stochastic Differential Equations

We can also go the other way around. Let $\mathcal{A}$ denote a differential operator associated with a convection-diffusion equation:

$$(\frac{\partial}{\partial t} + \mathcal{A})p_t = 0 \tag{32}$$

- The above equation is solved by introducing the Heat Kernel, $e^{-t\mathcal{A}}$. Given initial data $p_0$, the formal solution is then given by $e^{-t\mathcal{A}}(p_0)$:

$$\frac{\partial}{\partial t} \circ e^{-t\mathcal{A}}(p_0) = -\mathcal{A} \circ e^{-t\mathcal{A}}(p_0) \tag{33}$$

- Under suitable conditions, one can interpret $e^{-t\mathcal{A}}$ as the transition operator of a Markov process. For example, working in the "positional basis" we can write

$$U(x, y; t) \equiv \langle x| \, e^{-t\mathcal{A}} \, |y\rangle \tag{34}$$

which is the probability density of translations from sample point $y$ to sample point $x$ in an interval $t$.

In this language, it is natural to identify the operator $\mathcal{A}$ as a "Hamiltonian Operator," with $P_t = e^{-t\mathcal{A}}$ its associated "time evolution operator".

- Given a marginal probability distribution $f : S \to \mathbb{R}$ the action $P_t$ is given by

$$P_t(f)(x,t) = \int_S \mathrm{Vol}_S(y) U(x,y;t) f(y) \tag{35}$$

- If the action of $P_t$ is known, we can also derive a functional form for $\mathcal{A}$ by using the fact that $P_t$ is the exponentiation of $\mathcal{A}$:

$$\mathcal{A}(f) = \lim_{t\to 0} \frac{P_t(f) - f}{t} \tag{36}$$

- If $\mathcal{A}$ is of the form $\mathcal{A} = m^i \frac{\partial}{\partial x^i} + \frac{1}{2}\delta^{kl}\sigma^i_k \sigma^j_l \frac{\partial^2}{\partial x^i \partial x^j}$, we can associate solutions to the PDE with martingales adapted to the stochastic process

$$dX^i_t = m^i(X_t, t)dt + \sigma^i_j dW^j_t \tag{37}$$

# Overview

# The Bayesian Inversion Problem

- Let $S_Y$, and $S_U$ denote the sample spaces of random variables $Y$ and $U$, respectively.
- We assume that there exists a process by which the signal $U$ is mapped into the observable output $Y$, $G : S_U \to S_Y$:

$$Y = G(U) + N; \quad (\frac{dX}{dt} = m + \eta) \tag{38}$$

  Here $U \in S_U$ is called the *signal* and $N \in S_Y$ is *noise*.
- We are interested in the inversion problem of determining the signal, $u$, that predicated a measured output, $y$. However, it is often intractable or even analytically impossible to fully invert the process. Thus, it is more prudent to seek a *probability distribution* over the possible inputs. This leads to the idea of *Bayesian Inversion* [**dashti2017bayesian**].

# The Bayesian Inversion Problem continued

- To this end, we postulate for $U$ a prior measure $\pi_0 \in \mathcal{M}_U$.
- We also treat $N$ as a random variable independent from $Y$ governed by a distribution $p_0 \in \mathcal{M}_{S_Y}$.
- This allows us to define a conditional random variable:

$$Y \mid U = G(U) + N \tag{39}$$

and moreover, realize a channel $\phi : S_Y \to S_Y$ from *observed output* to *noise* given by

$$n = \phi(y) = y - G(u) \tag{40}$$

- Thus, we can define a conditional measure for $Y \mid U$ by pulling back $p_0$ by the map $\phi$. The resulting distribution, $p_{Y|U}$ provides us with the conditional probability distribution for the realized output given the signal.

$$p_{Y|U}(y \mid u) = p_0(y - G(u)) \tag{41}$$

# Dynamical Bayes

- Suppose $\{Y_t\}_{t=1}^T$ is a set of $T$ IID random variables drawn from the data generating distribution $p_Y^*$. Then, we have shown that the posterior distribution $\pi_T$ satisfies an integro-differential equation:

$$\frac{\partial}{\partial T}\pi_T(u) = -\left(D(p_{Y|U}) - \mathbb{E}_{\pi_T}(D(p_{Y|U}))\right)\pi_T(u) \tag{42}$$

- We shall assume that the data generating distribution can be realized in the form of the conditional density at a specific value of the underlying signal: $p_Y^*(y) = p_{Y|U}(y \mid u^*)$.
- One of the most significant insights of this approach is that for sufficiently large $T$ the $2n$-point functions of the $T$-posterior distribution are given by

$$\mathbb{E}_{\pi_T}\left(\prod_{j=1}^{2n}(U-u^*)^{i_j}\right) = \frac{1}{T^n}\sum_{p\in\mathcal{P}_{2n}^2}\prod_{(r,s)\in p}\mathcal{I}^{i_r i_s}\Big|_{u^*} \tag{43}$$

Here $\mathcal{I}^{ij}$ is the inverse of the Fisher metric for the family of probability distributions

# Bayesian Inversion for Normal Data

Consider the following Bayesian Inversion problem:

$$y = \mu + n \tag{44}$$

where $n$ is noise distributed according to $\mathcal{N}(0, \sigma^2)$. The process $G$ is simply the identity, and hence the inference problem is on the value of the mean parameter, $\mu$.

- In this case, the Dynamical Bayesian Inference equation can be solved exactly:

$$\pi_T(\mu) = \frac{1}{\sqrt{2\pi(\sigma^2/T)}} e^{-\frac{1}{2(\sigma^2/T)}(\mu-\mu^*)^2} \tag{45}$$

  where $\mu^*$ is the data generating signal.
- If we define the "inverse" time parameter $\tau = \frac{1}{T}$, this solution takes the suggestive form

$$\pi_\tau = \frac{1}{\sqrt{2\pi\sigma^2\tau}} e^{-\frac{(\mu-\mu^*)^2}{2\sigma^2\tau}} \tag{46}$$

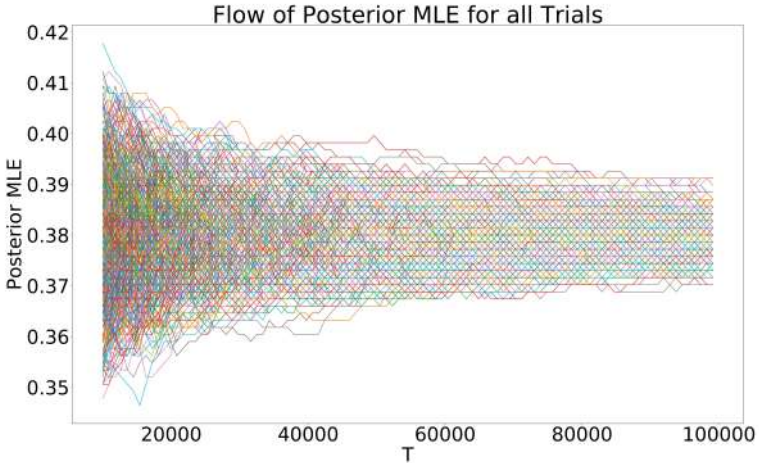
# Bayesian Diffusion

The posterior distribution in the inverse time parameter $\tau$ is of the form of the standard solution to the diffusion equation with diffusivity $\sigma^2$.

- This suggest a very interesting interpretation of what we call *Bayesian Diffusion*: As one moves *backwards* in a Bayesian inference by *throwing away* data instead of observing new data, one realizes a diffusion process in the prediction of the signal.
- It is important to notice that the "inverse" time parameter is given by $\frac{1}{T}$. This is consistent with the interpretation of decreasing the amount of observed data.
- It is also important to recognize the role played by $\sigma^2$ in setting a scale for the resulting diffusion process.
- Although the posterior distribution diffuses in the backwards inference time, it is always centered around the data generating signal $\mu^*$. This seems to suggest that there is no *drift*.

# Bayesian Drift and Scheme Independence

- The *trajectory* of the maximum a posteriori estimate should be regarded as a *scheme dependent* quantity.



# A Simple Argument

Let $E = \{Y_t\}_{t=1}^N$ denote a fixed set of IID random variables. Moreover, let $\pi \in \text{Perm}(N)$ denote a permutation of $N$ elements. Then, $\pi(E) = \{Y_{\pi(t)}\}_{t=1}^N$ is also a set of $N$ IID random variables. In fact, it consists of precisely the *same* random variables as $E$, only in a different order.

- Define
$$\mu(E; n) = \frac{1}{n} \sum_{i=1}^{n} Y_i \tag{47}$$
- Then, it is clear that for any $n < N$
$$\mu(E; n) \neq \mu(\pi(E); n) \tag{48}$$
- However, if we include all of the observations it will be the case that
$$\mu(E; N) = \mu(\pi(E); N), \quad \forall\, \pi \in \text{Perm}(N) \tag{49}$$
- More generally, after a sufficiently large number of observations are made, any sequence of IID observations will tend towards the same data generating distribution.

# Bayesian Drift

Thus, we see that the trajectory of the maximum a posteriori estimate is determined entirely by the sequence in which data is observed. The terminal distribution, however, is invariant under permutations of the order in which data is incorporated into the posterior model. This should be compared to scheme independence an ERG.

- We can incorporate the effect of different sequences of observation by specifying the trajectory of the maximum a posteriori estimate.
- Let $\gamma : \mathbb{R} \to S_U$ denote such a trajectory. It can be generated as the integral flow of a vector field $\dot{\gamma}$.
- In particular, we will consider the case in which the maximum a posteriori estimate follows a gradient descent with respect to the log-likelihood function $\Phi(u; y) = \ln(p_0(y - G(u)))$:

$$\dot{\gamma} = -\text{grad}_u \Phi(u; y)\Big|_\gamma \tag{50}$$

# A Partial Differential Equation for Bayesian Inference

We are now prepared to derive a partial differential equation governing the backwards drift-diffusion of the posterior predictive distribution.

- Given the $\tau$ posterior, $\pi_\tau(u)$, and the conditional density $p_{Y|U}$, we can define the $\tau$ dependent posterior predictive distribution:

$$p_\tau(y) = \int_{S_U} \mathrm{Vol}_U(u)\pi_\tau(u)p_0(y - G(u)) \tag{51}$$

- It is natural to interpret $\pi_\tau$ as the transitionary probability density. It can be associated with a Markov operator

$$P_\tau(p_0)(y) = \mathbb{E}_{\pi_\tau}(p_0(y - G(U))) \tag{52}$$

- We can now derive the differential operator exponentiating to $P_\tau$ by computing

$$\mathcal{A}(p_0) = \lim_{\tau \to 0} \frac{P_\tau(p_0) - p_0}{\tau} \tag{53}$$

# A Partial Differential Equation for Bayesian Inference

Taylor expanding around $u = 0$ we can write:

$$P_\tau(p_0)(y) = \mathbb{E}_{\pi_\tau}\left(p_0(y) - \frac{\partial p_0}{\partial y^i}\frac{\partial G^i}{\partial u^j}\dot{\gamma}^j\tau + \frac{1}{2}\frac{\partial^2 p_0}{\partial y^i \partial y^j}\frac{\partial G^i}{\partial u^k}\frac{\partial G^j}{\partial u^l}\delta u^k \delta u^l + \mathcal{O}(\delta u^3)\right) \quad (54)$$

- Using the previously stated formula for the $2n$-point functions, and assuming that the maximum a posteriori estimate follows a gradient descent with respect to the log-likelihood in the geometry defined by the Fisher Metric, we can compute explicitly

$$P_\tau(p_0)(y) = p_0(y) - m^i\frac{\partial p_0}{\partial y^i} + K^{ij}(\gamma)\frac{\partial^2 p_0}{\partial y^i \partial y^j} + \mathcal{O}(\tau^2) \quad (55)$$

where

$$K^{ij}(\gamma) = \frac{\partial G^i}{\partial u^k}\frac{\partial G^j}{\partial u^l}\mathcal{I}^{kl}(\gamma); \quad m^i = -K^{ij}(\gamma)\frac{\partial \Phi}{\partial y^j} \quad (56)$$

# A Partial Differential Equation for Bayesian Inference

Thus, we can easily compute:

$$\frac{\partial p_0}{\partial \tau} = \mathcal{A}(p_0) = -m^i \frac{\partial p_0}{\partial y^i} + K^{ij}(\gamma) \frac{\partial^2 p_0}{\partial y^i \partial y^j} \tag{57}$$

Which we can interpret as the differential equation satisfied by a function adapted to the stochastic process

$$dY_\tau = -\text{grad}\Phi d\tau + dW_\tau \tag{58}$$

in a geometry with metric $K$.

- We regard this equation as *defining* an ERG flow in the sense of a drift-diffusion process.

# The ERG/Dynamical Bayesian Inference Correspondence

| | ERG | Bayesian Diffusion |
|---|---|---|
| 1. | Time parameter: $t$ | Inverse Observation Time: $\tau = \frac{1}{T}$ |
| 2. | The inverse metric $g^{ij}$ | The inverse Fisher metric $\mathcal{I}^{ij}(\gamma_\tau)$ |
| 3. | The potential function $V$ | The Log Likelihood function $\Phi(\gamma_\tau; y)$ |
| 4. | The scheme function $\Sigma = S_q - S_p$ | The Log Likelihood ratio $\Sigma = \Phi(u; y) - \Phi(\gamma_\tau; y)$ |

- This dictionary allows one to translate between an ERG flow, and the diffusion process associated with the *backwards* trajectory of a dynamical Bayesian inference.
- Given a complete Dynamical Bayesian Inference trajectory, one can therefore move from the UV to the IR and visa versa, defining an meaningful notion of *inverting* an ERG flow.

# Renormalizability and Scale

Recall, in the field theory context the role of the inverse metric was played by the regulated two point function $\dot{C}_\Lambda$ which defines a running momentum scale for operators in the theory.

- In the Backwards Diffusion Process, the Fisher Metric $\mathcal{I}$ plays an analogous role as a generalized two point function.
- In this sense, we view the Fisher metric as defining a notion of an emergent scale.
- In Wilsonian RG, the question of renormalizability corresponds to whether or not higher order Feynman diagrams can be regulated by introducing a finite number of operator sourced counterterms.
- In the language of statistical inference, the operator content of a theory corresponds to the problem of model selection.
- In view of this fact, there seems to be a natural notion of renormalizability in terms of whether a probability distribution can be specified with only a finite number of connected moments, or, conversely, if one requires an infinite number of such moments.

# Quantum Error Correction

The correspondence introduced in this paper can be regarded as a classical instantiation of the correspondence between renormalization and quantum error correction [**cotler2019entanglement**; **furuya2020real**].

- Let $\mathcal{A}$ denote an operator algebra whose elements we regard as the observables of a theory. The set of *states* associated with $\mathcal{A}$ can be identified with its predual $\mathcal{A}_*$ which consists of linear functionals that, when paired with an observable, return a complex number corresponding to an expectation value.
- An RG flow can be formally understood as a quantum channel, which is a completely positive, trace preserving, linear map between states $\mathcal{E} : \mathcal{A}_* \to \mathcal{B}_*$.
- Such a map is a natural mathematical encoding of an RG flow due to the *Data Processing Inequality* [**ohya2004quantum**]

$$D_{KL}(\rho \parallel \rho') \leq D_{KL}(\mathcal{E}(\rho) \parallel \mathcal{E}(\rho')) \tag{59}$$

# Quantum Error Correction Continued

The task of *Quantum Error Correction* is to *recover* the information contained in a state $\rho$ after it has been noised by the channel $\mathcal{E}$.

- Studying Quantum Error Correction therefore provides several important lessons about how the information contained in a theory is corrupted under an RG flow, and under what circumstances that information can be exactly, or approximately recovered [**junge2018universal**].
- When the operator algebra, $\mathcal{A}$, is *commutative*, the space of states, $\mathcal{A}_*$, can be identified with the set of classical probability distributions assigning expectation values to random variables $a \in \mathcal{A}$.
- In this context, it can be shown the the universal recovery map, the *Petz Map*, exactly reproduces Bayes' Law.

# Diffusion Learning

Diffusion learning [**sohl2015deep**] provides an actionable response to the following question: Given an intractable distribution $p_0$, how can we effectively sample data?

- The implementation of the diffusion learning scheme bears a very strong resemblance to the ERG/Dynamical Bayesian inference correspondence.
- The initial data, $p_0$, is first run through a forward diffusion process generated by the operator $F$. This process destroys some fine grained information contained in $p_0$, but results in an analytically tractable distribution $p_F$ which is a fixed point distribution of $F$.
- The distribution $p_F$ is then run through a *backwards* diffusion process generated by an operator $B$. The operator $B$ is chosen such that the relative entropy between the terminal distribution of the backwards process, and the initial data $p_0$ is minimized.
- There may be a utility to using non-diffusive recovery channels. The improved accuracy of the reconstructed distribution must be balanced against the decreased efficiency of the recovery process.

# Discussion

Some takeaways:

- The proposed correspondence between ERG and Dynamical Bayesian Inference provides a clean quantification of the information content in an effective theory at every scale. $\rightarrow$ Applications in diagnosing the information lost during an RG flow, understanding RG monotonicity theorems.
- The notion of Bayesian inference as an inverse process to renormalization is reminicient of the error-correcting interpretation of ERG. $\rightarrow$ Applications in Bulk-Reconstruction, Holographic Renormalization, the AdS/CFT correspondence, and general information theoretic approaches to Quantum Gravity
- Viewing ERG abstractly as a one parameter family of probability distributions for a generic sample space allows the logic of RG to be carried over into contexts that are not immediately physical. $\rightarrow$ Applications in Machine Learning as a device for interpreting and constructing Artifical Intelligences.

# References I