Title: Aspects of RG flows and Bayesian Updating

Speakers: David Berman

Collection/Series: Theory + AI Workshop: Theoretical Physics for AI

Date: April 10, 2025 - 2:15 PM

URL: https://pirsa.org/25040108

Abstract:

We will examine the idea of Bayesian updating as an inverse diffusion like process and its relation to the exact renormalisation group. In particular we will look at the role of Fisher Information, its metric and possible physical interpretations.

Menu mainv2 - Copy.pdf Create Sign in
All tools Edit Convert E-Sign Find text or tools AI Assistant

Quantum Field Theory inspired learning

250y.xxxxx, 2402.00944, 2305.10491, 2212.11379 and 2204.12939

$$P(y) \approx \frac{1}{\sqrt{2\pi}} \exp\{-(1/2)(y-4)^2\}$$

David S. Berman, Marc S. Klinger, Jon J. Heckman and Alex G. Stapleton

1
36
1 / 36
34°F Light rain/snow Search ENG US 7:22 pm 10/4/2025

Motivation

We want to explore various QFT inspired ideas for learning.

- Data comes from some underlying probability distribution
- We construct parameterised models of such distributions and learning is about determining model parameters from data
- Bayes's Theorem provides the basis for such statistical inference
- In QFT, renormalisation flow describes the dependence of model parameters on energy
- Suggests a Bayes-RG flow dictionary
- Amount of data takes on the role of energy
- The Fisher information will play a central role in all that follows

34°F Light rain/snow Search ENG US 7:23 pm 10/4/2025

Menu mainv2 - Copy.pdf Create Sign in
All tools Edit Convert E-Sign Find text or tools AI Assistant

High Level View

- Cotler and Rezchikov showed Exact RG as Optimal Transport
- iterated Bayesian updating, *dynamical Bayes*, is described by a first order equation that maybe linked to a diffusion like equation
- Exact RG as diffusion and *inverse dynamical Bayes* as diffusion; Fokker-Planck description provides the bridge
- Identify ERG and *inverse Bayes*

34°F Light rain/snow Search ENG US 7:27 pm 10/4/2025

Plan

- Introduce information geometry basics
- Describe Exact Renormalisation Group and its flow
- Describe the idea of dynamical Bayesian Inference
- Then we will give the construction of an explicit map between renormalization and statistical learning.
- Inspired by the role of the Fisher information, we will examine the performance of an information pruned generative model

34°F Light rain/snow Search ENG US 7:28 pm 10/4/2025

Information Preliminaries

Consider a parametric family of probability distributions, $p_{Y|\Theta}(y \mid \theta)$. A priori, we should consider any allowed distribution for Y as a candidate for the data generating distribution.

Then consider the space of such distributions, $\mathcal{M} = \{p_{Y|\Theta}(y \mid \theta) \mid \theta \in \mathbb{R}^n\}$. So, $\mathcal{M}$ is a space whose points correspond to probability distributions for Y. By construction, this space has a local coordinate system given in terms of the parameters θ.

A natural basis for the tangent space of $\mathcal{M}$, is given to us in terms of the *score vectors* $\underline{\ell}_i = \frac{\partial}{\partial \theta^i} \ln\left(p_{Y|\Theta}(y \mid \theta)\right)$.

34°F Light rain/snow Search ENG US 7:28 pm 10/4/2025

The Fisher information matrix then provides a metric given by:

$$\mathcal{I}_{ij}(\theta) = \mathbb{E}_\theta(\ell_i \ell_j) = \mathbb{E}_\theta\left(\frac{\partial \ln\left(p_{Y|\Theta}(Y \mid \theta)\right)}{\partial \theta^i}\frac{\partial \ln\left(p_{Y|\Theta}(Y \mid \theta)\right)}{\partial \theta^j}\right). \tag{1}$$

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

$$D_{KL}(\theta \parallel \theta') = \mathbb{E}_\theta\left(\ln\left(\frac{p_{Y|\Theta}(Y \mid \theta)}{p_{Y|\Theta}(Y \mid \theta')}\right)\right). \tag{2}$$

measures the relative entropy between two distributions. D_{KL} is an information divergence, which means that $D_{KL}(\theta \parallel \theta') \geq 0$ and it is equal to zero if and only if $\theta = \theta'$. In the immediate neighborhood of a point $\theta \in \mathcal{M}$ the KL-divergence can be expanded to quadratic order as

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

Menu mainv2 - Copy.pdf Create Sign in
All tools Edit Convert E-Sign Find text or tools AI Assistant

Exact Renormalisation Group

A Euclidean QFT, is described by S_Λ which determines an unnormalized probability distribution:

$$\hat{P}_\Lambda[\Phi] = e^{-S_\Lambda[\Phi|\lambda]} \tag{4}$$

with λ^i providing coordinates for a space of such probability distributions.

- An ERG flow is a one parameter family of probability distributions, P_Λ, governed by a functional differential equation:

$$-\frac{d}{d\ln\Lambda} P_\Lambda = \mathcal{F}[P_\Lambda] \tag{5}$$

- ERG may be thought of as a diffusive flow on the space of probability distributions.

34°F Light rain/snow Search ENG US 7:31 pm 10/4/2025

Menu mainv2 - Copy.pdf Create Sign in
All tools Edit Convert E-Sign Find text or tools AI Assistant

Exact Renormalisation as Diffusion

Here Λ is a physically meaningful RG scale (typically associated with a momentum cutoff), and ϕ corresponds to the field configuration relevant to a given theory.
The guiding principle of the ERG is that the flow $P_\Lambda[\phi]$ must be chosen in such a way that the partition function is preserved:

$$\frac{d}{d\ln\Lambda}\int_{\mathcal{F}}\mathcal{D}\phi\, P_\Lambda[\phi] = 0. \tag{6}$$

The most familiar form of exact renormalization is the so-called Polchinski scheme.

$$P_\Lambda[\phi] \propto e^{-\frac{1}{2}\int \frac{d^d p}{(2\pi)^d}\,\phi(p)G(p^2)K_\Lambda^{-1}(p^2)\phi(-p)}\, e^{-S_{int,\Lambda}[\phi]}. \tag{7}$$

34°F Light rain/snow Search ENG US 7:31 pm 10/4/2025

The first term as the Gaussian distribution associated with a free field theory with propagator $G(p^2)$, but with the incorporation of a function $K_\Lambda^{-1}(p^2)$ which plays the role of a smooth cutoff function in momentum space.

$K_\Lambda^{-1}(p^2)$ suppresses the contribution of momentum modes above the cutoff scale Λ.
The second term is the exponential of the renormalized interacting action at the scale Λ.

In Polchinski's picture, $K_\Lambda(p^2)$ has a prescribed dependence on Λ, thus Polchinski's ERG equation arises by determining the equation which must be obeyed by $S_{int,\Lambda}[\phi]$ in order to satisfy the principle that the partition function is preserved.

34°F Light rain/snow Search ENG US 7:31 pm 10/4/2025

The resulting equation can be put into the form:

$$\frac{d}{d\ln\Lambda}P_\Lambda[\phi] = \int_{M\times M} d^dx d^dy \left\{ C^{Pol.}_\Lambda(x,y)\frac{\delta^2 P_\Lambda[\phi]}{\delta\phi(x)\delta\phi(y)} \right. \tag{8}$$

$$\left. + \frac{\delta}{\delta\phi(x)}\left(P_\Lambda[\phi]C^{Pol.}_\Lambda(x,y)\frac{\delta V^{Pol.}_\Lambda[\phi]}{\delta\phi(y)}\right)\right. \tag{9}$$

$$\equiv \Delta P_\Lambda[\phi] + \mathrm{div}\left(P_\Lambda[\phi]\mathrm{grad}_{C^{Pol.}_\Lambda} V^{Pol.}_\Lambda[\phi]\right), \tag{10}$$

where

$$C^{Pol.}_\Lambda(p^2) = (2\pi)^d G(p^2)^{-1}\frac{\partial K_\Lambda(p^2)}{\partial\ln\Lambda}; \quad V^{Pol.}_\Lambda[\phi] = \int \frac{d^dp}{(2\pi)^d}\phi(p)G(p^2)K^{-1}_\Lambda(p^2)\phi(-p). \tag{11}$$

The final expression is the *Fokker-Planck* equation with diffusion governed by $C^{Pol.}_\Lambda(p^2)$ and drift governed by the potential $V^{Pol.}_\Lambda[\phi]$.

This is the relationship between exact renormalization and diffusion so one can identify exact RG as a functional version of Fokker-Plank.

One can now also consider different schemes and regulators which lead to a more general ERG than Polchinski's This was done by Wegner and Morris.

The key point is that after the scheme is specified a Fokker-Plank equation is then determined with the data $C_\Lambda^{WM}(p^2)$ and drift governed by the potential $V_\Lambda^{WM}[\phi]$.

34°F Light rain/snow Search ENG US 7:32 pm 10/4/2025

Bayesian Inference

Bayesian inference is an approach to infer the data generating model from observations of the data Y.

- Consider a probabilistic model for Y depending upon a set of parameters, θ. For example, an exponential family with parameters θ^i:

$$p_{Y|\theta}(y \mid \theta) = e^{\theta^i q_i(y)} \tag{12}$$

- We assume some prior probability distribution, $\pi_0(\theta)$, for the parameters θ
- By observing the data $\{Y_t\}_{t=1}^T$, the distribution over parameters can be inferred through the implementation of Bayes' theorem, to give the posterior distribution:

$$\pi_T(\theta) \propto \pi_0(\theta) \prod_{t=1}^{T} p_{Y|\theta}(y_t \mid \theta) \tag{13}$$

34°F Light rain/snow Search ENG US 7:34 pm 10/4/2025

Dynamical Bayes

What does Bayesian inference would look like as a continuous time dynamical system. In other words, we think of data as being continuously observed and updates continuously applied. The posterior distribution is governed by:

$$\frac{\partial \pi_T(\theta)}{\partial T} = -\left(D_{KL}(\theta_* \parallel \theta) - \mathbb{E}_{\pi_T}\left(D_{KL}(\theta_* \parallel \theta)\right)\right)\pi_T(\theta). \tag{14}$$

Here θ_* is the parameter corresponding to ground truth. The equation (14) has a schematic solution

$$\pi_T(\theta) \propto e^{-TD_{KL}(\theta_* \| \theta)} \tag{15}$$

which we interpret as a Boltzmann distribution with "energy" given by the KL-divergence between the data generating model and a model at $\theta \in \mathcal{M}$.

34°F Light rain/snow Search ENG US 7:36 pm 10/4/2025

Menu mainv2 - Copy.pdf Create Sign in
All tools Edit Convert E-Sign Find text or tools AI Assistant

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

$$\pi_T(\theta) = \mathcal{N}(\mu_T, \frac{1}{T}\mathcal{I}(\mu_T)^{-1})(\theta). \tag{16}$$

Here μ_T is the T-path of the maximum a posterior (MAP) estimate.
μ_T as defines a new dynamical degree of freedom. One can then construct a potential function V for which μ_T is *defined* to be a gradient flow

$$\frac{d}{dT}\mu_T = \text{grad}_{\mathcal{I}} V|_{\mu_T}. \tag{17}$$

Eventually, as $T \to \infty$, $\mu_T \to \theta_*$.
One may interpret (16) as specifying that the posterior distribution π_T is localized around the MAP, μ_T, with a characteristic width given by $\frac{1}{T}\mathcal{I}(\mu_T)^{-1}$.

34°F Light rain/snow Search ENG US 7:43 pm 10/4/2025

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.

34°F Light rain/snow Search ENG US 7:45 pm 10/4/2025

Menu mainv2 - Copy.pdf Create Sign in
All tools Edit Convert E-Sign Find text or tools AI Assistant

Bayesian Inversion for Normal Data

Consider the following Bayesian Inversion problem:

$$y = \mu + n \tag{18}$$

where n is noise distributed according to $\mathcal{N}(0, \sigma^2)$. The inference problem is to find the parameter, μ.

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

$$\pi_T(\mu) = \frac{1}{\sqrt{2\pi(\sigma^2/T)}} e^{-\frac{1}{2(\sigma^2/T)}(\mu-\mu^*)^2} \tag{19}$$

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

$$\pi_\tau = \frac{1}{\sqrt{2\pi\sigma^2\tau}} e^{-\frac{(\mu-\mu^*)^2}{2\sigma^2\tau}} \tag{20}$$

16
36
34°F Light rain/snow Search ENG US 7:46 pm 10/4/2025

Menu mainv2 - Copy.pdf Create Sign in
All tools Edit Convert E-Sign Find text or tools AI Assistant

Bayesian Diffusion

The posterior distribution in the inverse time parameter τ is of the form of the standard solution to the diffusion equation with diffusivity σ^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 σ^2 in setting a scale for the resulting diffusion process.

17
36
34°F Light rain/snow Search ENG US 7:47 pm 10/4/2025

Menu mainv2 - Copy.pdf Create Sign in
All tools Edit Convert E-Sign Find text or tools AI Assistant

A Partial Differential Equation for Bayesian Inference

After some work (in the paper, please read) we can write down a general equation describing the drift-diffusion of the posterior predictive distribution which is again Fokker-Plank type.

$$\frac{\partial p}{\partial \tau} = -m^i \frac{\partial p}{\partial y^i} + \mathcal{I}^{ij}(\gamma) \frac{\partial^2 p}{\partial y^i \partial y^j} \tag{21}$$

- From this equation we can link to an ERG flow in the sense of a drift-diffusion process.

34°F Light rain/snow Search ENG US 7:49 pm 10/4/2025

Menu mainv2 - Copy.pdf Create Sign in
All tools Edit Convert E-Sign Find text or tools AI Assistant

The ERG/Dynamical Bayesian Inference Correspondence

	ERG	Bayesian Diffusion
1.	Scale parameter: $ln\Lambda$	Inverse Observation Time: $\tau = \frac{1}{T}$
2.	The regularised 2pt fn. C	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.

34°F Light rain/snow Search ENG US 7:49 pm 10/4/2025

Menu mainv2 - Copy.pdf Create Sign in
All tools Edit Convert E-Sign Find text or tools AI Assistant

Renormalizability and Scale

Recall, in the field theory context the role of the inverse metric was played by the regulated two point function C_Λ 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.

20
36
33°F Snow Search ENG US 7:51 pm 10/4/2025

Pruning using the Fisher Information

- We want to use the Fisher Information metric to prune a network
- Consider an generative encoder/decoder trained on MNIST used generate MNIST images
- Calculate the Information metric for the network parameters (ie the weights)
- Look for a clear hierarchy in information space
- Prune weights with little information
- Compare the outputs

33°F Snow Search ENG US 7:53 pm 10/4/2025

Generated data

Source images. Input to autoencoder

Cutoff $\Lambda = 0$. 15680 parameters. Unpruned autoencoder

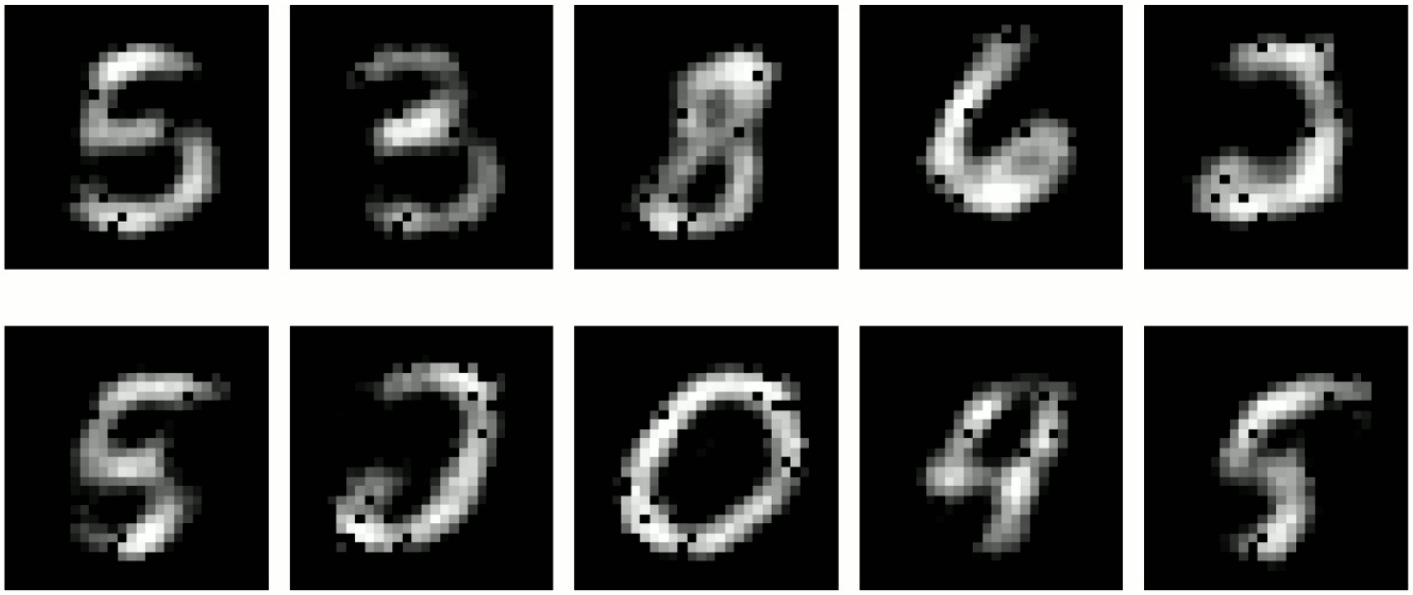

Figure: Cutoff $\Lambda = 0$. 15680 parameters. Unpruned autoencoder.

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

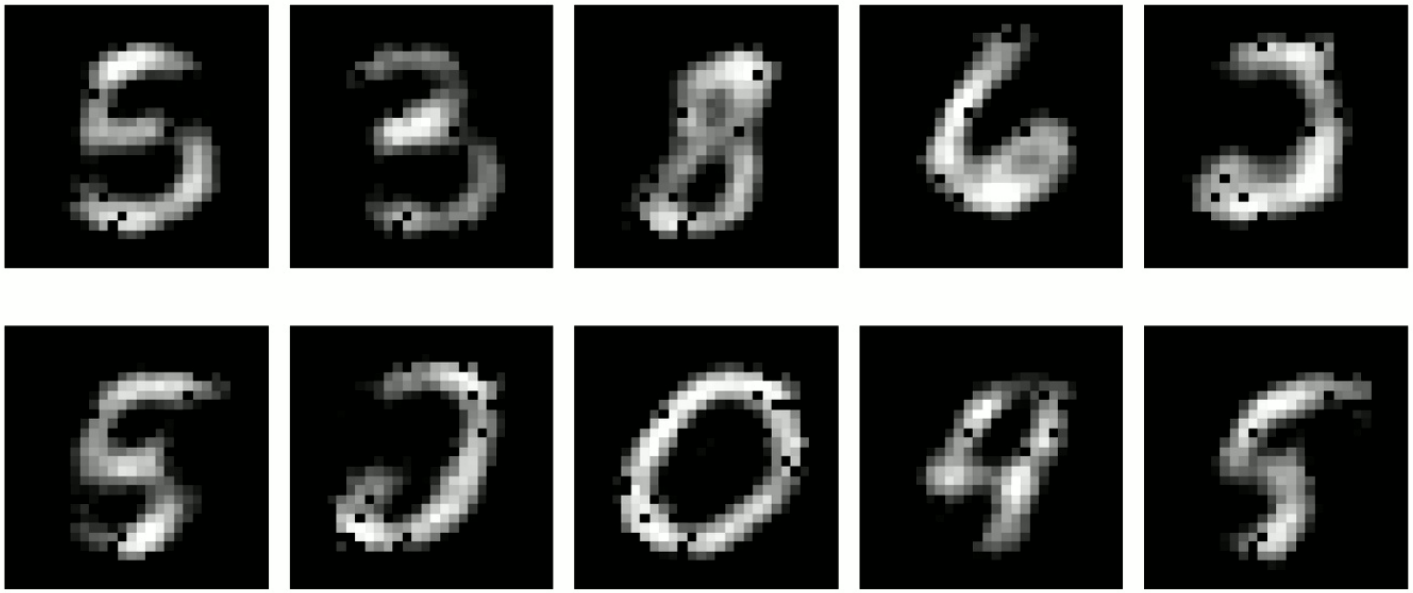

Figure: Cutoff $\Lambda = 0.125$. 10168 parameters removed (64.8%).

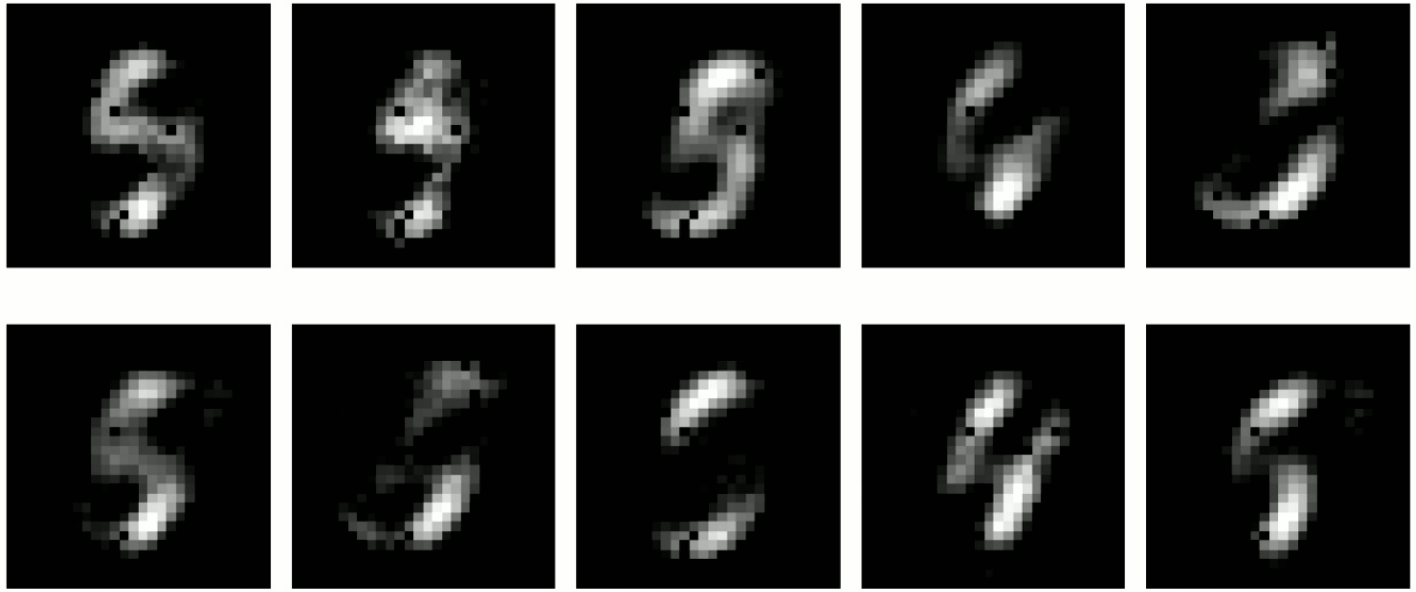

Figure: Cutoff $\Lambda = 1.25$. 13879 parameters removed (82.1%).

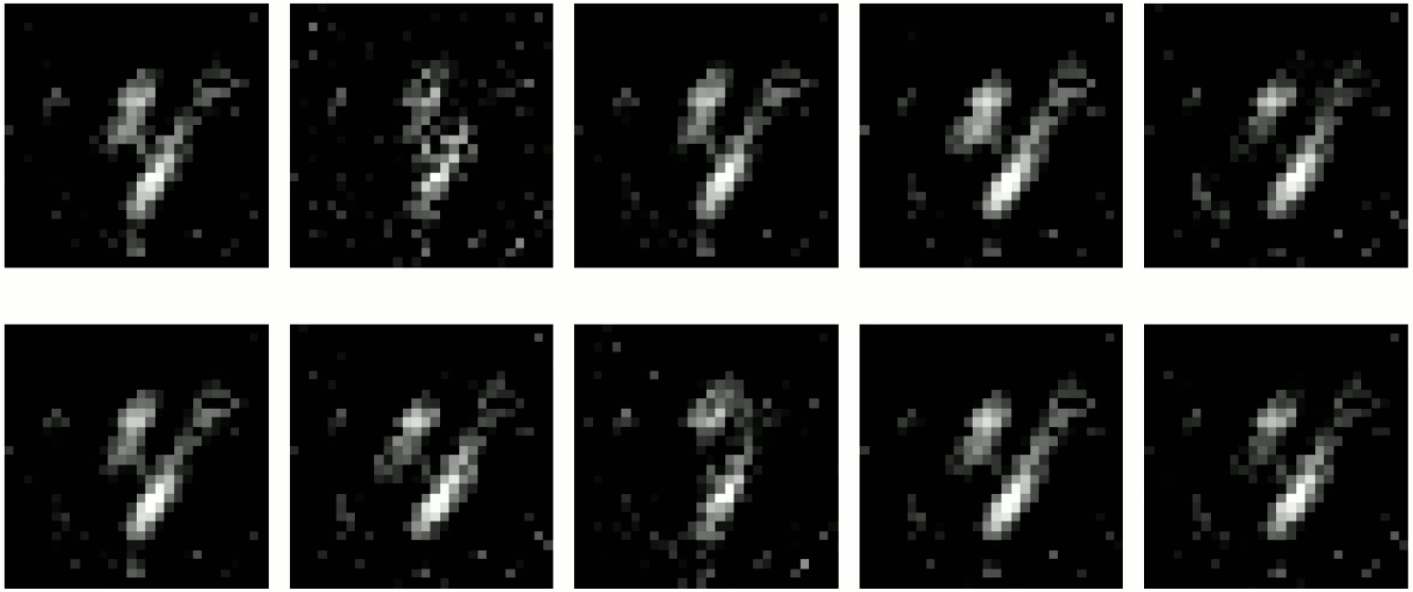

Figure: Cutoff $\Lambda = 1.75$. 15386 parameters removed (98.1%).

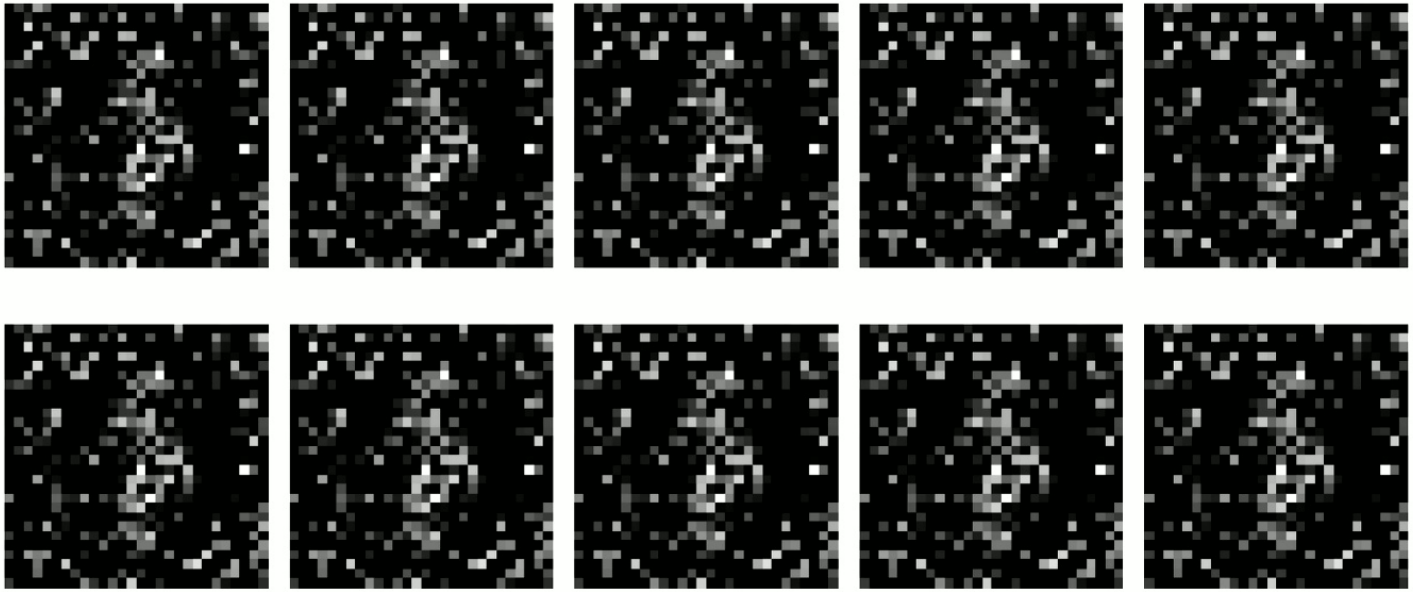

Figure: Cutoff $\Lambda = 2.375$. All (15680) parameters removed (100%).

Figure: Source images. Input to autoencoder.

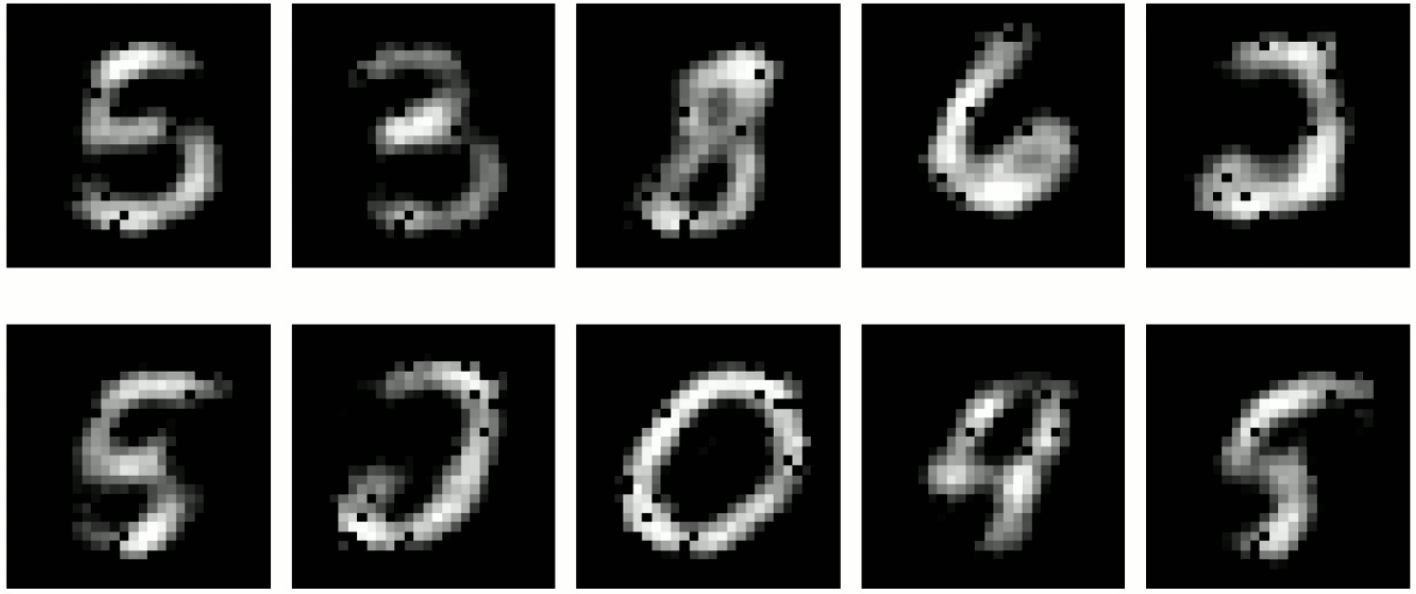

Figure: 15680 parameters. Unpruned autoencoder. Equivalent to cutoff $\Lambda = 0$.

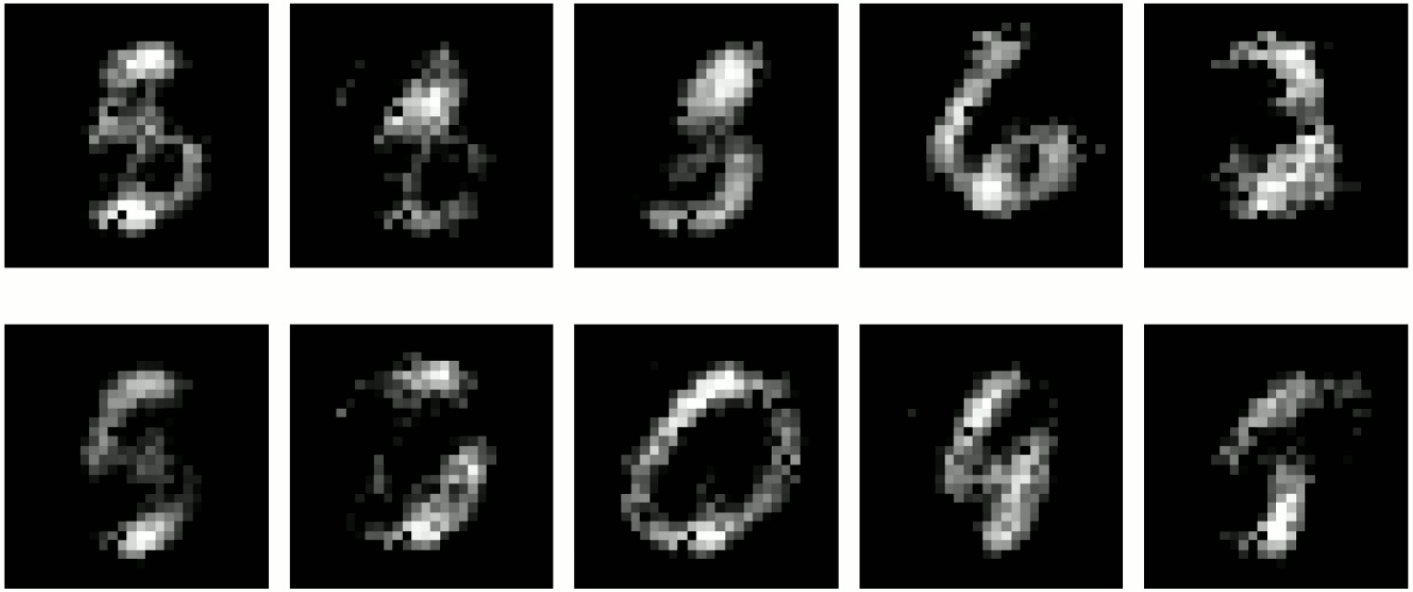

Figure: 10168 parameters removed (64.8%). Equivalent to cutoff $\Lambda = 0.125$.

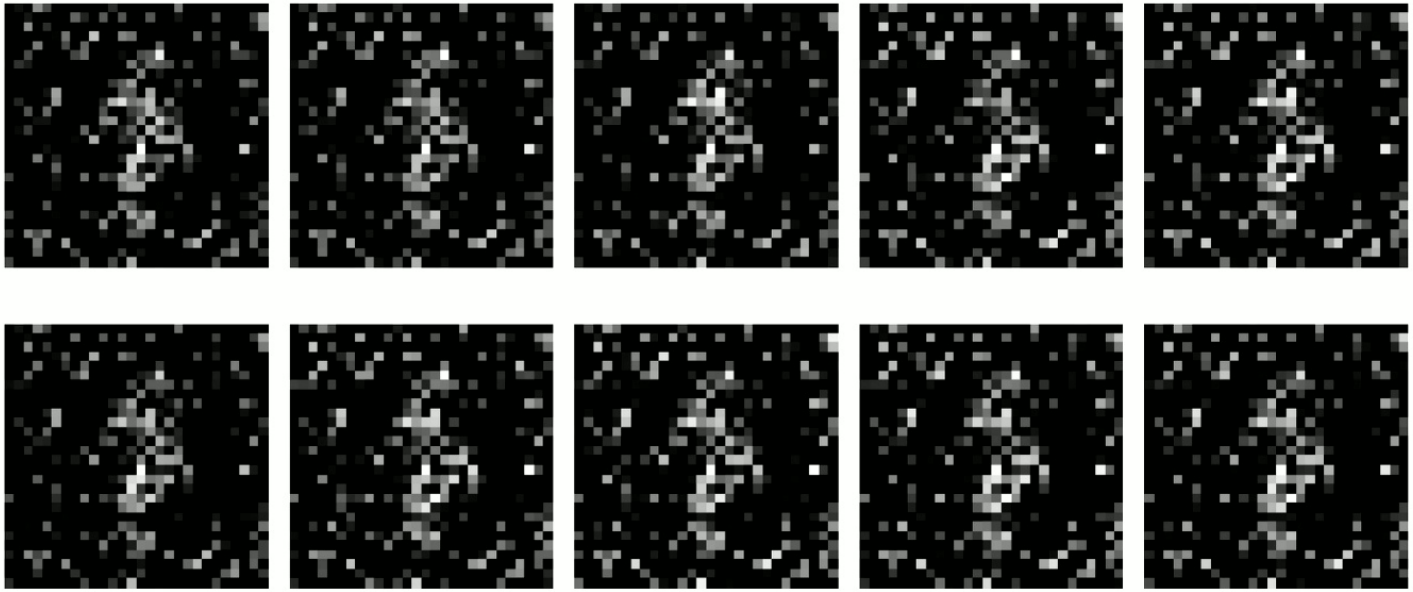

Figure: 15386 parameters removed (98.1%). Equivalent to cutoff $\Lambda = 1.75$.

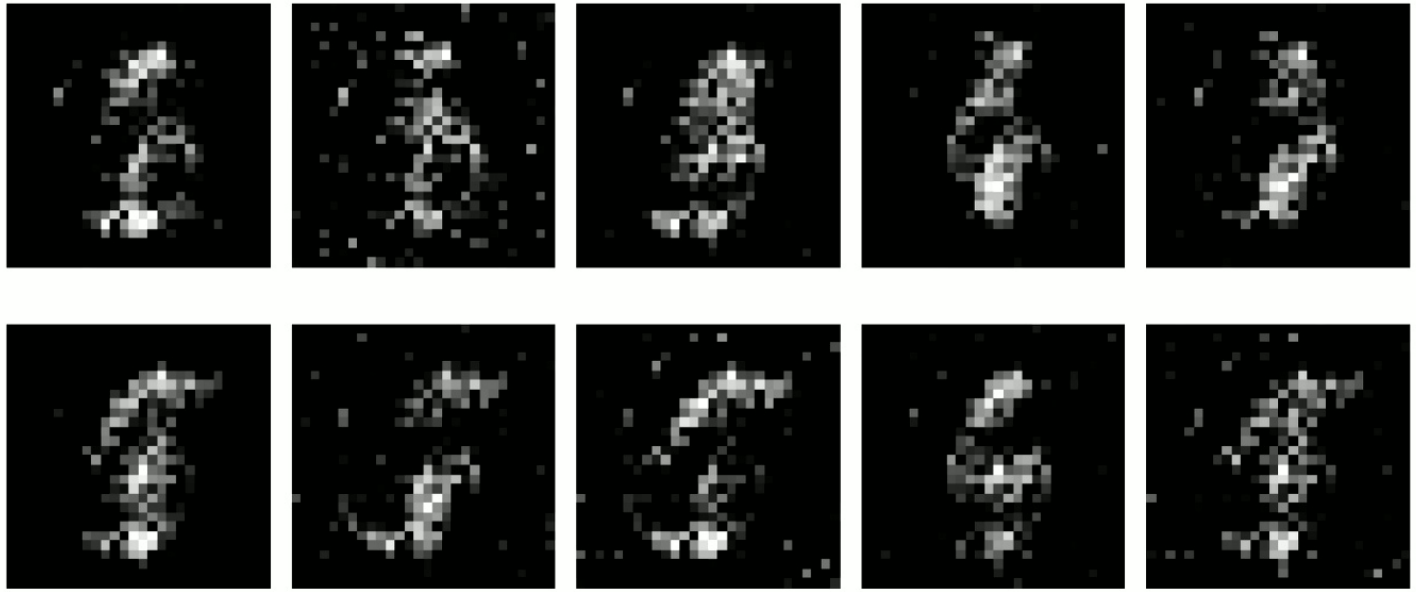

Figure: 13879 parameters removed (82.1%). Equivalent to cutoff $\Lambda = 1.25$.

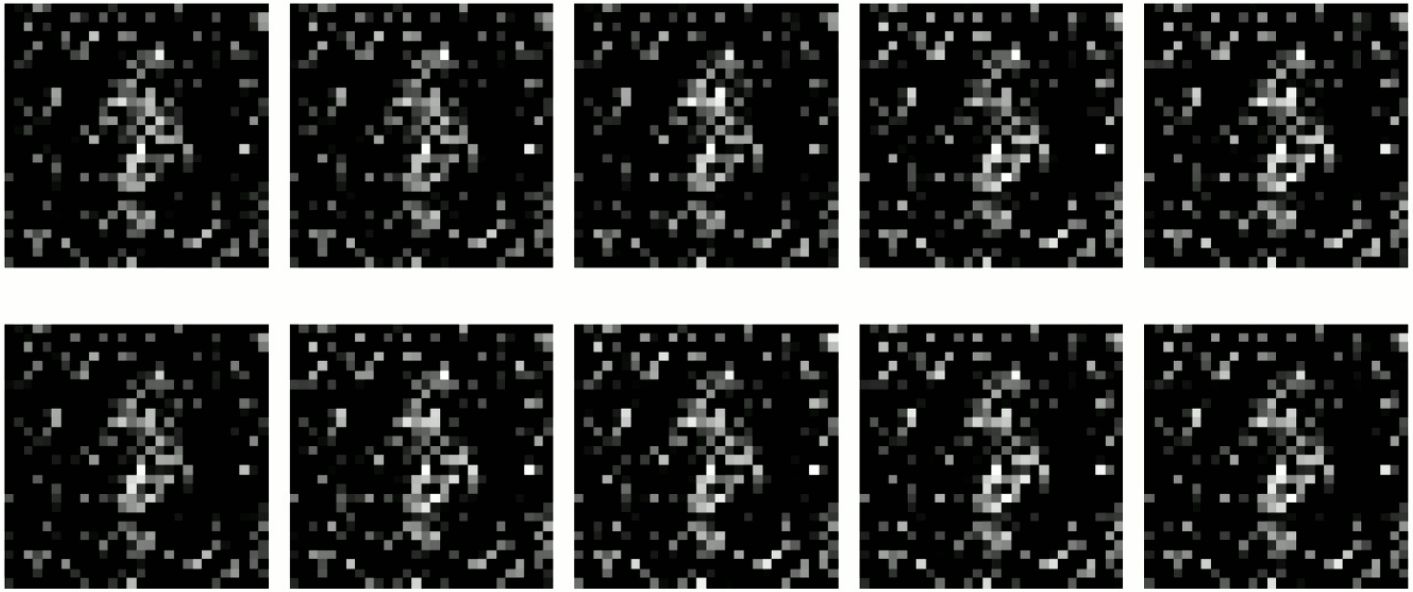

Figure: 15386 parameters removed (98.1%). Equivalent to cutoff $\Lambda = 1.75$.

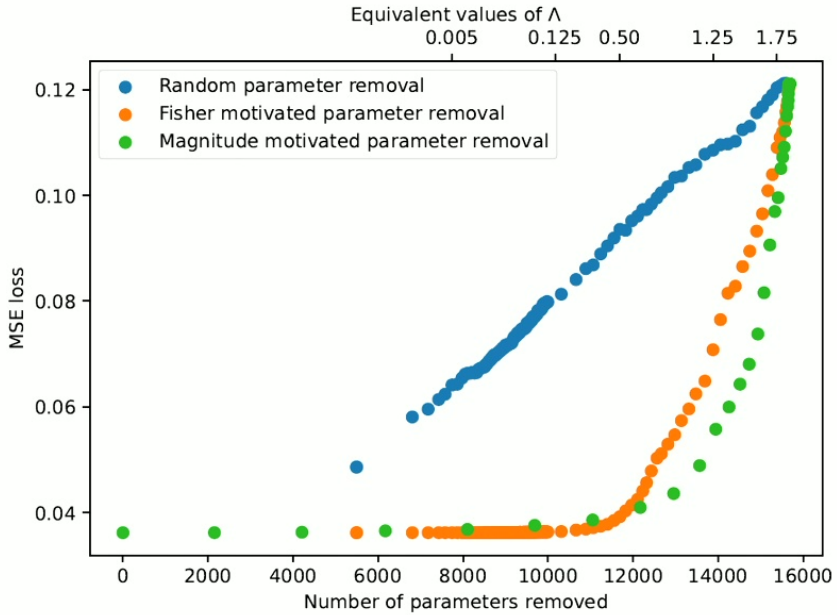

Figure: Mean square error loss between the true distribution and output of the pruned networks.

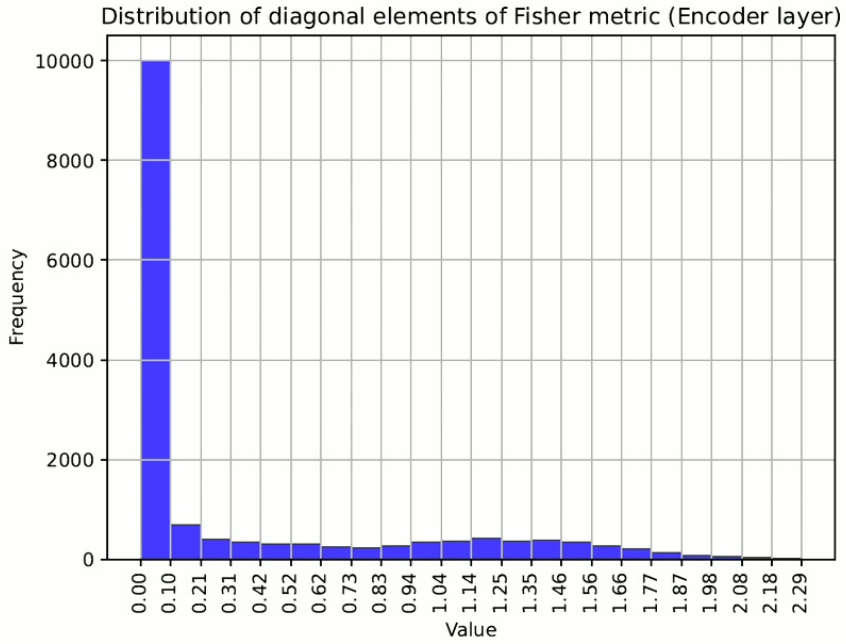

Figure: Histogram showing the distribution of diagonal Fisher matrix elements.

Menu mainv2 - Copy.pdf Create Sign in
All tools Edit Convert E-Sign Find text or tools AI Assistant

Discussion

- The proposed correspondence between ERG and Dynamical Bayesian Inference provides a quantification of the information content in an effective theory at every scale.
- The notion of Bayesian inference as an inverse process to renormalization inspires a host of questions around reinterpreting RG, for example 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. We are looking at various applications...

33°F Snow Search ENG US 8:00 pm 10/4/2025