Diffusion Models and Associative Memory: The Power and Plausibility of Imaginative Generation

Avery Louis, Sisely DeLisi

December 2023

1 Introduction

Associative memory refers to the capacity for identifying relationships between familiar and novel stimuli based on prior experience. Connections between concurrently encountered information clusters enables recall of entire associative networks with a single cue. For example, a pine-scented candle cues an associated memory of a complex event decades past: climbing pine trees with your childhood best friend. Neurologically, associative memory entails synthesis of information across diverse sensory modalities and cognitive processes. This is understood largely through involvement of long term synaptic plasticity, namely the Hebbian model of learning: 'neurons that fire together wire together'. However, reality is likely far more complex, and is not yet completely understood.

It's clear that any computational framework for long term associative memory must account not only for synaptic plasticity, but also a structural plasticity allowing for the temporally dynamic, continuous integration of new experiences into the existing memory landscape. Traditionally, Hopfield networks, utilizing Hebbian plasticity and attractor dynamics, have been the standard accepted model in terms of biological plausibility in neuroscience. However, they suffer from low storage capacity as a result of their direct method of encoding memories as attractor points. New systems in Machine Learning, Diffusion Models, seem to offer potential solutions to a more viable biological model. These models are capable of recovering lost data, but the beauty is that they are not engineered to do so - instead they're engineered to mimic the probability distribution function of a vast dataset in order to generate plausible members of the dataset. Given that we know our own associative memories to not be perfect content addressable memory systems, but rather sometimes fabricators or embellishers of detail, it's worth considering what diffusion models might have to contribute to theoretical neuroscience.

In this paper, we will first draw some mathematical analogies between Associative Memory and Diffusion Models, namely that they navigate analogous landscapes. Next, we will move to assessing the capabilities and limitations of the differing computational models in replicating the physical processes of human memory. Then, flipping the narrative, we will consider first the physical processes, then their corresponding models. Here, we'll delve into the dynamic role of cAMP Response Element-Binding protein in memory processes and how it influences the plausibility of these computational analogies.

2 Structural Analogies

2.1 Associative Memories

Associative Memories (AMs) are pattern-restoring models capable of recovering information from partial data. They are most often energy-based models, meaning their dynamics occur at the level of system energy, guided by an energy (*Lyapunov*) function. In this paper, we'll focus mostly on

Hopfield Networks (HNs), pioneered by John Hopfield in 1982 using principles from the Ising model of statistical mechanics. This network evolves dynamically in an energy landscape towards stable states of energetic minima that correspond to "memories". These memories, or more accurately, patterns, are stored in the values of the weights between neurons in the network. The energy function itself, then, is engineered such that from any point in state space, we are guaranteed to converge to a stable state corresponding to some pattern stored in the network. The energy function of a state x in Hopfield's original formulation of HNs is given by

$$E(x) = -\frac{1}{2}\sum_{i,j}^{N} w_{ij}x_i x_j + \sum_{i}^{N}\theta_i x_i \tag{1}$$

where w_{ij} is the connection weight between two neurons, x_i is the state of unit i and θ_i is the threshold of unit i [7].

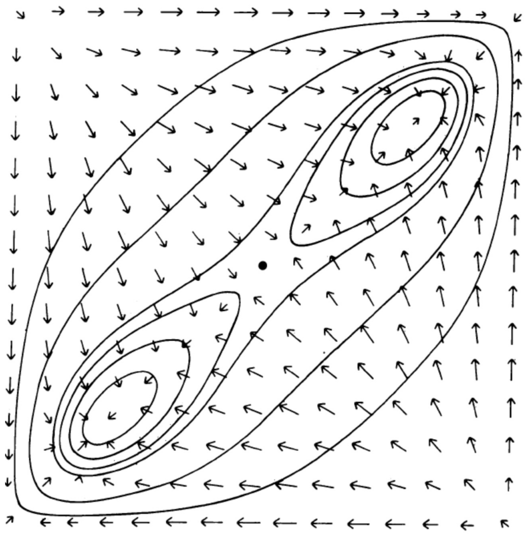

Figure 1: Energy contour map for a two-neuron, two-stable-state system (Hopfield, 1984)

Given an initial position in state space, $\mathbf{x}_t$ at time $t = 0$, the goal is to minimise the energy function by descending the energy gradient until we reach a fixed point. So we can create a position update function of the following form.

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{dt}{\tau}\nabla_x E(\mathbf{x}_t) \tag{2}$$

Where τ is a time constant determining the speed of evolution. Thus we have the following differential equation governing the dynamics of an HN:

$$\tau\frac{d\mathbf{x}}{dt} = -\nabla_x E(\mathbf{x}) \tag{3}$$

Now, HNs use Hebbian learning (as evidenced by the $w_{ij}x_ix_j$ term in (1)), which is precisely what makes them a form of *content-addressable memory* (CAM). The weights of the network are stored in a weight matrix ($\mathbf{W}$), which is computed using Hebbian learning principles. Consider a pattern represented as a binary vector $\mathbf{v} = [v_1, v_2, ...v_N]$ where v_i is the state of the i^{th} neuron (+1 or -1 for firing or not firing), then the outer product of this vector with itself is a matrix given by $\mathbf{v}\mathbf{v}^\top$. Each element, w_{ij} of this matrix represents the weight between neurons i and j for a specific pattern. For multiple patterns stored, we get the following weight matrix:

$$\mathbf{W} = \sum_{p=1}^{k} \mathbf{v}_p \mathbf{v}_p^\top \tag{4}$$

where $\mathbf{v}_p$ is the p^{th} pattern vector. Writing this out explicitly makes it clear that HNs directly encode patterns into the weight matrix above, representing the synaptic weights of the network. This also makes it clear that HNs can very easily suffer from ”over-saturation” - too many patterns and too few neurons will cause performance to degrade drastically. Traditional HNs can store $\sim 0.138N$ patterns where N is the number of neurons, which gives 138 patterns for every 1000 neurons [4]. That said, Modern Hopfield Networks use an updated energy function that accounts for continuous variables, and can store a number of patterns that is roughly on the order of the number of neurons. In contrast, the human brain is capable of storing and retrieving an enormous number of memories of varying form and detail, which might undermine our confidence in HNs as biological models of memory.

2.2 Diffusion Models

Diffusion Models (DMs) are a class of generative models within machine learning that aim to create data by reversing a diffusion process. In this paper, we'll focus mostly on the Denoising Diffusion Probabilistic Model (DDPM) identified by Jonathan Ho et al in their 2020 paper [5]. DDPMs first inject small amounts of Gaussian noise to an image and learn the parameters that best allow the model to predict the original data from the noisy data at each timestep. This reverse process effectively denoises the data, resulting in the generation of a new sample that resembles the training data. Similarly to AMs, DMs attempt to find stationary points in a landscape, but rather than finding minima in an energy function, they search for maxima in a log-probability function.

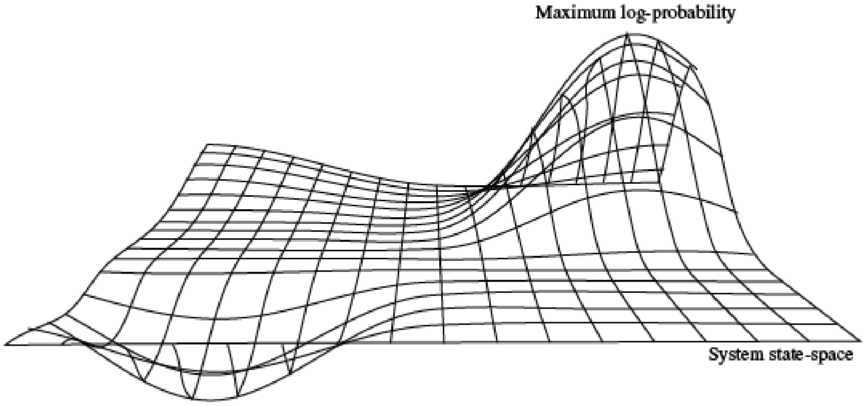

Figure 2: A log-probability landscape for a DM

Diffusion models, like AMs learn by matching their ”landscape”, or *score-function* ($\mathbf{F}(\mathbf{x})$), to the data received. However, instead of encoding specific neural patterns directly into the synaptic weights of the system, they model their score-function to match the probability distribution of the dataset they are trained on. In other words, they create their own internal probability density function which they train to match that of the data so that when it comes time to generate an image, the peaks of their score-function will correspond to possible images that are **most** likely to be found in the actual dataset. Both AMs and DMs use principles from statistical mechanics to govern their dynamics, so thanks to the Boltzmann Distribution, we can relate the score-function in DMs to the energy function in AMs:

$$p(\mathbf{x}) = \frac{1}{Z}e^{-E(x)} \tag{5}$$

Where Z is a normalisation constant that ensures $\int p(\mathbf{x})d\mathbf{x} = 1$. With some rearranging, we can see the relationship between minimising energy and maximising log-probability.

$$E(\mathbf{x}) = -\log(Z) - \log(p(\mathbf{x})) \tag{6}$$

$$E(\mathbf{x}) \propto -\log(p(\mathbf{x})) \tag{7}$$

Minimising the energy function and maximising the log-probability seem to be mathematically analogous. Further, given an initial position in state space, $\mathbf{x}_t$ at time $t = 0$, the goal is to maximise the log-probability function by ascending the log-probability gradient until we reach a maximum point. We can create a position update function using the score-function $\mathbf{F}(\mathbf{x}) = \nabla_{\mathbf{x}} \log p(\mathbf{x}) = -\nabla_{\mathbf{x}} E(\mathbf{x})$.

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \alpha\mathbf{F}(\mathbf{x}) \tag{8}$$

Where α represents a step size in ascending the gradient. [11] use $\alpha = \dfrac{\sigma^2(t)}{2}$ for diffusion co-efficient $\sigma(t)$ which improves efficacy by allowing larger amounts of noise to be removed the further you are from the original probability distribution. We can then define $\tau(t) \triangleq \dfrac{2}{\sigma^2(t)}$ such that the differential equation of a Diffusion Model bears exact resemblance, up to a time-step modulator, to that of an Associative Memory system (3).

$$\tau(t)\frac{d\mathbf{x}}{dt} = \nabla_x p(\mathbf{x}) \tag{9}$$

$$= -\nabla_x E(\mathbf{x}) \tag{10}$$

2.3 Implications

Given the striking similarities between the underlying mathematics governing Diffusion Models and Associative Memory, it's tempting to say that they are simply two different ways of ”doing the same thing”, but that would be misplaced. It is rather the differences between these two systems that are truly enlightening.

AMs are retrieval systems, while DMs are generative. Therefore, as we have seen, Hopfield Networks' biggest pitfall is their low capacity, which in constrast to DMs, we can attribute to the fact that Hopfield Networks have no *generative capacity.* They either converge upon a fixed point or remain very near to the ”superposition-of-all-training-data” state. This can be seen below where we have trained a modern hopfield network [10] to recognise 10 patterns from the MNIST dataset and then 1000 patterns.

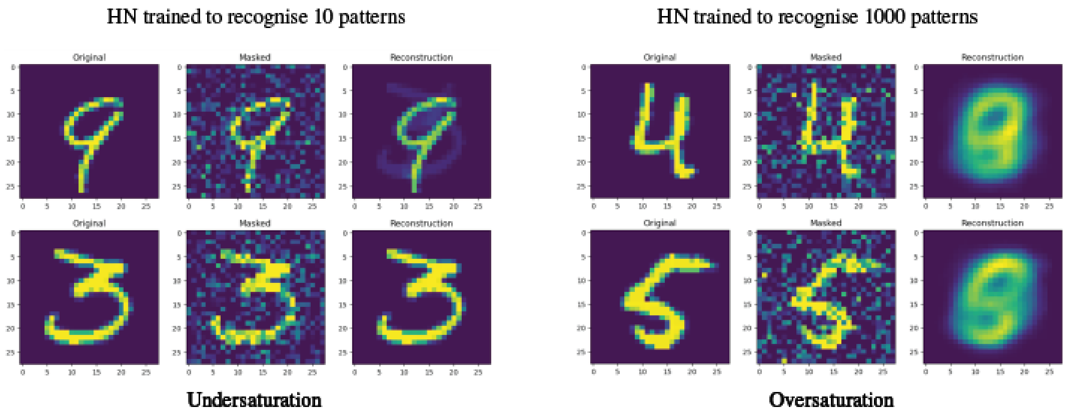

Figure 3: HN saturation becomes clear on the right hand side with the superposition of states being output regardless of the input

DMs are capable of overcoming this issue to the extent that they encode their data not directly, but by way of distributed synaptic patterns. This is a feature that arises from the fact that the energy function is not tractable for DMs [6]. Instead, they navigate a log-probability landscape where the system learns a *propensity map*, but AMs use the energy landscape and learn its contours. Although DMs' probability landscapes contain stationary points, they don't discretise memory into such points, but are instead free to map dispersed memory structures of arbitrary dimensionality [1][2]. DMs do this via dimensionality reduction - they are attracted towards a low-dimensional manifold that is found within a high-dimensional space containing all the desired high-dimensional data (like images or text). Although many AMs including Hopfield Networks perform dimensionality reduction, the use of the energy (*lyapunov*) function hinders this by restricting the system's end states to predetermined attractor points.

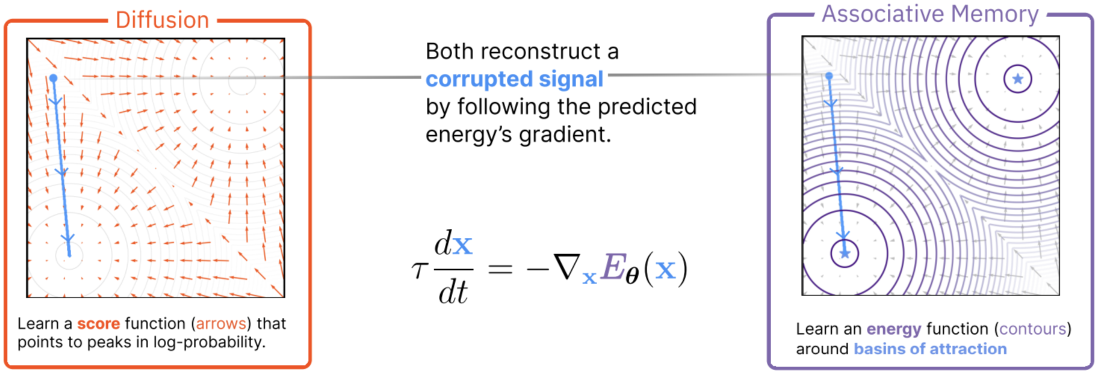
Diffusion
Learn a score function (arrows) that points to peaks in log-probability.
Both reconstruct a corrupted signal by following the predicted energy's gradient.
$\tau \frac{d\mathbf{x}}{dt} = -\nabla_{\mathbf{x}} \boldsymbol{E}_{\boldsymbol{\theta}}(\mathbf{x})$
Associative Memory
Learn an energy function (contours) around basins of attraction

Figure 4: Comparing the learned features of DM and AM landscapes [6]

It is worth asking, then, whether diffusion models might, with some adjustment, make better candidates for a biological model of memory recall.

3 Models as Physical Systems

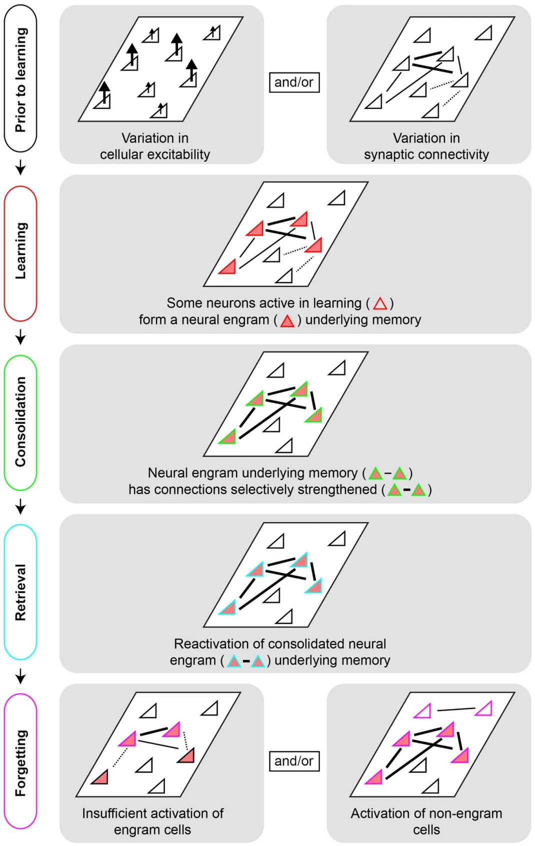
Prior to learning
Variation in cellular excitability
and/or
Variation in synaptic connectivity
Learning
Some neurons active in learning (△)
form a neural engram (▲) underlying memory
Consolidation
Neural engram underlying memory (▲–▲)
has connections selectively strengthened (▲–▲)
Retrieval
Reactivation of consolidated neural
engram (▲–▲) underlying memory
Forgetting
Insufficient activation of engram cells
and/or
Activation of non-engram cells

Figure 5: Depicts the transformation of engrams in each stage of memory, as described by Guskjolen and Cembrowsk [3]

The term 'engram' refers to the physical manifestation of a memory, i.e. a change within the nervous system that occurs in response to learned experience. During memory formation, neurons are recruited into an engram with varying probabilities largely attributed to their intrinsic neuronal excitability [3]. This intrinsic excitability refers to the neuron's propensity to fire an action potential in response to stimulus. It is notably temporally dynamic, modulated by various factors including both synaptic input and interstellar signaling pathways [9]. The crucial point to our discussion is just that: the idea that the selection of neurons into an engram is based on both synaptic and non-synaptic forms of plasticity. Static, discrete models like Hopfield networks only account for the prior. In this light, DDPMs look far more biologically plausible, accommodating dynamic non-synaptic processes and ensemble-based memory representations. A successful model of long-term associative memory must be able to reflect synaptic changes as well as the internal molecular states and processes that govern these changes.

3.1 Synaptic Plasticity and Network Dynamics: The Implications of CREB

CREB (cAMP Response Element-Binding protein) is a key modulator of neuronal intrinsic excitability. Heightened levels have been shown to increase the likelihood that a given neuron will undergo long term potentiation (the primary process modeled in Hebbian learning). However, it is crucial to note that this modulation is twofold at least. Beyond increasing individual neuronal excitability, CREB also orchestrates complex patterns of connectivity throughout neural networks [12]. It is able to achieve this through transcriptional control of genes vital to synaptic formation and plasticity [8]. Such differential gene regulation in response to varied stimuli facilitated by CREB conflicts with the biological plausibility of static network models like those posited by Hopfield Networks.

To reiterate in relation to prior discussion, Hopfield Networks model memory using predetermined attractor points within a static energy landscape, implying a fixed, unchanging structure of synaptic connections. However, CREB-induced changes in synaptic plasticity and connectivity suggest a more fluid, adaptive neural network. This adaptability is poorly represented in the static attractor points of Hopfield Networks, where the potential for continuous, dynamic reconfiguration of synaptic connections - a hallmark of biological neural networks influenced by CREB - is not adequately captured.

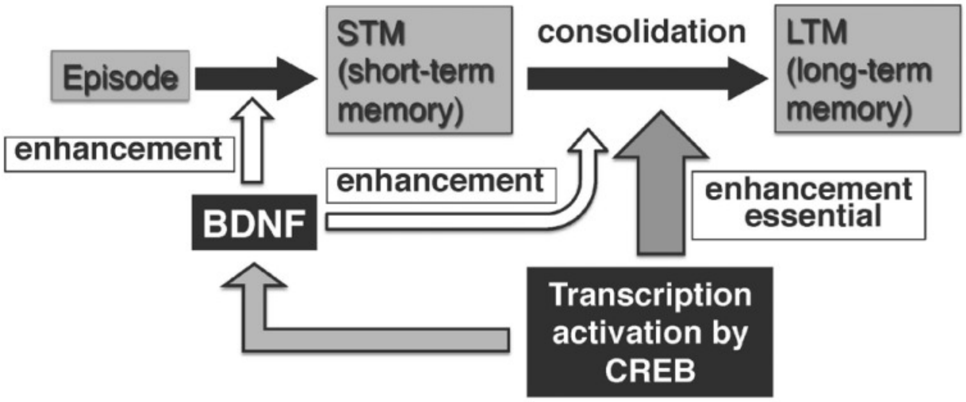

Figure 6: CREB regulates memory formation by controlling gene expression critical for memory consolidation. Its transcriptional activation strength determines memory robustness. CREB also indirectly influences short-term memory through the regulation of its target gene BDNF, enhancing memory consolidation via a positive feedback loop. [8]

3.2 Reconceptualizing Memory Processes in Light of CREB Dynamics

The dynamic regulation of synaptic connectivity by CREB challenges the notion of memory encoding as a static process suggested by Hopfield Networks. Instead, it aligns more closely with the principles of DDPMs, which navigate a log-probability landscape. As previously discussed, this landscape represents a learning model where the system acquires a propensity map reflecting the dispersed, dynamic nature of memory structures. In biological terms, this might be likened to the way CREB dynamically modulates synaptic connections and neuronal excitability, leading to a more dispersed and adaptable encoding of memories. Fluctuations in CREB activity during different stages of memory formation – from encoding to retrieval, correlate with the evolving nature of memory itself [3]. This aspect would be modeled as a time-varying influence on the memory landscape. The memory trace evolves rather than being confined to predetermined points: a crucial function in the physical process which is accommodated in DDPM models, but not in Hopfields.

3.3 DDPMs: Physical Mechanisms and Computational Analogues

Modulating synaptic connections and neuronal excitability offers a biological parallel to the concept of a propensity maps in DDPMs. These represent a probabilistic distribution of data points in a high-dimensional space which guides the model's learning process towards areas of higher probability. Analogously, CREB-induced changes in synaptic strength and excitability alter the 'landscape' of neuronal connections, increasing the likelihood of certain neuronal pathways being activated in response to stimuli. This CREB-mediated modulation can be seen as 'mapping' out a probability landscape within the neural network, where certain pathways become more likely to contribute to memory encoding based on their enhanced excitability and synaptic strength. In this context, the propensity map in DDPMs may be taken as akin to the neural network's evolving connectivity pattern under CREB's influence. Just as DDPMs navigate through a landscape based on learned probabilities, CREB guides the neural network through a landscape of synaptic modifications and excitability changes. This dynamic process results in a more dispersed and adaptable encoding of memories, moving away from the static, predetermined patterns of Hopfield Networks to a model that mirrors the probabilistic and evolving nature of biological memory processes.

4 Conclusion

Exploration into the structural analogies between Hopfield Networks and Denoising Diffusion Probabilistic Models reveals a compelling narrative about the nature of memory encoding in both biological and computational realms. Our analysis underscores the limitations of Hopfield Networks in capturing the dynamic, evolving nature of biological memory processes, particularly in light of the role of CREB in synaptic plasticity and neuronal network dynamics. In contrast, DDPMs, with their capacity for modeling continuous and probabilistic processes, offer a more accurate and biologically plausible framework. This shift towards DDPMs not only enhances our understanding of memory processes but also opens new frontiers in the development of computational models that can emulate the complex and adaptive nature of biological systems. These insights encourage further exploration around the integration of neurobiological mechanisms and computational models, suggesting that both domains might benefit from, as well as themselves catalyze, advances in their respective other.

References

[1] Ambrogioni, L. (2023a). In search of dispersed memories: Generative diffusion models are associative memory networks.

[2] Ambrogioni, L. (2023b). The statistical thermodynamics of generative diffusion models.

[3] Guskjolen, A. and Cembrowski, M. S. (2023). Engram neurons: Encoding, consolidation, retrieval, and forgetting of memory. *Molecular Psychiatry*, 28(8):3207–3219.

[4] Hertz, J. A., Krogh, A. S., and Palmer, R. G. (2019). *Introduction to the theory of neural computation*. CRC Press, London, England.

[5] Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models.

[6] Hoover, B., Strobelt, H., Krotov, D., Hoffman, J., Kira, Z., and Chau, D. H. (2023). Memory in plain sight: A survey of the uncanny resemblances between diffusion models and associative memories.

[7] Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U. S. A.*, 79(8):2554–2558.

[8] Kida, S. (2012). A functional role for creb as a positive regulator of memory formation and ltp. *Experimental Neurobiology*, 21:136 – 140.

[9] Mozzachiodi, R. and Byrne, J. H. (2010). More than synaptic plasticity: role of nonsynaptic plasticity in learning and memory. *Trends in Neurosciences*, 33(1):17–26.

[10] Pulfer, B. (2023). Personal short implementations of machine learning papers.

[11] Song, Y. and Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[12] Yiu, A. P., Mercaldo, V., Yan, C., Richards, B., Rashid, A. J., Hsiang, H.-L. L., Pressey, J., Mahadevan, V., Tran, M. M., Kushner, S. A., Woodin, M. A., Frankland, P. W., and Josselyn, S. A. (2014). Neurons are recruited to a memory trace based on relative neuronal excitability immediately before training. *Neuron*, 83(3):722–735.