Neural Entropy

Akhil Premkumar
Kavli Institute for Cosmological Physics
University of Chicago
Chicago, IL 60637
akhilprem@uchicago.edu

Abstract

We examine the connection between deep learning and information theory through the paradigm of diffusion models. Using well-established principles from non-equilibrium thermodynamics we can characterize the amount of information required to reverse a diffusive process. Neural networks store this information and operate in a manner reminiscent of Maxwell's demon during the generative stage. We illustrate this cycle using a novel diffusion scheme we call the entropy matching model, wherein the information conveyed to the network during training exactly corresponds to the entropy that must be negated during reversal. We demonstrate that this entropy can be used to analyze the encoding efficiency and storage capacity of the network, and raises the prospect of applying diffusion models as a test bench to understand neural networks.

1 Introduction

In this paper we explore whether ideas from thermodynamics and information theory can be used to understand the behavior of neural networks. Diffusion models serve as a natural bridge to draw the connection between thermodynamics and machine learning, because they were originally developed by synthesizing ideas from these disciplines [1]. Very briefly, samples from a training dataset are incrementally noised till they are distributed as a generic Gaussian, while a neural network learns to reverse these noising steps. Once trained, the network can transform a random Gaussian vector into a highly structured output that resembles a typical member of the training data. In the continuum limit, the noising and denoising stages become diffusive processes [2], the thermodynamic properties of which are well established. A generative model based on the diffusion paradigm must also follow the rules of thermodynamics, specifically the Second Law—to create structure out of noise in one part of the system the model must either produce disorder elsewhere, or apply information to negate the entropy like Maxwell's demon. The latter is how diffusion models operate.

Diffusion gradually wipes out information from the system over time, but the process can be reversed by reinstating the lost information. In a diffusion model the neural network stores this information. Entropy is a measure of 'missing information', so the total entropy produced during the forward process also quantifies the amount of information the network needs to remember and put back during reversal (see Fig. 1). This is is what we refer to as *neural entropy*.

The entropic point of view can be demonstrated with a reparameterization of the reverse diffusion process from the popular denoising score matching formulation [4]. We call this the *entropy matching model*. The neural entropy of this model is the total dissipation in the forward diffusion process, which has a lower bound related to the L^2-Wasserstein distance between the initial data distribution and the final Gaussian [5]. Thus we establish a fundamental connection between diffusion models and the theory of optimal transport, opening up a new space of design choices for future diffusion models. Furthermore, diffusion models are infinitely deep variational autoencoders [6], therefore neural entropy offers a way to characterize a neural network's performance as an encoder.

Machine Learning and the Physical Sciences Workshop, NeurIPS 2024.

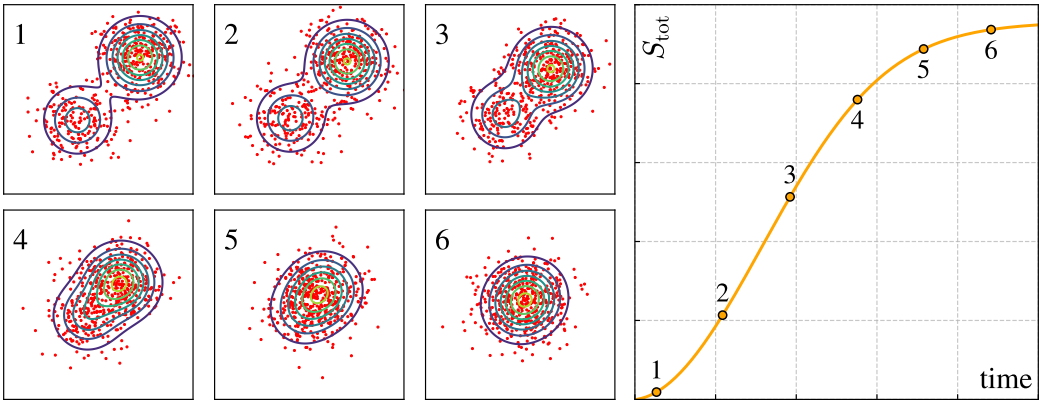


Figure 1: Diffusion is a non-equilibrium process that generates entropy over time. On the left we see snapshots of a diffusive process (Ornstein-Uhlenbeck). In the forward direction the distribution evolves from $1 \to 6$, and entropy produced till that point in time is indicated on the right. In the reverse direction, $6 \to 1$, a diffusion model removes this excess entropy using information it learned during training. Note that $S_{\rm tot}$ is the total entropy produced, which is different from the change in Gibbs entropy of the distribution [3].

2 Entropy Matching

Consider the problem of converting a generic prior distribution p_0 to a highly structured target distribution $p_{\rm d}$. We would like to learn how to do this given only a set of samples $y_{\rm d} \sim p_{\rm d}$. Diffusion models accomplish this by applying a diffusive process on $\{y_{\rm d}\}$ (see App. A for notation),

$$\mathrm{d}Y = b_+(Y, s)\mathrm{d}s + \sigma(s)\mathrm{d}\hat{B}_s, \tag{1}$$

which distributes the samples as p_0 after some time T. We will assume that b_+ is a confining drift term. Ideally, this process can be reversed with the SDE [7, 8]

$$\mathrm{d}X = -b_-(X, T-t)\mathrm{d}t + \sigma(T-t)\mathrm{d}B_t, \tag{2}$$

where $t = T - s$, and the drift term is

$$b_- = b_+ - \sigma^2 \nabla \log \overleftarrow{p}. \tag{3}$$

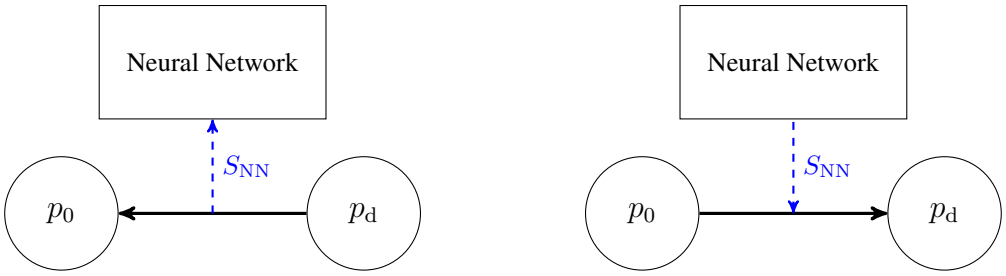
All the nontrivial information needed to transform $p_0 \to p_{\rm d}$ is contained in the *score function* $\nabla \log \overleftarrow{p}$, where the overhead arrow highlights that this object must be learned from the forward process, Eq. (1). This is precisely what score matching models do, by approximating the scores with a neural network $\boldsymbol{s_\theta}$. Replacing b_- with $b_+ - \sigma^2 \boldsymbol{s_\theta}$, it can be shown that the training objective for such models upper bounds the KL divergence between the data distribution and the generated distribution $p_{\boldsymbol{\theta}}$ [9, 6],

$$\int_0^T \mathrm{d}t\, \frac{\sigma^2}{2} \mathbb{E}_p \left[\| \boldsymbol{s_\theta} - \nabla \log p \|^2 \right] \geq \mathrm{D}_{KL}\left(p_{\rm d}(\cdot) \| p_{\boldsymbol{\theta}}(\cdot, T)\right). \tag{4}$$

Alternatively, if we substitute b_- with $-b_+ - \sigma^2 \boldsymbol{e_\theta}$, where $\boldsymbol{e_\theta}$ represents a new neural network in what we call the *entropy matching* model, we obtain the inequality

$$\int_0^T \mathrm{d}t\, \frac{\sigma^2}{2} \mathbb{E}_p \left[\left\| \frac{2b_+}{\sigma^2} - \nabla \log p + \boldsymbol{e_\theta} \right\|^2 \right] \geq \mathrm{D}_{KL}\left(p_{\rm d}(\cdot) \| p_{\boldsymbol{\theta}}(\cdot, T)\right). \tag{5}$$

If the network is disconnected, so that $\boldsymbol{e_\theta} = 0$, the l.h.s. becomes the total entropy produced by the forward process Eq. (1) in converting $p_{\rm d} \to p_0$ [10]. This is $S_{\rm tot}$ from Fig. 1. This quantity has the following physical interpretation: As diffusion progresses, our knowledge of the system diminishes over time. $S_{\rm tot}$ is a measure of this information loss. In entropy matching, the model records this information in its network during training, so the content of the original distribution is not truly lost. Then, we can recover this distribution by re-introducing the stored information back into the system.



(a) Forward diffusion/training (b) Reverse diffusion/generation

Figure 2: An idealized cycle of a diffusion model. During the forward process information lost as $p_{\rm d}$ transforms to p_0 is captured by a neural network. This information is applied to lower the entropy of the system and recover $p_{\rm d}$ in the reverse process. The dashed arrow indicates the flow of information.

The entropy matching model allows us to identify precisely how much information is delivered to the network during training. We may then associate the information in the network with the *neural entropy*

$$S_{\rm NN}(T) := \int_0^T {\rm d}t\, \frac{1}{2\sigma^2} \mathbb{E}_p \left[\left\| 2b_+ - \sigma^2 \nabla \log p \right\|^2 \right]. \tag{6}$$

It is important to stress that $S_{\rm NN}$ quantifies the maximum information stored in a perfectly trained network; it is *not* the entropy of an internal phase space density over the neural network's microstates. In a diffusion model the network is configured as a memory, and it must be able to retain $S_{\rm NN}/\log 2$ bits of information. However, the storage capacity of the network is not simply the bit count of the parameters of the network—depending on how the network encodes information, it works as an *effective memory* that furnishes a functional bit count that is different from the sum of its parts.

Attempting to give an entropy interpretation to the bound in Eq. (4) will lead to physical inconsistencies. For example, if we set $\boldsymbol{s_\theta} = 0$ in Eq. (4), the l.h.s. will be a non-zero positive number even in the special case $p_{\rm d} = p_0$ where no information would be lost under diffusion. Therefore, the upper bound in that equation cannot be interpreted as an entropy. A more rigorous justification for the concept of neural entropy, based on the Vaikuntanathan-Jarzynski relation [11], is given in [12].

3 Experiments

A perfectly trained network absorbs and encodes $S_{\rm NN}$ worth of information during the forward process, and uses it to exhume $p_{\rm d}$ from p_0 during reversal (see Fig. 2). However, in practice not all of $S_{\rm NN}$ is retained by the network. A useful probe of network retention is the KL divergence from Eq. (5)—we expect that a network that catches more information during training would reconstruct a $p_{\boldsymbol{\theta}}$ that is closer to $p_{\rm d}$, compared to one that remembers less. We can test this hypothesis using synthetic data for which the true $p_{\rm d}$ is known.

Our experiments are designed to test how a neural network responds to different amounts of information, as characterized by the neural entropy of an entropy matching model, $S_{\rm NN}^{\rm em}$. The latter is a function of $p_{\rm d}, p_0$, and the forward process, Eq. (1). We vary $S_{\rm NN}^{\rm em}$ by changing $p_{\rm d}$, whilst keeping p_0, b_+, and σ fixed. Our $p_{\rm d}$ are Gaussian mixtures. These are low dimensional distributions, typically $D = 4$ to 8, and our network is a simple MLP with Gaussian random feature layers for x and t [13]. The structure of the network is *kept fixed* in all the experiments, so its capacity to store information is the same in all cases. The diffusion model is trained on data sampled from $p_{\rm d}$. These sample are noised by a VP process [2], with $\beta_{\rm min} = 0.1$ and $\beta_{\rm max} = 8$. Then, we compute ${\rm D}_{KL}(p_{\rm d}(\cdot)\|p_{\boldsymbol{\theta}}(\cdot, T))$ on a new set of samples from $p_{\rm d}$. The main output of our experiments are plots of this KL against $S_{\rm NN}^{\rm em}$ (see Fig. 3). There is a clear deterioration in the network performance as we try to push more information into it.

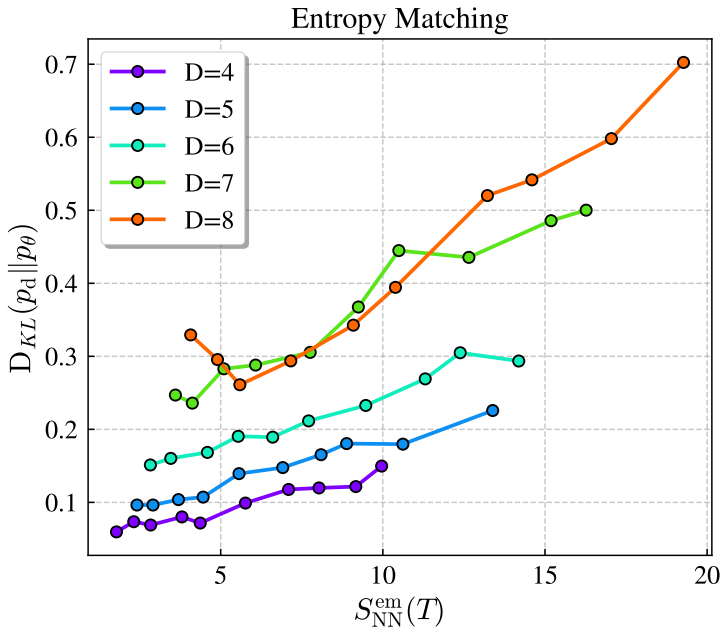


Figure 3: Plots of the KL divergence between the data distribution $p_\mathrm{d}(\cdot)$ and the generated distribution $p_{\boldsymbol{\theta}}(\cdot, T)$ against the neural entropy delivered in an entropy matching model. Each point corresponds to a p_d distribution associated with a different neural entropy. In this plot the p_d's are randomly generated Gaussian mixtures in D dimensions. Lower values of the KL indicate better performance of the diffusion model. Pushing more information into the neural network deteriorates model performance.

4 Conclusions

We have introduced the idea of neural entropy and demonstrated that, with entropy matching diffusion models, it can be used to quantify the information delivered to a neural network. We can gauge the network's effectiveness at encoding and storing this information by how well it reconstructs the data distribution. This paradigm serves as a test bench for neural network architectures, which may be used to assess their performance against different kinds of data and training parameters. The main paper [12] also makes several important conceptual connections, some of which are summarized below:

Optimal transport: The entropy matching model establishes a link between diffusion models, thermodynamics, and optimal transport, through the *Benamou-Brenier formula* [5]. It related the entropy produced in a diffusive process with the L^2-Wasserstein distance between p_0 and p_d [14]. For the diffusion process Eq. (1),

$$S_{\mathrm{tot}} \geq \frac{\mathcal{W}_2(p_\mathrm{d}, p_0)^2}{2\sigma^2 T}, \tag{7}$$

with the diffusion coefficient σ^2 set to a constant. The main consequence of Eq. (7) is that, given the same p_d and p_0, there exists some way of diffusing $p_\mathrm{d} \to p_0$ that incurs minimum entropy production. In practice, we are limited to using forward processes with no drift or an affine drift term, which is unlikely to saturate Eq. (7).

Maxwell's demon: The operational cycle of a diffusion model (see Fig. 2) bears a strong likeness to the famous thought experiment known as Maxwell's demon [15]. In both cases information from a prior measurement is used to alter the entropy of the system in a way that appears to defy the Second Law of Thermodynamics. However, in the reverse diffusion case the information is applied via the drift term, which means the 'diffusion demon' interacts with the diffusing particles.

Score matching: The difficulty with defining neural entropy for score matching models is demonstrated with more detailed experiments in the main paper.

A Notation

Notation: We use the time variable s for the forward diffusion process, which runs from right ($s = 0$) to left ($s = T$) in Fig. 4. Sometimes we indicate functions of s as $\overleftarrow{f}$ to remove ambiguity when the same function is also expressed in terms of time variable $t = T - s$. That is, $\overleftarrow{f}(s) = \overleftarrow{f}(T-t) = f(t)$. $\hat{B}_s$ and B_t denote the Brownian motions associated with the forward and reverse/controlled SDEs, respectively. ∇ is the gradient with respect the spatial coordinates, and ∂_t, ∂_s are partial time derivatives. Throughout the paper, we set Boltzmann's constant to unity, $k_{\mathrm{B}} = 1$. $\log$ is the natural logarithm. p_{d} and p_0 denote the initial ($s = 0$) and final ($s = T$) densities for the forward process. $p_{\boldsymbol{\theta}}(\cdot, 0)$ and $p_{\boldsymbol{\theta}}(\cdot, T)$ are the initial ($t = 0$) and final ($t = T$) densities of the generative process. In this extended abstract we have taken $p_{\boldsymbol{\theta}}(\cdot, 0) = p_0(\cdot)$ for simplicity.

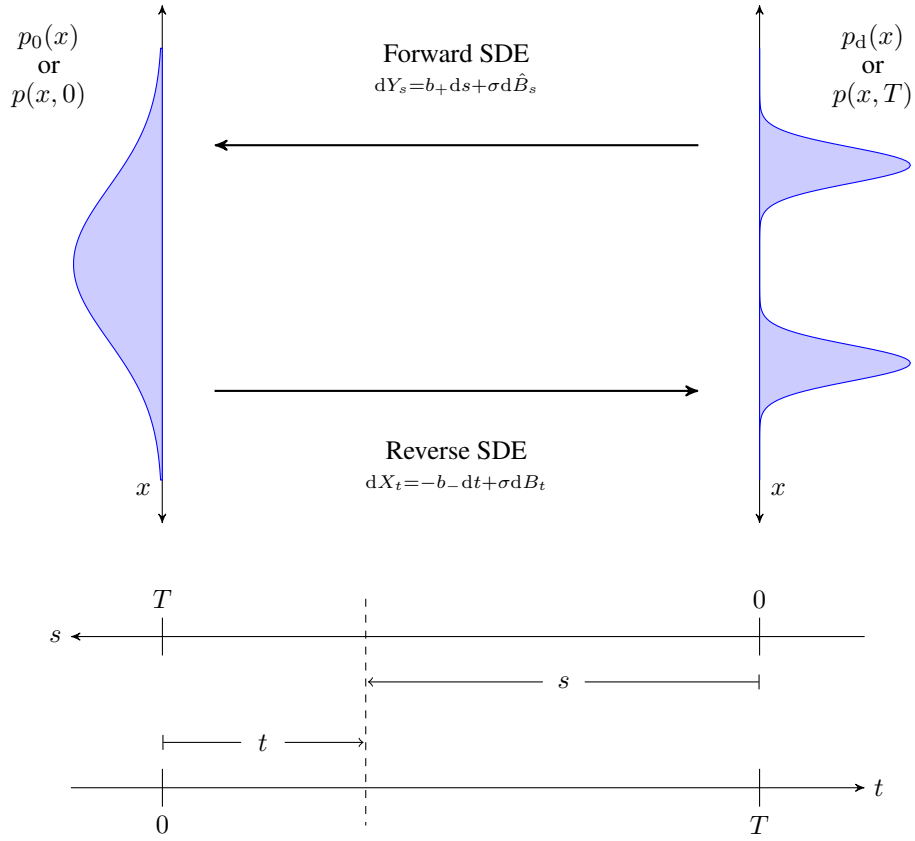


Figure 4: A schematic of the forward and reverse diffusion processes.

References

[1] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

[2] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

[3] Udo Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on Progress in Physics*, 75(12):126001, nov 2012. doi: 10.1088/0034-4885/75/12/126001. URL https://dx.doi.org/10.1088/0034-4885/75/12/126001.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

[5] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, Jan 2000. ISSN 0945-3245. doi: 10.1007/s002110050002. URL https://doi.org/10.1007/s002110050002.

[6] Chin-Wei Huang, Jae Hyun Lim, and Aaron C Courville. A variational perspective on diffusion-based generative models and score matching. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 22863–22876. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/c11abfd29e4d9b4d4b566b01114d8486-Paper.pdf.

[7] Edward Nelson. Derivation of the schrödinger equation from newtonian mechanics. *Phys. Rev.*, 150:1079–1085, Oct 1966. doi: 10.1103/PhysRev.150.1079. URL https://link.aps.org/doi/10.1103/PhysRev.150.1079.

[8] Brian D.O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. ISSN 0304-4149. doi: https://doi.org/10.1016/0304-4149(82)90051-5. URL https://www.sciencedirect.com/science/article/pii/0304414982900515.

[9] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 1415–1428. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/0a9fdbb17feb6ccb7ec405cfb85222c4-Paper.pdf.

[10] Udo Seifert. Entropy production along a stochastic trajectory and an integral fluctuation theorem. *Phys. Rev. Lett.*, 95:040602, Jul 2005. doi: 10.1103/PhysRevLett.95.040602. URL https://link.aps.org/doi/10.1103/PhysRevLett.95.040602.

[11] S. Vaikuntanathan and C. Jarzynski. Dissipation and lag in irreversible processes. *Europhysics Letters*, 87(6):60005, oct 2009. doi: 10.1209/0295-5075/87/60005. URL https://dx.doi.org/10.1209/0295-5075/87/60005.

[12] Akhil Premkumar. Neural entropy, 2024. URL https://arxiv.org/abs/2409.03817.

[13] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7537–7547. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/55053683268957697aa39fba6f231c68-Paper.pdf.

[14] Tan Van Vu and Keiji Saito. Thermodynamic unification of optimal transport: Thermodynamic uncertainty relation, minimum dissipation, and thermodynamic speed limits. *Phys. Rev. X*, 13:011013, Feb 2023. doi: 10.1103/PhysRevX.13.011013. URL https://link.aps.org/doi/10.1103/PhysRevX.13.011013.

[15] J. C. Maxwell. Life and scientific work of peter guthrie tait. In C. G. Knott, editor, *Life and Scientific Work of Peter Guthrie Tait*, page 213. Cambridge University Press, London, 1911.