

Article

Stochastic Thermodynamics of Learning Parametric Probabilistic Models

Shervin S. Parsi ^{1,2}

¹ Physics Program, The Graduate Center, City University of New York, New York, NY 10016, USA; shsparsi@gmail.com

² Initiative for the Theoretical Sciences, The Graduate Center, City University of New York, New York, NY 10016, USA

Abstract: We have formulated a family of machine learning problems as the time evolution of parametric probabilistic models (PPMs), inherently rendering a thermodynamic process. Our primary motivation is to leverage the rich toolbox of thermodynamics of information to assess the information-theoretic content of learning a probabilistic model. We first introduce two information-theoretic metrics, memorized information (M-info) and learned information (L-info), which trace the flow of information during the learning process of PPMs. Then, we demonstrate that the accumulation of L-info during the learning process is associated with entropy production, and the parameters serve as a heat reservoir in this process, capturing learned information in the form of M-info.

Keywords: parametric generative models; machine learning; thermodynamics of information; entropy production; information theory

1. Introduction

Starting nearly half a century ago, physicists began to learn that information is a physical entity [1–3]. Today, the information-theoretic perspective has significantly impacted various fields of physics, including quantum computing [4], cosmology [5], and thermodynamics [6]. Simultaneously, recent years have witnessed the remarkable success of an algorithmic approach known as machine learning, which is adept at learning information from data. This paper is propelled by a straightforward proposition: if “information is physical”, then the process of learning information must inherently be a physical process.

The concepts of memory, prediction, and information exchange between subsystems have undergone extensive exploration within the realms of thermodynamics of information [6] and stochastic thermodynamics [7]. For instance, Still et al. [8] delved into the thermodynamics of prediction, and the role of information exchange between thermodynamic subsystems has been studied by Sagawa and Ueda [9] and Esposito et al. [10]. This rich toolbox of thermodynamics of information is our main venue to study the physics of the machine learning process, with motivation to assess the information content of the learning process.

The type of machine learning problem we consider in this study encompasses any algorithmic approach that *evolves* a parametric probabilistic model (PPM) or simply the model toward a desirable target distribution through gradient-based loss function minimization. To establish our notation, consider a set of observations denoted by the training dataset B , drawn from an unknown target distribution p^* . The PPM, without loss of generality, can be written as the follows:

$$p_{\theta}(X = x) = e^{-\phi_{\theta}(x)} \quad (1)$$

This distribution is parameterized by a set of parameters $\theta \in \mathbb{R}^M$. The objective of learning is to find a set of parameters such that the samples drawn from the model exhibit a desirable



Citation: Parsi, S.S. Stochastic Thermodynamics of Learning Parametric Probabilistic Models. *Entropy* **2024**, *26*, 112. <https://doi.org/10.3390/e26020112>

Academic Editor: Antonio Prados

Received: 23 November 2023

Revised: 3 January 2024

Accepted: 17 January 2024

Published: 25 January 2024



Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

statistical characteristics. In machine learning practice, one constructs the function $\phi_\theta(x)$ with a (deep) neural network and leaves the parameter selection task to an optimizer that minimizes a loss function. Examples include energy-based models [11], Large Language Models (LLMs), Softmax classifiers, and Variational Autoencoders (VAEs) [12].

While the information-theoretic approach to this problem is prevalent in the field [13–16], it has also faced criticism [17]. Our primary motivation for framing learning in a PPM as a thermodynamic process is to facilitate the assessment of the information content inherent in the learning process. The structure of this paper is as follows. Section 2 briefly discusses prior information-theoretic approaches to the learning problem with PPM and the challenges they encounter. Subsequently, we introduce our own information-theoretic metrics. Finally, Sections 3 and 6 employ the thermodynamic framework to address these information-theoretic inquiries.

2. Information Content of PPMs

Locating information within the parametric model (i.e., the neural network) remains a fundamental question in machine learning [18]. This challenge is central to any information-theoretic perspective on machine learning problems. In a pioneering study, Shwartz-Ziv et al. [19] quantified the internal information within neural networks by estimating the mutual information between inputs and the activities of hidden neurons. Moreover, they employed the information bottleneck theory to interpret the decrease in this mutual information as evidence of data compression during learning. This perspective garnered significant attention in the field [15,16], reinforcing the view of neural networks as an information channel. However, the study encountered critiques [17,20]. A primary problem was that the hidden neurons' activity in a neural network constitutes a deterministic function of the input. Such determinism inherently possesses a trivial mutual information value, even prior to any learning. The challenge of defining a well-defined and interpretable (Shannon) information metric in deterministic neural networks has prompted the proposition that neural network information processing is geometric in nature [17] (given that inputs are mapped to a latent space of differing dimensions) rather than information-theoretical.

In a distinct research direction, the authors of [21] addressed the significance of assessing the information content of the model's parameters. In our study, we echo this view, emphasizing that the parameters are the primary carriers of learned information within neural networks. Consequently, any information-theoretic measure of learned information by the model should be grounded in the parameters rather than the deterministic activity of hidden neurons. However, quantifying the information within parameters poses challenges, primarily due to the elusive nature of their distribution [22]. In this section, we introduce two information-theoretic metrics crafted to assess the information content within the learning process of a PPM. This paves the way for computing these quantities within the thermodynamic framework.

To avoid introducing new notation, we also denote B as the ground truth random variable associated with the target distribution p^* from which the training dataset is sampled. Subsequently, we represent the action of the optimizer as a map between this ground truth random variable and the desired set of parameters after n optimization steps:

$$\Theta_{t_n} = \Lambda_n(B) \quad (2)$$

The map Λ_n incorporates the structure of the loss function, the optimization algorithm, and any hyperparameters related to the optimizer's action. We exclude the initial parameters' value from this map's argument under the assumption that as n increases, the final set of parameters becomes independent of its initial condition. In information theory terminology, this map corresponds to a *statistic* of the ground truth random variable [23]. Moreover, the outcome of this map defines a model from which the final model-generated sample is sampled: $X_{t_n} \sim p(x|\theta_{t_n})$. Considering that the model-generated sample becomes independent

of the ground truth random variable given the parameters, we can express the following Markov chain governing the learning process:

$$B \rightarrow \Theta_{t_n} \rightarrow X_{t_n}. \quad (3)$$

The data processing inequality (DPI) associated with this Markov chain serves as our framework to define two information-theoretic metrics that gauge the information content of the model:

$$\underbrace{I_{\Theta;B}(t_n)}_{\text{M-info}} \geq \underbrace{I_{X;B}(t_n)}_{\text{L-info}}. \quad (4)$$

We used the notations presented in Table 1. The left-hand side of this inequality quantifies the accumulation of mutual information between the parameters and the training dataset, while the right-hand side characterizes the performance of the generative model, which gauges the accumulation of mutual information between the model's generated samples and the training dataset. We refer to the former as memorized information (M-info) and the latter as learned information (L-info). We also note that both of these quantities start at zero before the training begins. Thus, their measurements at t_n reveal accumulation of information during the learning process.

Table 1. A list of notations used in this paper.

$\Delta_{t_n} f(t) := f(t_n) - f(0)$	Change over the interval $[0, t_n]$
$\langle f(x) \rangle_{p(x)} := \int dx p(x) f(x)$	Average over $p(x)$
$s_X(t) := s[p_t(x)] := -\ln p_t(x)$	Surprisal of $p_t(x)$
$S_X(t) := S[p_t(x)] := \langle -\ln p_t(x) \rangle_{p_t(x)}$	Shannon entropy of $p_t(x)$
$I_{X;\Theta}(t) := I[X_t; \Theta_t] := S_X(t) - S_{X \Theta}(t)$	Mutual information between X and Θ at time t

In the context of the learning problem, the DPI, as referenced in Equation (4), suggests that what is *memorized* is always greater than or equal to what is *learned*. The L-info metric is task-oriented. For example, in the realm of image generation, it quantifies the statistical resemblance between the model's outputs and the genuine images. In the case of classification tasks, L-info would encapsulate only the pertinent information for label prediction. In contrast, M-info can encompass information not directly pertinent to the current task. For instance, it might capture intricate pixel configurations in an image dataset, which are not crucial for identifying distinct patterns like human faces. The DPI in Equation (4) neatly illustrates the risk of overfitting, which is when a model starts to incorporate extraneous information that does not align with the learning objective. The necessity of constraining the information in a model's parameters is highlighted in [21], echoing the minimum description length principle [24]. Additionally, studies suggest that the SGD optimizer tends to favor models with minimal information in their parameters [22]. The recent work by the authors of [25] even proposed an upper limit for minimizing parameter information to bolster generalization capabilities. These findings suggest that the learning process seeks to minimize the left-hand side of the DPI inequality while simultaneously maximizing the right-hand side, which measures the model performance. This leads us to an ideal scenario where $I_{\Theta;B}(t_n) = I_{X;B}(t_n)$, signifying that all memorized information is relevant to the learning task.

We now take one step further in our definition of M-info and L-info. First, the presence of the optimizer map, as referenced in Equation (2) and connecting the ground truth source of the training dataset to the parameters, allows us to simplify M-info as follows:

$$\text{M-info} := I_{\Theta;B}(t_n) = S(\Theta_{t_n}) \quad (5)$$

Thus, the parameters naturally emerge as the model's *memory*, where its Shannon entropy measures the stored information during the learning process.

Second, we swap B for Θ_{t_n} in the definition of L-info at the cost of losing some information:

$$\begin{aligned} \text{L-info} &:= I_{X;B}(t_n) = I_{X;\Lambda_{t_n}(B)} + \epsilon \\ &= I_{X;\Theta}(t_n) + \epsilon \end{aligned} \quad (6)$$

where ϵ is a non-negative number that equals zero only when Λ_n is a *sufficient statistic* of B . For the above expression, the condition of a sufficient statistic can be eased when Θ_{t_n} becomes sufficient only with respect to X . In other words, when Λ_n preserves all information in B that is also mutual in X_{t_n} , i.e., preserving all task-oriented information. Indeed, we are interested in this type of preservative map in machine learning practice. Therefore, for a well-chosen optimizer, we consider $I_{X;\Theta}$ to be a reasonable proxy to L-info, and we use the two interchangeably:

$$\text{L-info} := I_{X;\Theta}(t_n) \quad (7)$$

3. The Learning Trajectory of a PPM

The time evolution of the PPM is the first clue to frame the learning process as a thermodynamic process. To illustrate this, consider a discretized time interval $[0, t_n]$, which represents the time needed for n optimization steps of the parameters. During this time, the optimizer algorithm draws a sequence of independent and identically distributed samples from the training dataset. We denote this sequence by $\mathbf{b}_n := \{b_{t_1}, b_{t_2}, \dots, b_{t_n}\}$ and refer to it as the “input trajectory”. Then, the outcome of the optimization defines a sequence of parameters called the “parameters’ trajectory”: $\theta_n := \{\theta_0, \theta_{t_1}, \theta_{t_2}, \dots, \theta_{t_n}\}$. Each realization of parameters defines a specific PPM. Consequently, the parameters’ trajectory produces a sequence of PPMs:

$$\mathcal{T} := \{p(X|\theta_0), p(X|\theta_{t_1}), p(X|\theta_{t_2}), \dots, p(X|\theta_{t_n})\} \quad (8)$$

We refer to this sequence as the *learning trajectory*, depicted in Figure 1. On the other hand, a thermodynamic process can be constructed solely from the time evolution of a distribution [26]. Therefore, we see $\mathcal{T}$ as a thermodynamic process. The physics of this process is encoded in the transition rates governing the master equation of this time evolution. Finding the transition rate associated to learning a PPM is our main task in this section.

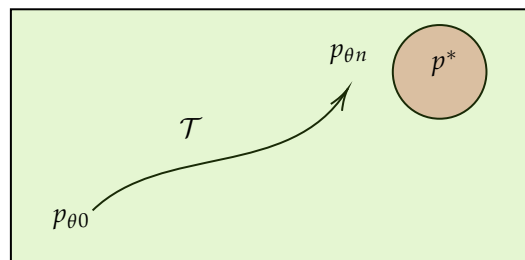


Figure 1. The learning trajectory $\mathcal{T}$ depicts the thermodynamic process that takes the initial model state to the final state. The green area shows the space of family of distribution accessible to the PPM. The red area considers the possibility that the target distribution p^* is not in this family.

3.1. The Model Subsystem

We refer to the subsystem that goes under the thermodynamic process $\mathcal{T}$ as the *model subsystem*. This subsystem has X degrees of freedom, and its microscopic states’ realization along the learning trajectory represent model-generated samples at each time step. We denote the stochastic trajectory of the model subsystem (model-generated samples) by $\mathbf{x}_n := \{x_{t_1}, x_{t_2}, \dots, x_{t_n}\}$. To avoid confusion with our notation, we consider the probability

functions $p(x_{t_i}|\theta_{t_i})$ and $p(x_{t_{i-1}}|\theta_{t_i})$, which represent the probability of observing $x_{t_i} \in \mathbf{x}_n$ and $x_{t_{i-1}} \in \mathbf{x}_n$ at time $t = t_i$, respectively. Here, the time index of θ aligns with the time index of the PPM (i.e., $p_{t_i}(X|\theta_{t_i}) \equiv p(X|\theta_{t_i})$) because the PPM is fully defined upon observing the parameters. In contrast, the time index on x denotes a specific observation within $\mathbf{x}_n$. To simplify our notation, the absence of a time index on x denotes a generic realization of the random variable X , and we write $p(x|\theta_{t_i})$ instead of $p(X|\theta_{t_i})$.

3.2. The Parameter Subsystem

The parameters of the neural network at each step of optimization represent realization of the parameter subsystem with Θ degrees of freedom and the stochastic trajectory θ_n . The statistical state of the parameter subsystem is given with the marginal $p(\theta_{t_i})$ at time step $t = t_i$. This marginal state represents the statistic of all possible outcomes of training a PPM on a specific learning objective. We can think of training an ensemble of computers on the same machine learning task. This allows us to think about the time evolution of the marginal $p(\theta_{t_i})$ and the joint distribution $p(x|\theta_{t_i})p(\theta_{t_i})$ during the learning process. We refer to this view as the *ensemble view* of the learning process. In practice, however, we train the PPM only once, and we do not have access to the marginal $p(\theta_{t_i})$. Thus, our model-generated samples are conditioned on specific observations of parameters $\theta_{t_i} \sim \Theta_{t_i}$. This defines the *conditional view* of the learning process, which is fully described by the learning trajectory of the PPM.

In machine learning practice, it is desirable for a training process to exhibit a robust outcome, regardless of who is running the code. One way to achieve this is by imposing a *low variance condition* on the statistics of the parameters across the ensemble of all learning trials. This condition asserts that the parameters' trajectory across the ensemble is confined to a small region $D_n \subset \mathcal{R}^M$. As n grows larger, this region shrinks and becomes associated with the area surrounding the target distribution as depicted in Figure 1. Under this condition, we can express

$$\langle f(\theta) \rangle_{p(\theta_n)} \approx f(\theta^*), \quad \forall \theta^* \in D_n \quad (9)$$

The above approximation becomes exact when $p(\theta_n)$ assumes the form of a Dirac delta function, indicating a zero-variance condition in the parameters' dynamics.

The low-variance condition proves invaluable when computing the information-theoretic measurements introduced in Section 2. This is because the computation of the M-info $I_{B;\Theta}$ and L-info $I_{X;\Theta}$ necessitates averaging over the parameters' distribution. However, since we typically train our model just once, we lack direct access to the parameters' distribution throughout the learning trajectory. To overcome this challenge, we introduce the partially averaged L-info:

$$I_{X;\Theta}(\theta_t) = \int dx p(x|\theta_t) \ln \frac{p(x|\theta_t)}{p_t(x)} \quad (10)$$

Subsequently, under the low-variance condition of the parameter subsystem, we can measure the partially averaged L-info as a proxy for the L-info, where $I_{X;\Theta}(t) = \langle I_{X;\Theta}(\theta) \rangle_{p_t(\theta)} \approx I_{X;\Theta}(\theta_t)$. In Section 6, we will delve deeper into the evidence supporting the low-variance condition of the subsystem Θ .

We refer to the joint (X, Θ) as the *learning system*, which embodies the thermodynamic process of learning a PPM. In this section, we will demonstrate that the thermodynamic exchange between the model subsystem and parameter subsystem is the primary source for producing M-info and L-info during the learning process. Before delving deeper, we establish two interconnected assumptions about the parameter subsystem: (1) The PPM is over-parameterized; specifically, the subsystem Θ has a much higher dimension compared with the subsystem X . (2) The parameter subsystem evolves in a quasi-static fashion (slow dynamics).

The foundation for these assumptions in machine learning is clear. Training over-parameterized models represents a pivotal achievement of machine learning algorithms, and the slow dynamics (often termed as lazy dynamics) of these over-parametrized models are well documented [27,28]. These characteristics underscore the significant role of the parameter subsystem in the learning process, akin to that of a heat reservoir. Over-parameterization implies a higher heat capacity for this subsystem compared with the model subsystem. Additionally, the quasi-dynamics align with the behavior of an ideal heat reservoir, which does not contribute to entropy production [29]. The role of the parameter subsystem as a reservoir aligns with the assumption of a low-variance condition for this subsystem. This is because we expect the stochastic dynamics of the reservoir to be low variance across the ensemble of all trials.

In this study, we attribute the role of an ideal heat reservoir to the parameter subsystem with an inverse temperature $\beta^{-1} = 1$. In Section 6, we delve deeper into the rationale behind this assumption by examining the stochastic dynamics of parameters under a vanilla stochastic gradient descent optimizer and highlighting potential limitations of this assumption.

3.3. Lagged Bipartite Dynamics

We want to emphasize that the dynamics of subsystem X are not a mere conjecture or an arbitrary component in this study. Rather, it is an integral part of training a generative PPM. These dynamics are inherent in the optimizer action, necessitating a fresh set of model-generated samples to compute the loss function or its gradients after each parameter update. For instance, in the context of EBM, a Langevin Monte Carlo (LMC) sampler can be employed to generate new samples from the model [30]. The computational cost of producing a fresh set of model-generated samples introduces a time delay in the parameter dynamics. For instance, when using an LMC sampler, the number of Monte Carlo steps dictates this lag time. Conversely, in the case of a language model, since the computation of the loss function relies on inferring subsequent tokens, the inference latency signifies the time delay.

We denote the lag time parameters as τ . Here, the model subsystem evolves on the timescale δt , while the parameter subsystem evolves on the timescale $\alpha = \tau \delta t$. In the thermodynamic context, this parameter represents the *relaxation time* of the subsystem X under the fixed microscopic state of the subsystem Θ . Conceptually, parameter τ acts as a complexity metric, quantifying the computational resources required for each parameter optimization step. Moreover, the dynamics of the joint (X, Θ) exhibit a bipartite property. This implies that simultaneous transitions in the states of X and Θ are not allowed, given that the observation of a new set of model-generated samples occurs only after a parameter update.

The lagged bipartite dynamics described above can be represented using two time resolutions: δt and α . In the finer time resolution of δt , the Markov chain within the time interval $[t_i, t_{i+1}]$ is as follows:

$$(x_{t_i}, \theta_{t_i}) \rightarrow (x_{t_i}, \theta_{t_{i+1}}) \rightarrow (x_{t_i+\delta t}, \theta_{t_{i+1}}) \dots \rightarrow (x_{t_i+\tau\delta t}, \theta_{t_{i+1}}) \equiv (x_{t_{i+1}}, \theta_{t_{i+1}}). \quad (11)$$

We can also analyze this dynamics at a coarser time resolution of α . Within the interval $[t_0, t_n]$, the Markov Chain appears as

$$(x_0, \theta_0) \dashrightarrow (x_{t_1}, \theta_{t_1}) \dots (x_{t_{n-1}}, \theta_{t_{n-1}}) \dashrightarrow (x_{t_n}, \theta_{t_n}). \quad (12)$$

In the above Markov chain, the dashed arrows remind us of the *ignorance* of the intermediate steps in the high resolution picture (Equation (11)). Figure 2 illustrates the lagged bipartite dynamics of the learning system. An important observation is that the learning trajectory $\mathcal{T}$, as defined in Equation (8), is written in the low-resolution picture. Therefore, studying the learning trajectory means studying the dynamics of the system (X, Θ) in the low-resolution picture.

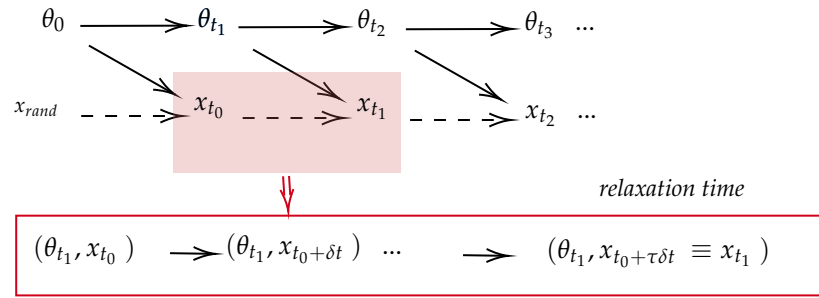


Figure 2. This figure shows the Bayesian network for the joint trajectory probability $P[x_n, \theta_n]$ based on the lagged bipartite dynamics.

3.4. Trajectory Probabilities

To set the stage for the application of the fluctuation theorem (FT) to learning a PPM, we define the trajectory probability of the joint (x_n, θ_n) as the probability of observing the trajectory of model-generated samples and parameters during the learning process:

$$P[x_n, \theta_n] := p(x_0, x_{t_1}, \dots, x_{t_n}, \theta_0, \theta_{t_1}, \dots, \theta_{t_n}) \quad (13)$$

Additionally, we can consider the time reversal of the model's trajectory and parameters' trajectory as $\tilde{x}_n := \{x_{t_n}, x_{t_{n-1}}, \dots, x_{t_1}\}$ and $\tilde{\theta}_n := \{\theta_{t_n}, \theta_{t_{n-1}}, \dots, \theta_{t_1}\}$, respectively. Then, the probability of observing the backward trajectory is denoted by $P[\tilde{x}_n, \tilde{\theta}_n]$.

Here, $P[x_n, \theta_n]$ represents the trajectory probability of the learning system in the ensemble view. In practice, however, we typically train our model only once, and we often lack access to the parameters' distribution. Therefore, our model is conditioned on the observation of a specific parameters' trajectory θ_n . This defines the trajectory probability in the conditional view:

$$P[x_n | \theta_n] := \frac{P[x_n, \theta_n]}{P[\theta_n]} \quad (14)$$

where

$$P[\theta_n] = p(\theta_0, \theta_{t_1}, \dots, \theta_{t_n}). \quad (15)$$

Similarly, the backward conditional trajectory probability is the probability of observing the time reversal model's trajectory, conditioned on the observation of the time reversal parameters' trajectory $P[\tilde{x}_n | \tilde{\theta}_n] = \frac{P[\tilde{x}_n, \tilde{\theta}_n]}{P[\tilde{\theta}_n]}$.

We now use the Markov property of the Markov chain in Equation (12) to decompose the conditional trajectory probability and the marginal trajectory probability, respectively, as follows:

$$\begin{aligned} P[x_n | \theta_n] &= p(x_{t_n} | x_{t_{n-1}}, \theta_{t_n}) \dots p(x_{t_1} | x_0, \theta_{t_1}) p(x_0 | \theta_0), \\ P[\theta_n] &= p(\theta_{t_n} | \theta_{t_{n-1}}) \dots p(\theta_{t_1} | \theta_{t_0}) p(\theta_0), \end{aligned} \quad (16)$$

where expressions such as $p(x_{t_n} | x_{t_{n-1}}, \theta_{t_n})$ and $p(\theta_{t_n} | \theta_{t_{n-1}})$ represent the transition probabilities that determine the probability of moving from one microscopic state to another. Additionally, we define two probability trajectories conditioned on the initial conditions, which will be used later in the formulation of the FT:

$$\begin{aligned} P[(x_n | \theta_n) | (x_0 | \theta_0)] &:= P[(x_n | \theta_n)] / p(x_0 | \theta_0), \\ P[\theta_n | \theta_0] &:= P[\theta_n | \theta_0] / p(\theta_0). \end{aligned} \quad (17)$$

3.5. Local Detailed Balance (LDB) for Learning PPMs

The transition probabilities, represented in Equation (16), capture the physics of the learning problem. When considering a Markov property (i.e., memoryless process) for

time evolution of the model subsystem, the transition rate for this subsystem is reduced to the PPM:

$$p(x_{t_i}|x_{t_{i-1}}, \theta_{t_i}) = p(x_{t_i}|\theta_{t_i}). \quad (18)$$

The above expression suggests that the transition rate between two microscopic states $x_{t_{i-1}}$ and x_{t_i} under the fixed θ_{t_i} is equivalent to the probability of observing x_{t_i} by the PPM itself at $t = t_i$. To reiterate, this is the Markov property that suggests the elements inside x_n are independently and freshly drawn from the PPM specified with given parameters along the learning trajectory $\mathcal{T}$. This is especially true when $\tau \gg 1$. We can generalize this observation for the backward transition probability $p(x_{t_{i-1}}|x_{t_i}, \theta_{t_i})$, which represents the probability of the backward transition $(x_{t_i}, \theta_{t_i}) \dashrightarrow (x_{t_{i-1}}, \theta_{t_i})$ under a fixed θ_{t_i} , as follows:

$$p(x_{t_{i-1}}|x_{t_i}, \theta_{t_i}) = p(x_{t_{i-1}}|\theta_{t_i}). \quad (19)$$

The above expression tells us that the probability of backward transition is equivalent to the probability of observing the sample generated at $t = t_{i-1}$ in x_n with the PPM at time $t = t_i$.

Finally, we write the log ratio of the forward and backward transitions:

$$\ln \frac{p(x_{t_i}|x_{t_{i-1}}, \theta_{t_i})}{p(x_{t_{i-1}}|x_{t_i}, \theta_{t_i})} = \ln \frac{p(x_{t_i}|\theta_{t_i})}{p(x_{t_{i-1}}|\theta_{t_i})} = -(\phi_{\theta_{t_i}}(x_{t_i}) - \phi_{\theta_{t_i}}(x_{t_{i-1}})), \quad (20)$$

where the second equality is due to Equation (1). The above expression resembles the celebrated local detailed balance (LDB) [31], which relates the log ratio of the forward and backward transition probabilities to the difference in potential energy of the initial and final states in the transition. The heat reservoir that supports the legitimacy of the above LBD expression for the learning PPM is the parameter subsystem, whose temperature has been set to one, as we will discuss in more detail in Section 6. We emphasize that the above LBD emerged naturally under assumption of the Markov property and a relaxation time for learning a generic generative PPM. It is also important to note that the above LBD is only valid in the low-resolution picture.

The LBD relation, presented in Equation (20), has a profound consequence. This allows us to write the forward conditional probability trajectory $P[x_n|\theta_n]$ and the backward conditional probability trajectory $P[\tilde{x}_n|\tilde{\theta}_n]$ solely based on the series of PPMs in the learning trajectory $\mathcal{T}$:

$$\begin{aligned} P[x_n|\theta_n] &= p(x_{t_n}|\theta_{t_n}) \dots p(x_{t_1}|\theta_{t_1}) p(x_0|\theta_0) \\ P[\tilde{x}_n|\tilde{\theta}_n] &= p(x_0|\theta_{t_1}) \dots p(x_{t_{n-1}}|\theta_{t_n}) p(x_{t_n}|\theta_{t_n}) \end{aligned} \quad (21)$$

This is significant because it renders the application of the FT framework to the learning PPMs practically, as we have access to elements of the learning trajectory during the learning process.

4. L-Info and Entropy Production

The version of the fluctuation theorem we are about to apply to the learning PPMs is known as the detailed fluctuation theorem (DFT) [32]. We also note that the machinery we are about to present for measuring the information flow in PPMs was developed to study the information exchange between thermodynamic subsystems [9]. The novelty here lies merely in the application of this machinery to the learning process of a PPM. In this section, we extensively use the notations presented in Table 1. Also, note that the temperature of

the parametric reservoir is set to one. Applying the DFT in the conditional view (i.e., the conditional forward and backward trajectories defined in Equation (21)) results in

$$\begin{aligned}\sigma_{x_n|\theta_n} &= \ln \frac{P[x_n|\theta_n]}{P[\tilde{x}_n|\tilde{\theta}_n]} \\ &= \ln \frac{P[(x_n|\theta_n)|(x_0|\theta_0)]}{P[(\tilde{x}_n|\tilde{\theta}_n)|(x_{t_n}|\theta_{t_n})]} + \ln \frac{p(x_0|\theta_0)}{p(x_{t_n}|\theta_{t_n})} \\ &= -q_{x_n}(\theta_n) + s[p(x_{t_n}|\theta_{t_n})] - s[p(x_{t_0}|\theta_{t_0})]\end{aligned}\quad (22)$$

The first line is due to the DFT, which defines the stochastic Entropy Production (EP) to be the logarithm of the ratio of the forward and backward trajectory probabilities. The second line is due to the decomposition presented in Equation (17). Finally, the third line is the consequence of the LDB relation in Equation (20), and the definition of the stochastic heat flow $q_{x_n}(\theta_n)$ is the change in the energy of the subsystem X due to alterations in its microscopic state configuration:

$$q_{x_n}(\theta_n) := -\ln \frac{P[(x_n|\theta_n)|(x_0|\theta_0)]}{P[(\tilde{x}_n|\tilde{\theta}_n)|(x_{t_n}|\theta_{t_n})]} = \sum_{i=1}^n \phi_{\theta_{t_i}}(x_i) - \phi_{\theta_{t_i}}(x_{i-1}). \quad (23)$$

Note that our convention defines $q_{x_n} > 0$ as the heat absorbed by the subsystem X.

The second law arises from averaging Equation (22) over the forward trajectory distribution $P[x_n|\theta_n]$ and recalling the nonnegativity property of the KL divergence to establish the nonnegativity of the averaged EP; $\Sigma_{X|\Theta}(\theta_n) := \langle \ln \frac{P_F[x_n|\theta_n]}{P_B[\tilde{x}_n|\tilde{\theta}_n]} \rangle_{P_F[x_n|\theta_n]} > 0$. We note that this still a partially averaged EP, conditioned on the stochastic trajectory of the parameters. This is indeed the consequence of working in the conditional view.

Motivated to compute the L-info, in the next step, we rearrange Equation (22) as follows:

$$\mathcal{I}[x_{t_n} : \theta_{t_n}] - \mathcal{I}[x_0 : \theta_0] = -q_{x_n}(\theta_n) + s[p(x_{t_n})] - s[p(x_{t_0})] - \sigma_{x_n|\theta_n}, \quad (24)$$

where $\mathcal{I}[x_{t_n} : \theta_{t_n}] := s[p(x_{t_n})] - s[p(x_{t_n}|\theta_{t_n})]$ is the mutual content (or stochastic mutual information) at $t = t_n$. We now arrive at the partially averaged L-info (Equation (10)) by averaging Equation (24) over $P[x_n|\theta_n]$:

$$\begin{aligned}I_{X;\Theta}(\theta_{t_n}) - I_{X;\Theta}(\theta_0) &= -Q_X(\theta_n) + (S_X(\theta_{t_n}) - S_X(\theta_0)) - \Sigma_{X|\Theta}(\theta_n) \\ &= \Sigma_X(\theta_n) - \Sigma_{X|\Theta}(\theta_n).\end{aligned}\quad (25)$$

This defines a series of Partially Averaged (PA) quantities:

$$\begin{aligned}Q_X(\theta_n) &:= \sum_{i=1}^n \langle \phi_{\theta_{t_i}}(x) \rangle_{p(x|\theta_{t_i})} - \langle \phi_{\theta_{t_i}}(x) \rangle_{p(x|\theta_{t_{i-1}})} && \text{(PA Heat flow)} \\ S_{X|\Theta}(\theta_{t_i}) &:= \langle -\log(p(x|\theta_{t_i})) \rangle_{p(x|\theta_{t_i})} && \text{(PA Conditional Entropy)} \\ S_X(\theta_{t_i}) &:= \langle -\log(p(x)) \rangle_{p(x|\theta_{t_i})} && \text{(PA Marginal Entropy)} \\ \Sigma_X(\theta_n) &:= (S_X(\theta_{t_n}) - S_X(\theta_0)) - Q_X(\theta_n) && \text{(PA Marginal EP)}\end{aligned}$$

We note that all PA quantities are conditioned on the parameters' trajectory (i.e., the choice of θ_n from the ensemble). This is a direct consequence of working in the conditional view. However, this also signifies that all thermodynamic quantities mentioned above are computable in the practice of machine learning, as they only require access to the time evolution of one PPM. Fortunately, thanks to the low-variance condition in Equation (9), we can use the partially averaged L-info as a proxy to the L-info.

Equation (25) equates the L-info to the difference between the marginal EP, and the conditional EP. We refer to this difference as the *ignorance* EP:

$$\Sigma_{ign}(\theta_n) := \Sigma_X(\theta_n) - \Sigma_{X|\Theta}(\theta_n) \quad (26)$$

It is important to note that both the marginal EP and the conditional EP measure the EP of the same process, which is the time evolution of the subsystem X . However, the conditional EP measures this quantity with a lower time resolution of α , which is conditioned on the parameter's subsystem. On the other hand, the marginal EP measures this quantity with a higher time resolution of δt , including the relaxation time of the subsystem X between each parameter's update. Therefore, the term "ignorance" refers to ignorance of the full dynamic of X , and the origin of the L-info is the EP between each consecutive parameter's update (i.e., the EP of generating fresh samples represented with the Markov chain in Equation (11)).

5. M-Info and the Parameter Subsystem

We can also apply the DFT to subsystem Θ :

$$\begin{aligned} \sigma_{\theta_n} &= \log \frac{P[\theta_n]}{P[\tilde{\theta}_n]} \\ &= -q_{\theta_n} + s[p(\theta_{t_n})] - s[p(\theta_{t_0})]. \end{aligned} \quad (27)$$

In the above expression, the second line is due to the decomposition in Equation (17), and the definition of the stochastic heat flow for the parameter subsystem is $q_{\theta_n} := \log P[\theta_n|\theta_0]/P[\tilde{\theta}_n|\theta_{t_n}]$.

Under the assumption that the subsystem Θ evolves quasi-statically, the EP of this subsystem is zero, as expected for an ideal reservoir. This results in $q_{\theta_n} = \Delta_{t_n}s[p(\theta_t)]$. Furthermore, in the closed system of (X, Θ) , the heat flow of the subsystem X must be provided with an inverse flow of the subsystem Θ (i.e., $q_{x_n}(\theta_n) = -q_{\theta_n}$). Thus, we arrive at the stochastic version of the Clausius relation for the heat reservoir:

$$\Delta_{t_n}s[p(\theta_t)] = -q_{x_n}(\theta_n) \quad (28)$$

This relation states that the heat dissipation in subsystem X ($q_{x_n}(\theta_n) < 0$) is compensated by an increase in the parameter subsystem entropy. We recall the definition of M-info from Equation (5) as the entropy of the subsystem Θ . Since heat dissipation is a source of L-info accumulation (see Equation (25)), the above Clausius relation states that this information is stored in the parameters by increasing the entropy of this subsystem (i.e., the M-info), confirming the role of parameters as the memory space of the PPM.

We can also take the ensemble average of Equation (28) (i.e., averaging over $P[x_n, \theta_n]$):

$$\Delta_{t_n}S[\Theta_t] = -Q_X(t_n), \quad (29)$$

where $Q_X(t_n) := \sum_{x_n, \theta_n} P[x_n, \theta_n] q_{x_n}(\theta_n) = \sum_{\theta_n} P[\theta_n] Q_X(\theta_n)$ is the fully averaged dissipated heat from the subsystem X . However, under the low-variance condition of learning (Equation (9)), we expect $Q_X(\theta_n)$ to be independent of a specific parameters' trajectory from the ensemble. Thus, we can write $Q_X(t_n) \approx Q_X(\theta_n)$.

The Ideal Learning Process

The learning objective necessitates an increase in L-info to enhance the model's performance while simultaneously reducing the M-info to minimize the generalization error and prevent overfitting. As previously mentioned in Section 2, the ideal scenario is achieved when all the stored information in the parameters (M-info) matches the task-oriented information learned by the model (L-info). Now that we have studied the machinery for

computing these two information-theoretic quantities through the computation of entropy production, we can formally examine this optimal learning condition.

Maximizing the L-info, as described in Equation (25), is equivalent to maximizing the marginal EP while minimizing the conditional EP. Given that the conditional EP is always nonnegative, the “ideal” scenario would involve achieving a conditional EP of zero (i.e., $\Sigma_{X|\Theta}(t_n) = 0$). This condition can be realized through a quasi-static time evolution of the PPM occurring on the lower-resolution timescale α , presented in the Markov chain in Equation (12). In the context of generative models, this condition is akin to achieving perfect sampling. Under these circumstances, all EP of the subsystem X transforms into L-info, resulting in $\Delta_{t_n} I_{X;\Theta}(\theta_t) = \Sigma_X(t_n)$.

Thermodynamically, the condition of quasi-static time evolution of the PPM (and consequently the zero conditional EP) can be realized by having a large relaxation parameter $\tau \gg 1$, which allows the model to reach equilibrium after each optimization step. However, a high relaxation parameter comes at the cost of requiring more computational resources and longer computation times. This introduces a fundamental trade-off between the time required to run a learning process and its efficiency, a concept central to thermodynamics and reminiscent of the Carnot cycle, representing an ideal engine that requires an infinite operation time.

6. The Parameters’ Reservoir

In the formulation of the previous sections, we make the assumption that the subsystem Θ behaves as an ideal reservoir. In this section, we delve deeper into the premises of this assumption by studying the dynamics of the parameter subsystem. To facilitate our formulation, we adapt a negative log-likelihood as a fairly general form for the loss function,

$$\ell(b_t, \theta) := -\frac{1}{|b_t|} \sum_{x \in b_t} \log(p_\theta(x)) = \phi_\theta(b_t). \quad (30)$$

Here, the loss function is computed according to the empirical average of a random mini-batch $b_t \in B$ drawn from the training dataset at time step t . The last equality is due to the PPM defined in Equation (1) and $\phi_\theta(b_t) := \frac{1}{|b_t|} \sum_{x \in b_t} \phi_\theta(x)$, where the notation $|\cdot|$ shows the size of a set. We also use a vanilla stochastic gradient descent (SGD) optimizer with the learning rate r to take gradient steps iteratively for n steps in the direction of loss function minimization:

$$\theta_{t+1} = \theta_t - r \nabla_\theta \phi_\theta(b_t)|_{\theta=\theta_t} \quad (31)$$

To render the dynamics of parameters in the form of a conventional overdamped Langevin dynamic, we introduce the following conservative potential, defined by the entire training dataset B :

$$U_B(\theta) := \frac{1}{|B|} \sum_{x \in B} \phi_\theta(x). \quad (32)$$

The negative gradient of this potential gives rise to a deterministic vector force. Additionally, we define the fluctuation term, which represents the source of random forces due to selection of a mini-batch at time step t_n :

$$\eta(t_n) := -\nabla_\theta \phi_{\theta_t}(b_{t_n}) + \nabla_\theta U_B(\theta_{t_n}).$$

We now reformulate the SGD optimizer in Equation (31) in the guise of overdamped Langevin dynamics, dividing it by the parameters’ update timescale α to convert the learning protocol into a dynamic over time:

$$\frac{\theta_{t_{n+1}} - \theta_{t_n}}{\alpha} = -\mu \nabla_\theta U_B(\theta_{t_n}) + \mu \eta(t_n), \quad (33)$$

where $\mu := r/\alpha$ is known as the mobility constant in the context of Brownian motion.

We note that Equation (33) is merely a rearrangement of the standard SGD. For us to interpret it as a Langevin equation, the term $\eta(t_n)$ must represent a stationary stochastic process to serve as the *noise* term in the Langevin equation. To demonstrate this property of $\eta(t_n)$, we must examine the characteristic of its time correlation function (TCF) [33]: $C_{i,j}(t, t - t') := \delta_{i,j} \langle \eta_i(t) \eta_j(t') \rangle$, where the indices i, j represent different components of the vector θ and $\delta_{i,j}$ is the Kronecker delta.

If the fluctuation term η satisfies the condition of the white noise (uncorrelated stationary random process), and assuming that Equation (33) describes a motion akin to Brownian motion, then we can apply the fluctuation-dissipation theorem to write

$$\langle \eta_i(t) \eta_j(t') \rangle = \frac{2k_B T}{\mu} \delta(t - t') \delta_{i,j} \quad (34)$$

Here, $\delta(t - t')$ is a delta Dirac, k_B stands for the Boltzmann constant, and the constant T symbolizes the temperature. To render our framework as unitless, we treat the product of the Boltzmann factor and temperature as dimensionless. Moreover, regardless of the noise width, we set $T = 1$, and henceforth it will not appear in our formulation. This is possible by adjusting the Boltzmann factor according to the noise width (i.e., $k_B = \mu \langle \eta_i(t) \eta_i(t) \rangle / 2$).

We still need to investigate if the fluctuation term indeed describes an uncorrelated stationary random process, as presented in Equation (34). To this end, we conducted an experiment by training an ensemble of 50 models for the classification of the MNIST dataset. To induce different levels of stochastic behavior (i.e., different “temperatures”), we considered three different mini-batch sizes. A smaller mini-batch size leads to a bigger deviation in the fluctuation term, consequently amplifying the influence of random forces. The results are presented in Figure 3.

The plot in Figure 3c represents the TCF function at no time lag $t = t'$ (i.e., a variance of $\eta(t)$) as a function of time. The constant value of variance suggests the stationary property of $\eta(t)$. Moreover, Figure 3d illustrates the autocorrelation of $\eta(t)$ at different time lags, indicating the white noise characteristic for this term.

However, it would be naive to draw a generic conclusion regarding the nature of the fluctuation term as an uncorrelated stationary random process solely based on a simple experiment. Indeed, research has demonstrated that the noise term can be influenced by the Hessian matrix of the loss function [34]. This observation aligns with our definition of the fluctuation term presented in Equation (33), where η is defined in relation to the gradient of the loss itself. Consequently, as the optimizer explores the landscape of the loss function, the characteristics of the fluctuation term η can vary. We can grasp this concept in the context of Brownian motion by envisioning a Brownian particle transitioning from one medium to another, each with distinct characteristics. This implies that there could be intervals during training where η stays independent of the loss function and exhibits a stationary behavior.

Moreover, we overlooked the fact that $\eta(t)$ is also a function of θ itself. This could potentially jeopardize its stationary property. To address this issue, we refer to the slow dynamic (lazy dynamic) [27,28] of over-parameterized models under SGD optimization. This slow dynamic allows us to write the Taylor expansion (similar to what has been performed in neural tangent kernel theory [35], but with a different purpose) of the loss function around a microscopic state θ^* , sampled from its current state $p_t(\theta)$:

$$\phi_{\theta_t}(b_t) = \phi_{\theta^*}(b_t) + (\theta_t - \theta^*) \nabla_{\theta} \phi_{\theta^*}(b_t) \quad (35)$$

As a result, the gradient of the loss $\nabla_{\theta} \phi_{\theta_t}(b_t) = \nabla_{\theta} \phi_{\theta^*}(b_t)$, signifying an independent behavior from the specific value of the parameter θ_t at a given time t . We can extend this concept to the deterministic force $-\nabla_{\theta} U_B(\theta_t) = -\nabla_{\theta} U_B(\theta^*) = F(\theta^*)$, which indicates a conservative force in the lazy dynamics regime, denoted as $F(\theta^*)$. This indicates a conservative force. The key point here is that the value of this force is independent of the exact microscopic state θ_t , but rather on any typical sample θ^* from Θ_t . In Appendix A,

we illustrate how the condition of lazy dynamics leads to a thermodynamically reversible dynamic of the subsystem Θ .

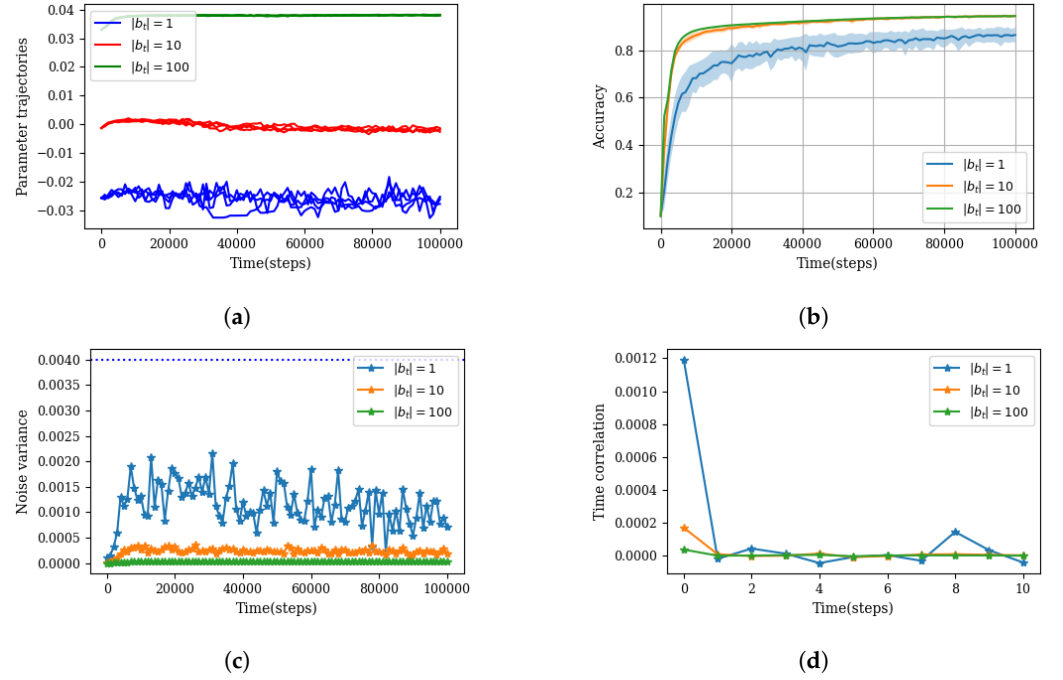


Figure 3. This experiment contrasts the parameter dynamics with three different mini-batch sizes: $|b_t| = 1$, $|b_t| = 10$ and $|b_t| = 100$. The model under consideration is a four-layer feedforward neural network with a uniform width of 200 neurons. It was trained on the MNIST classification task using a vanilla SGD optimizer. The experiment was replicated over 50 trials to generate an ensemble of parameters. (a) One random parameter from the model’s last layer was chosen for each batch size scenario, and four of its dynamic realizations are depicted. (b) Illustration of both the average accuracy (solid line) and the variance in accuracy within the ensemble (shaded area), emphasizing the low-variance condition, which asserts that macroscopic quantities such as accuracy have low variance statistics across the ensemble. (c) The noise variance averaged over all parameters (i.e., $\frac{1}{\dim(\theta)} \sum_{i=0}^{\dim(\theta)} C_{i,i}(t, 0)$) for each mini-batch size scenario, underscoring the stationary nature of η . This part also highlights the role of the mini-batch size in determining the noise width (i.e., the temperature of the environment). The horizontal dashed line indicates the maximum absolute value observed from $\nabla_{\theta} U_B(\theta_{t_n})$, serving as a reference point for the magnitude of the noise. (d) The autocorrelation of the term η averaged over all parameters. For instance, computing this quantity at step 1000 reads as follows: $\frac{1}{\dim(\theta)} \sum_{i=0}^{\dim(\theta)} C_{i,i}(t = 1000, t' - t)$. The rapid decline in autocorrelation with time lag indicates the white noise characteristic of η .

6.1. Naive Parametric Reservoir

The stationary state of the subsystem Θ , under the dynamic of Equation (33) and satisfying the fluctuation-dissipation relation in Equation (34), corresponds to the thermal equilibrium state (the canonical state):

$$p^{eq} = e^{-U_B(\theta) + F_{\Theta}} \quad (36)$$

where $F_{\Theta} := -\log(\int d\theta e^{-U_B(\theta)})$ is the free energy of the subsystem θ . Recall that the temperature was set to one. This state also satisfies the detailed balance condition, which defines the log ratio between the forward and backward transition probabilities (presented in Equation (16)) as follows:

$$\log \frac{p(\theta_{t_i} | \theta_{t_{i-1}})}{p(\theta_{t_{i-1}} | \theta_{t_i})} = -(U_B(\theta_{t_i}) - U_B(\theta_{t_{i-1}})) \quad (37)$$

The standard plot of the loss function versus the optimization steps in machine learning practice can help us to visualize the dynamics of the subsystem Θ . A rapid decline in the loss function signals a swift relaxation of the subsystem Θ to its equilibrium state. It is important to note that this *self-equilibrating property* is determined by the training dataset B through the definition of the potential function $U_B(\theta)$. These swift and self-equilibrating properties mirror the characteristics of a heat reservoir in thermodynamics [29]. This observation supports the foundations on which we designated the subsystem Θ as the “parametric reservoir”. After a swift decline, a gradual reduction of the loss function can be a sign of a quasi-statistic process, where the subsystem Θ evolves from one equilibrium state to another. This can be due to the lazy dynamic condition, as discussed in Appendix A. Additionally, the requirement of a high heat capacity for the reservoir ($\dim(\Theta) \gg \dim(X)$), offers a thermodynamic justification for the use of over-parameterized models in machine learning.

6.2. Realistic Parametric Reservoir

We refer to the assumption of the parametric reservoir with an equilibrium state expressed in Equation (36) as the “naive assumption” due to several issues that were previously sidestepped. The first issue stems from the assumption that all components of the parameter vector θ are subject to the same temperature (i.e., $\langle \eta_i(t) \eta_i(t) \rangle = \frac{2k_B T}{\mu}$ for all indexes i). In practice, we might find different values for the noise width, particularly with respect to different layers of a deep neural network. Furthermore, the weights or biases within a specific layer might experience different amounts of fluctuation. This scenario is entirely acceptable if we consider each group of parameters as a subsystem that contributes to the formation of the parametric reservoir. Consequently, each subsystem possesses different environmental temperatures and distinct stationary states. This observation may explain, in thermodynamic terms, why a deep neural network can offer a richer model. As it encompasses multiple heat reservoirs at varying temperatures, it presents a perfect paradigm for the emergence of non-equilibrium thermodynamic properties.

Second, the fluctuation term η may exhibit an autocorrelation property that characterizes colored noise, as presented in [36]. While this introduces additional richness to the problem, potentially displaying non-Markovian properties, it does not impede us from deriving the equilibrium state, as demonstrated in [37].

We also overlooked the irregular behavior of the loss function, such as spikes or step-like patterns. These irregularities are considered abnormal as we typically expect the loss function to exhibit a monotonous decline, but in practice, such behaviors are quite common. These anomalies may be associated with a more intricate process experienced by the reservoir, such as a phase transition or a shock. Nevertheless, we can still uphold the basic parametric reservoir assumption during the time intervals between these irregular behaviors.

The mentioned issues are attributed to a richer and more complex dynamic of the subsystem Θ and do not fundamentally contradict the potential role of subsystem Θ as a reservoir. Examples of these richer dynamics can be found in a recent study [38], which shows the limitation of the Langevin formulation of the SGD, and Ref. [39], which investigates the exotic non-equilibrium characteristic of parameters’ dynamics under SGD optimization.

Before closing this section, it is worth mentioning that the experimental results presented in Figure 3 support the assumption of a low-variance condition for the stochastic dynamics of the subsystem Θ . For instance, panel (a) shows that even in the high noise regime ($|b_i| = 1$), the dynamics of the parameters remained confined to a small region across the ensemble. Furthermore, panel (b) demonstrates the low-variance characteristics of the model’s performance accuracy. Finally, the large magnitude of the deterministic force (dashed line in panel (c)) to the random force is evidence of low-variance dynamics.

7. Discussion

In this study, we delved into the thermodynamic aspects of machine learning algorithms. Our approach involved first formulating the learning problem as the time evolution of a PPM. Consequently, the learning process naturally emerged as a thermodynamic process. This process is driven by the work of the optimizer, which can be considered a thermodynamic work since parameter optimization constantly changes the system's energy landscape through $\phi_\theta(x)$. The optimizer action is fueled by the input trajectory and a series of samples drawn from the ground truth system. The work and heat exchange of the subsystem X can be computed practically along the learning trajectory $\mathcal{T}$ as outlined below:

$$W_X(\theta_n) = \sum_{t=1}^n \langle \phi_{\theta_t}(x) \rangle_{p(x|\theta_{t-1})} - \langle \phi_{\theta_{t-1}}(x) \rangle_{p(x|\theta_{t-1})} \quad (38)$$

$$Q_X(\theta_n) = \sum_{t=1}^n \langle \phi_{\theta_t}(x) \rangle_{p(x|\theta_t)} - \langle \phi_{\theta_t}(x) \rangle_{p(x|\theta_{t-1})} \quad (39)$$

We use the term “practically” because when running a machine learning algorithm, we have access to the function $\phi_\theta(x)$ at each training instance. We also note that these quantities are conditioned on a specific parameter's trajectory as a result of working in the conditional view. Finally, the learning process can be summarized as follows: *The model learns by dissipating heat, and the dissipated heat increases the entropy of the parameters, which act as the heat reservoir (a memory space) for learned information.* This means the learning process must be irreversible, as this is the only way to increase the mutual information between the two subsystem X and Θ [10].

It is important to note that despite the wealth of research highlighting the significance of information content in parameters [21,22,25], calculating these quantities remains difficult due to the lack of access to the parameter distribution. In contrast, the thermodynamic approach computes the information-theoretic metrics indirectly as the heat and work of the process. Moreover, the mysterious success of over-parametrized models can be explained within the thermodynamic framework, where over-parameterization plays a crucial role in allowing the parameter subsystem to function as a heat reservoir.

At the same time, we are aware of the strong assumptions made during this study. Addressing each of these assumptions or providing justifications for them represents a direction for future research. For instance, we assumed slow dynamics for over-parameterized models using the SGD optimizer. This formed the basis for treating the parameters' degrees of freedom as an ideal heat reservoir, evolving in a thermodynamically reversible manner. Breaking this assumption leads to entropy production of parameter subsystem, changing contribution of entropy production to accumulation of L-info.

We also sidestepped the role of changes in the marginal entropy of the model's subsystem $\Delta_{t_n} S_X(t)$. This term can be estimated by computing the entropy of the empirical distribution of generated samples. For a model initialized randomly, this term is always negative, as the initial model produces uncorrelated patterns with maximum entropy. Then, the negative value of this term must converge when the entropy of the generated patterns reaches the entropy of the training dataset, or a representation of the training dataset that is most relevant to the learning objective. However, if we look at Equation (25) as an optimization objective to maximize the L-info, then an increase in the model's generated samples $S_X(t)$ is favorable. This might act as a regularization term to improve the generalization power of the model by forcing it to avoid easy replication of the dataset.

Funding: This research was funded by the National Institutes of Health BRAIN Initiative (R01EB026943), and the ITS-Simons Foundation fellowship (Fall 2022 and Spring 2023).

Data Availability Statement: This is a more theoretical work, data sharing is not applicable to this article.

Acknowledgments: The author expresses gratitude for the support provided by David J. Schawab during this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Reversibility under the Lazy Dynamic Regime

In this appendix, we establish the thermodynamic reversibility of parameter evaluation as a consequence of training an over-parameterized model with lazy dynamics. The forward action of the optimizer can be summarized as follows: $\mathbf{b}_n \rightarrow \boldsymbol{\theta}_n$, where the optimizer samples an independent and identically distributed trajectory of inputs from the training dataset, $\mathbf{b}_n = \{b_{t_1}, b_{t_2}, \dots, b_{t_n}\}$, to generate a trajectory of the updated parameters, $\mathbf{b}_n = \{\theta_0, \theta_{t_1}, \dots, \theta_{t_n}\}$.

The backward (time reversal) action of the optimizer is defined as $\tilde{\mathbf{b}}_n \rightarrow \boldsymbol{\theta}_n^\dagger$, where $\tilde{\mathbf{b}}_n = \{b_{t_n}, b_{t_{n-1}}, \dots, b_{t_1}\}$ represents the time reversal of the input trajectory and the gradient descent is reversed to a gradient ascent, resulting in a new parameter trajectory $\boldsymbol{\theta}_n^\dagger$.

In general, the backward action of the SGD does not yield the time reversal of the forward parameter trajectory:

$$\boldsymbol{\theta}_n^\dagger \neq \tilde{\boldsymbol{\theta}}_n = \{\theta_{t_n}, \theta_{t_{n-1}}, \dots, \theta_{t_0}\}$$

To illustrate this, let us examine a single forward and backward action of the optimizer:

$$\theta_{t+1} = \theta_t - r \nabla_{\theta} \phi_{\theta}(b_t)|_{\theta=\theta_t} \quad (\text{Forward step})$$

$$\theta_t^\dagger = \theta_{t+1} + r \nabla_{\theta} \phi_{\theta}(b_t)|_{\theta=\theta_{t+1}} \quad (\text{Backward step})$$

This discrepancy arises due to the gradient step's dependence on the current values of the parameters in both the forward and backward optimizations (i.e., $\nabla_{\theta} \phi_{\theta}(b_t)|_{\theta=\theta_t} \neq \nabla_{\theta} \phi_{\theta}(b_t)|_{\theta=\theta_{t+1}}$).

However, the key observation here is that under the lazy dynamic regime (as described in Equation (35)), this dependency vanishes, and both forward and backward gradient step can be written as $\nabla_{\theta} \phi_{\theta}(b_t)|_{\theta=\theta^*}$, where θ^* is a typical sample from the stationary state (or slowly varying state) of parameters. Under such conditions, the backward action of the SGD (running the learning protocol backward) results in a time reversal of the parameters' trajectory, namely $\boldsymbol{\theta}_n^\dagger = \tilde{\boldsymbol{\theta}}_n$, signifying the thermodynamic reversibility of the parameters' subsystem under lazy dynamic conditions. Thus, the lazy dynamics lead to a quasi-static evolution of the parameter subsystem, meaning that the subsystem Θ itself does not contribute to entropy production and acts as an ideal heat reservoir. See Ref. [40] for a detailed discussion on thermodynamic reversibility.

References

- Landauer, R. Irreversibility and heat generation in the computing process. *IBM J. Res. Dev.* **1961**, *5*, 183–191. [\[CrossRef\]](#)
- Szilard, L. On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings. *Z. Phys.* **1929**, *53*, 840. [\[CrossRef\]](#)
- Bennett, C.H. The thermodynamics of computation—A review. *Int. J. Theor. Phys.* **1982**, *21*, 905–940. [\[CrossRef\]](#)
- Nielsen, M.A.; Chuang, I.L. *Quantum Computation and Quantum Information: 10th Anniversary Edition*; Cambridge University Press: Cambridge, MA, USA, 2010.
- Almheiri, A.; Hartman, T.; Maldacena, J.; Shaghoulian, E.; Tajdini, A. The entropy of hawking radiation. *Rev. Mod. Phys.* **2021**, *93*, 035002. [\[CrossRef\]](#)
- Parrondo, T.S.J.; Horowitz, J. Thermodynamics of information. *Nat. Phys.* **2015**, *2*, 131–139. [\[CrossRef\]](#)
- Peliti, L.; Pigolotti, S. *Stochastic Thermodynamics: An Introduction*; Princeton University Press: Princeton, NJ, USA, 2021.
- Still, S.; Sivak, D.A.; Bell, A.J.; Crooks, G.E. Thermodynamics of prediction. *Phys. Rev. Lett.* **2012**, *109*, 120604. [\[CrossRef\]](#) [\[PubMed\]](#)
- Sagawa, T.; Ueda, M. Fluctuation theorem with information exchange: Role of correlations in stochastic thermodynamics. *Phys. Rev. Lett.* **2012**, *109*, 180602. [\[CrossRef\]](#) [\[PubMed\]](#)
- Esposito, M.; Lindenberg, K.; den Broeck, C.V. Entropy production as correlation between system and reservoir. *New J. Phys.* **2010**, *1*, 013013. [\[CrossRef\]](#)
- Song, Y.; Kingma, D.P. How to train your energy-based models. *arXiv* **2021**, arXiv:2101.03288.

12. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
13. Jeon, H.J.; Zhu, Y.; Roy, B.V. An information-theoretic framework for supervised learning. *arXiv* **2022**, arXiv:2203.00246.
14. Yi, J.; Zhang, Q.; Chen, Z.; Liu, Q.; Shao, W. Mutual information learned classifiers: An information-theoretic viewpoint of training deep learning classification systems. *arXiv* **2022**, arXiv:2210.01000.
15. Shwartz-Ziv, R.; LeCun, Y. To compress or not to compress- self-supervised learning and information theory: A review. *arXiv* **2023**, arXiv:2304.09355.
16. Yu, S.; Giraldo, L.G.S.; Príncipe, J.C. Information-theoretic methods in deep neural networks: Recent advances and emerging opportunities. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 19–27 August 2021.
17. Geiger, B.C. On information plane analyses of neural network classifiers—A review. *arXiv* **2021**, arXiv:2003.09671.
18. Achille, A.; Paolini, G.; Soatto, S. Where is the information in a deep neural network? *arXiv* **2019**, arXiv:1905.12213.
19. Shwartz-Ziv, R.; Tishby, N. Opening the black box of deep neural networks via information. *arXiv* **2017**, arXiv:1703.00810.
20. Andrew, M.; Bansal, Y.; Dapello, J.; Advani, M.; Kolchinsky, A.; Tracey, B.D.; Cox, D.D. On the information bottleneck theory of deep learning. *J. Stat. Mech. Theory Exp.* **2018**, 2018, 124020.
21. Hinton, G.E.; Camp, D.V. Keeping the neural networks simple by minimizing the description length of the weights. In Proceedings of the Sixth Annual Conference on Computational Learning Theory, Santa Cruz, CA, USA, 26–28 July 1993; pp. 5–13.
22. Achille, A.; Soatto, S. Emergence of invariance and disentanglement in deep representations. *J. Mach. Learn. Res.* **2018**, 19, 1947–1980.
23. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley and Sons: New York, NY, 1991.
24. Rissanen, J. Stochastic complexity and modeling. *Ann. Stat.* **1986**, 14, 1080–1100. [[CrossRef](#)]
25. Bu, Y.; Zou, S.; Veeravalli, V.V. Tightening mutual information-based bounds on generalization error. *IEEE J. Sel. Areas Inf. Theory* **2020**, 1, 121–130. [[CrossRef](#)]
26. den Broeck, C.V.; Esposito, M. Ensemble and trajectory thermodynamics: A brief introduction. *Phys. A Stat. Mech. Its Appl.* **2015**, 418, 6–16. [[CrossRef](#)]
27. Du, S.S.; Zhai, X.; Poczos, B.; Singh, A. Gradient descent provably optimizes over-parameterized neural networks. *arXiv* **2018**, arXiv:1810.02054.
28. Li, Y.; Liang, Y. Learning overparameterized neural networks via stochastic gradient descent on structured data. *Adv. Neural Inf. Process. Syst.* **2018**, 31. [[CrossRef](#)]
29. Deffner, S.; Jarzynski, C. Information processing and the second law of thermodynamics: An inclusive, hamiltonian approach. *Phys. Rev. X* **2013**, 3, 041003. [[CrossRef](#)]
30. Du, Y.; Mordatch, I. Implicit generation and generalization in energy-based models. *arXiv* **2019**, arXiv:1903.08689.
31. Maes, C. *Local Detailed Balance*; SciPost Physics Lecture Notes; SciPost: Amsterdam, The Netherlands, 2021.
32. Rao, R.; Esposito, M. Detailed fluctuation theorems: A unifying perspective. *Entropy* **2018**, 20, 635. [[CrossRef](#)] [[PubMed](#)]
33. Zwanzig, R. *Nonequilibrium Statistical Mechanics*; Oxford University Press: Oxford, UK, 2001.
34. Wei, M.; Schwab, D.J. How noise affects the hessian spectrum in overparameterized neural networks. *arXiv* **2019**, arXiv:1910.00195.
35. Jacot, A.; Gabriel, F.; Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Adv. Neural Inf. Process. Syst.* **2018**, 31.
36. Kühn, M.; Rosenow, B. Correlated noise in epoch-based stochastic gradient descent: Implications for weight variances. *arXiv* **2023**, arXiv:2306.05300.
37. Ceriotti, M.; Bussi, G.; Parrinello, M. Langevin equation with colored noise for constant-temperature molecular dynamics simulations. *Phys. Rev. Lett.* **2009**, 102, 020601. [[CrossRef](#)]
38. Ziyin, L.; Li, H.; Ueda, M. Law of balance and stationary distribution of stochastic gradient descent. *arXiv* **2023**, arXiv:2308.06671.
39. Adhikari, S.; Kabakçioğlu, A.; Strang, A.; Yuret, D.; Hinczewski, M. Machine learning in and out of equilibrium. *arXiv* **2023**, arXiv:2306.03521.
40. Sagawa, T. Thermodynamic and logical reversibilities revisited. *J. Stat. Mech. Theory Exp.* **2014**, 2014, P03025. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.