ON LOW-DIMENSIONAL REPRESENTATIONS AND ASSOCIATIVE MEMORY ENERGY FUNCTIONS

David Lipshutz
Department of Neuroscience, Baylor College of Medicine
Neuroengineering Initiative, Rice University
david.lipshutz@bcm.edu

ABSTRACT

Associative memory (AM) models encode memories as local minima of energy functions and are generally analyzed in terms of their storage capacity. We provide a complementary perspective: minimizing the AM energy function is related to minimizing the dimensionality of the model's representation. Starting from a representational objective that balances input reconstruction with maximization of the inverse participation ratio (IPR), we relate the objective to minimization of the classical AM energy function. We further generalize the relation to dense AM energy functions using kernel methods.

1 INTRODUCTION

Associative memory (AM) models are influential models of memory storage and retrieval (Amari, 1972; Little, 1974; Hopfield, 1982; Krotov, 2023). In these models, memories are stored as isolated minima of an energy function and can be retrieved from partial information by descending the energy function. These models are trained on uncorrupted storage patterns and are analyzed through the lens of storage capacity; that is, the number of distinct minima of the energy function.

We revisit the AM energy function from a complementary perspective. We show that minimizing the AM energy function is related to a objective for reducing the dimensionality of the model's representation, which may be useful for learning structure from noisy observations. Starting from a representational objective that balances input reconstruction and maximization of the inverse participation ratio (IPR), we derive a relation to minimization of the classical AM energy function (section 3). We further generalize the objective using the kernel method to establish a relation with dense AM energy functions (section 4).

Our results demonstrate that the AM energy function can serve as a principled objective for unsupervised learning of low-dimensional latent structures from noisy observations, offering an expanded normative interpretation for AM models: when the weights are learned from relatively few clean and uncorrelated storage patterns, the models learn isolated zero-dimensional point attractors; however, when weights are learned from many noisy correlated patterns, the model learns low-dimensional continuous attractors. This complements recent works establishing connections between AM energy functions and denoising-based generative diffusion models across different data regimes (Hoover et al., 2023; Raya & Ambrogioni, 2023; Ambrogioni, 2024; Pham et al., 2024; Smart et al., 2025). However, in contrast to supervised denoising models, we assume the model does not have access to clean signals, which is relevant for biological systems that learn from noisy observations.

2 CLASSICAL CONTINUOUS AM NETWORK

We first recall the classical continuous AM network (Hopfield, 1984), which is a recurrent network with N neurons whose states are described by an N-dimensional vector $\mathbf{y} = (y_1, \dots, y_N)$. For the case that the activation function is the identity, the dynamics of neuron i are described by

$$\tau \frac{dy_i}{dt} = \sum_{j \neq i} w_{ij} y_j - c_i y_i + x_i, \tag{1}$$

where $\tau > 0$ is a time constant, w_{ij} is the strength of connection between neuron j and neuron i for $j \neq i$, x_i represents a constant input to neuron i, and c_i represents the leak for neuron i. When the matrix $\mathbf{W} = (w_{ij})$ is symmetric, the network's dynamics are naturally interpreted as gradient flow with respect to y_i of the classical quadratic AM energy function (see Hopfield, 1984, equation 7)

$$E_{\text{classical}}(\mathbf{W}, \mathbf{c}, \mathbf{x}, \mathbf{y}) = -\frac{1}{2}\mathbf{y}^\top \mathbf{W}\mathbf{y} + \frac{1}{2}\mathbf{y}^\top \text{diag}(\mathbf{c})\mathbf{y} - \langle \mathbf{x}, \mathbf{y} \rangle. \tag{2}$$

where $\mathbf{x} = (x_i)$ and $\mathbf{y} = (y_i)$ and the weights satisfy the Hebbian rule $w_{ij} = \sum_\mu \xi_i^\mu \xi_j^\mu$ on a set of uncorrupted storage patterns $\{\boldsymbol{\xi}^\mu\}$. In the original formulation, the diagonal entries w_{ii} are assumed to be zero; however, this can be relaxed without loss of generality since the diagonal entries w_{ii} can be absorbed into c_i.

3 Unsupervised learning from noisy observations

In this section, we show that the classical AM energy function (equation 2) arises when minimizing a representational objective for learning low-dimensional structure from noisy data. Similar to the previous section, we consider a recurrent neural network with N neurons whose states are described by an N-dimensional vector $\mathbf{y}$ taking values in $\mathbb{R}^N$, as shown in Figure 1. The network receives inputs $\mathbf{x}$ that are sampled from a distribution $p(\mathbf{x})$ and generates responses $\mathbf{y}$. We assume that the responses $\mathbf{y}$ are a deterministic function of $\mathbf{x}$—that is, $\mathbf{y} = f(\mathbf{x})$ for some function $f(\cdot)$ that is optimized over. For mathematical convenience, we assume $\mathbf{x}$ and $\mathbf{y}$ are centered.

3.1 Representational objective

To state our objective, we first need to quantify the (inverse) dimensionality of a representation. We use the IPR (Abbott et al., 2010; Gao et al., 2017), which is defined by

$$D_{\text{IPR}} := \frac{\text{Tr}(\mathbf{C}^2)}{\text{Tr}(\mathbf{C})^2} = \frac{\sum_i \lambda_i^2}{(\sum_i \lambda_i)^2}, \tag{3}$$

where $\mathbf{C} = \text{Cov}(\mathbf{y})$ is the $N \times N$ covariance matrix of $\mathbf{y}$ and $\{\lambda_i\}$ are the eigenvalues of $\mathbf{C}$. The IPR is a continuous and homogeneous function of $\mathbf{C}$ that takes values between $1/N$ and 1. It is an inverse measure of dimensionality in the sense that the IPR is equal to $1/d$ when $\mathbf{C}$ has d eigenvalues equal to one and the rest are zero. Importantly, a larger IPR corresponds to a lower-dimensional response.

We propose the following representational objective (or loss) that balances input reconstruction with maximization of the IPR:

$$L_{\text{classical}} = \mathbb{E}\left[\|\mathbf{x} - \mathbf{y}\|^2\right] - \frac{1}{2} D_{\text{IPR}} \tag{4}$$

subject to a constraint of the form $\text{Tr}(\mathbf{C}) = 1$, which is equivalent to a constraint on the average (centered) response magnitude: $\mathbb{E}[\|\mathbf{y}\|^2] = 1$. Expectations are with respect to the distribution $p(\mathbf{x})$ and $\mathbf{y}$ is implicitly assumed to be a deterministic function of $\mathbf{x}$. Minimizing the first term in $L_{\text{classical}}$ minimizes the average Euclidean distortion between the inputs $\mathbf{x}$ and responses $\mathbf{y}$ while minimizing the second term encourages low-dimensional responses $\mathbf{y}$.

3.2 Relationship to the classical AM energy function

We now show that under the constraint $\text{Tr}(\mathbf{C}) = 1$, the objective function L can be rewritten in terms of the AM energy function $E_{\text{classical}}$. To see this, we rewrite $\text{Tr}(\mathbf{C}^2)$ using a Legendre transform (Pehlevan et al., 2018; Duong et al., 2026):

$$\text{Tr}(\mathbf{C}^2) = \max_{\mathbf{W}} \left\{ 2\,\text{Tr}(\mathbf{C}\mathbf{W}) - \|\mathbf{W}\|_F^2 \right\} = \max_{\mathbf{W}} \left\{ \mathbb{E}\left[2\mathbf{y}^\top \mathbf{W}\mathbf{y}\right] - \|\mathbf{W}\|_F^2 \right\}.$$

The first equality can be verified via differentiating and noting that the maximum is attained at $\mathbf{W} = \mathbf{C}$. The second equality uses the relation $\mathbf{C} = \mathbb{E}[\mathbf{y}\mathbf{y}^\top]$ (when $\mathbf{y}$ is centered) and the cyclic property of the trace operator. Therefore, under the constraint $\text{Tr}(\mathbf{C}) = 1$, we can rewrite the IPR as follows:

$$D_{\text{IPR}} = \max_{\mathbf{W}} \left\{ \mathbb{E}\left[2\mathbf{y}^\top \mathbf{W}\mathbf{y}\right] - \|\mathbf{W}\|_F^2 \right\}.$$

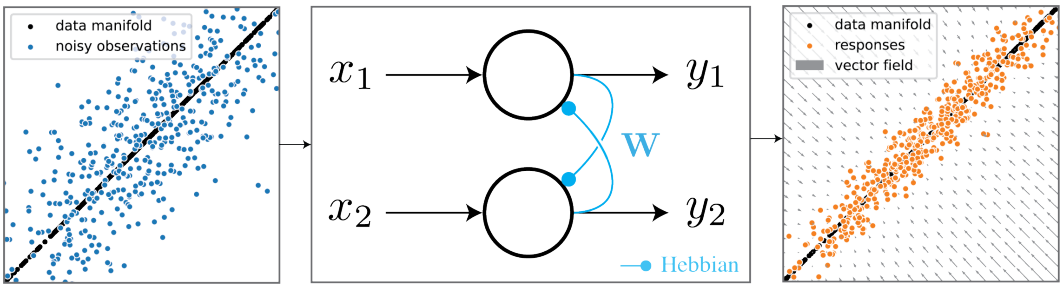


Figure 1: AM network with $N = 2$ neurons learns a one-dimensional line attractor from noisy data. **Left:** the blue dots are noisy observations $\mathbf{x}$ of the data manifold (black line). Specifically, data points $\mathbf{s}$ are sampled as $\mathbf{s} \sim \mathcal{N}(\mathbf{0}, \mathbf{E})$, where $\mathbf{E}$ is the matrix of ones, and $\mathbf{x} = \mathbf{s} + 0.5\mathbf{n}$, where $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. **Center:** AM network with Hebbian recurrent synapses $\mathbf{W}$. **Right:** the vector field $f(\mathbf{y}) = (\mathbf{W} - \text{diag}(\mathbf{c}))\mathbf{y}$ shows the learned neural dynamics and the orange dots correspond to the low-dimensional network responses $\mathbf{y}$. The axes in the left and right panels are both $[-2, 2] \times [-2, 2]$.

Substituting back into equation 4, expanding $\|\mathbf{x} - \mathbf{y}\|^2$, and using the fact that $\mathbb{E}[\|\mathbf{y}\|^2] = 1$, we get

$$L_{\text{classical}} = \min_{\mathbf{W}} \mathbb{E}\left[-2\langle \mathbf{x}, \mathbf{y}\rangle - \mathbf{y}^\top \mathbf{W}\mathbf{y} + \frac{1}{2}\|\mathbf{W}\|_F^2\right] + \mathbb{E}[\|\mathbf{x}\|^2] + 1.$$

To ensure $\text{Tr}(\mathbf{C}) = 1$, we could directly enforce the equality using a Lagrange multiplier. However, to obtain a more biologically plausible network implementation, we instead enforce the constraint $\text{Var}(y_i) = 1/N$ for $i = 1, \ldots, N$, which implies $\text{Tr}(\mathbf{C}) = 1$. We introduce Lagrange multipliers $\mathbf{c} = (c_1, \ldots, c_N)$ and then minimize with respect to $\mathbf{y}$:

$$\min_{\mathbf{y}=f(\mathbf{x})} \max_{\mathbf{c}} \min_{\mathbf{W}} \mathbb{E}\left[-2\langle \mathbf{x}, \mathbf{y}\rangle - \mathbf{y}^\top \mathbf{W}\mathbf{y} + \frac{1}{2}\|\mathbf{W}\|_F^2 + \sum_{i=1}^{n} \frac{c_i}{2}\left(y_i^2 - 1/N\right)\right],$$

where we have also dropped constant terms. Finally, we interchange the order of optimization and express the nested optimization problem in terms of the classical AM energy function:

$$\min_{\mathbf{W}} \max_{\mathbf{c}} \left\{2\mathbb{E}\left[\min_{\mathbf{y}} E_{\text{classical}}(\mathbf{W}, \mathbf{c}, \mathbf{x}, \mathbf{y})\right] + \frac{1}{2}\|\mathbf{W}\|_F^2 - \frac{1}{2N}\langle \mathbf{c}, \mathbf{1}\rangle\right\},$$

where the optimal function f is implicitly defined by $f(\mathbf{x}) = \arg\min_{\mathbf{y}} E_{\text{classical}}(\mathbf{W}, \mathbf{c}, \mathbf{x}, \mathbf{y})$. However, in contrast to the classical Hopfield network, the recurrent weights will be learned from noisy observations rather than clean storage patterns.

3.3 Optimization in a recurrent network with Hebbian plasticity

We can optimize the objective in a recurrent network with N neurons with fast neural dynamics and slow updates to the synapses and leak, Figure 1. We assume that at each time step, the network receives an input $\mathbf{x}$ and runs the gradient descent neural dynamics that coincide with equation 1, which equilibrate at $\mathbf{y} = (\text{diag}(\mathbf{c}) - \mathbf{W})^{-1}\mathbf{x}$. We then update the synapses by descending the gradient of $E_{\text{classical}}$ with respect to $\mathbf{W}$ and we update the leaks by ascending the gradient of $E_{\text{classical}}$ with respect to $\mathbf{c}$. This results in the update rules:

$$\Delta w_{ij} = \eta_w(y_i y_j - w_{ij}), \qquad \Delta c_i = \eta_r\left(y_i^2 - 1/N\right),$$

for $i, j = 1, \ldots, N$, where $\eta_w, \eta_r > 0$ are the learning rates for the synaptic weights and leak terms. The synaptic weight updates includes Hebbian and homeostatic terms and the leak update for a neuron is local in the sense that it only depends on that neuron's activity. Figure 1 illustrates the optimized circuit that learns a one-dimensional line attractor.

4 Kernel IPR and dense AMs

This section generalizes the representational objective in Section 3.1 to establish a correspondence with dense AM energy functions. The key insight is that while classical AMs minimize linear dimensionality, dense AMs minimize dimensionality in a high-dimensional nonlinear feature space.

4.1 REPRESENTATIONAL OBJECTIVE USING KERNEL METHODS

To establish a correspondence with dense AMs, we generalize the IPR using the kernel method. Given a symmetric positive definite kernel $K(\mathbf{x}, \mathbf{y})$, the kernel IPR for $\mathbf{y} \sim p(\mathbf{y})$ is defined by

$$D_{\text{IPR},K} = \frac{\mathbb{E}_{(\mathbf{y},\mathbf{y}')\sim p(\mathbf{y})\times p(\mathbf{y}')}\left[K^2(\mathbf{y},\mathbf{y}')\right]}{\mathbb{E}_{\mathbf{y}\sim p(\mathbf{y})}\left[K(\mathbf{y},\mathbf{y})\right]^2}, \tag{5}$$

which reduces to equation 3 when $K(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$. We propose the following generalization of the representational objective in equation 4:

$$L_{\text{dense}} = -2\mathbb{E}[K(\mathbf{x}, \mathbf{y})] - \frac{1}{2} D_{\text{IPR},K}, \tag{6}$$

subject to a constraint $\mathbb{E}[K(\mathbf{y}, \mathbf{y})] = 1$. The first term aligns $\mathbf{y}$ with the inputs $\mathbf{x}$ while the second term maximizes the kernel IPR. Note that if we replace $-2K(\mathbf{x}, \mathbf{y})$ with $m(\mathbf{x}, \mathbf{y}) := K(\mathbf{x}, \mathbf{x}) - 2K(\mathbf{x}, \mathbf{y}) + K(\mathbf{y}, \mathbf{y})$ in the previous display, then the loss L_{dense} corresponds exactly to the loss in equation 4 when $K(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$.

4.2 RELATION TO DENSE AM ENERGY FUNCTIONS

Since K is a symmetric positive definite kernel, there exists a Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and a feature map $\Phi : \mathbb{R}^N \mapsto \mathcal{H}$ such that $K(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle_{\mathcal{H}}$ (Aronszajn, 1950). As before, using the constraint $\mathbb{E}_{\mathbf{y}\sim p(\mathbf{y})}[K(\mathbf{y}, \mathbf{y})] = 1$ and a Legendre transform, we can write

$$D_{\text{IPR},K} = \max_{A} \left\{ 2\langle A\Phi(\mathbf{y}), \Phi(\mathbf{y}) \rangle_{\mathcal{H}} - \|A\|_{\text{HS}} \right\},$$

where $\| \cdot \|_{\text{HS}}$ is the Hilbert-Schmidt norm on the space of bounded linear operators $\mathcal{L}(\mathcal{H})$ and the maximization is over $A \in \mathcal{L}(\mathcal{H})$. See Appendix A for a detailed derivation. Introducing a Lagrange multiplier g to enforce the constraint $\mathbb{E}[K(\mathbf{y}, \mathbf{y})] = 1$, minimizing with respect to $\mathbf{y}$, and interchanging the order of optimization, we get

$$\min_{A} \max_{g} \left\{ \mathbb{E}\left[\min_{\mathbf{y}} E_{\text{dense}}(A, g, \mathbf{x}, \mathbf{y}) \right] - g + \frac{1}{2}\|A\|_{\text{HS}}^2 \right\}, \tag{7}$$

where we've defined the dense AM energy function as

$$E_{\text{dense}}(A, g, \mathbf{x}, \mathbf{y}) := -2\langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle_{\mathcal{H}} - \langle A\Phi(\mathbf{y}), \Phi(\mathbf{y}) \rangle_{\mathcal{H}} + g\langle \Phi(\mathbf{y}), \Phi(\mathbf{y}) \rangle_{\mathcal{H}}.$$

To more explicitly see the relationship to dense AM energy functions, suppose the linear operator A is the covariance operator for a set of storage patterns $\{\boldsymbol{\xi}^\mu\}$; that is, $A = \sum_\mu \Phi(\boldsymbol{\xi}^\mu) \otimes \Phi(\boldsymbol{\xi}^\mu)$. Then we can express the energy function in terms of the kernel K as follows:

$$E_{\text{dense}}(A, g, \mathbf{x}, \mathbf{y}) = -2K(\mathbf{x}, \mathbf{y}) - \sum_{\mu} K^2(\boldsymbol{\xi}^\mu, \mathbf{y}) + gK(\mathbf{y}, \mathbf{y}). \tag{8}$$

From equation 8, we see that when $K^2(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle^n$ or $K^2(\mathbf{x}, \mathbf{y}) = \exp(\beta \langle \mathbf{x}, \mathbf{y} \rangle)$, the second term respectively corresponds to polynomial energy function proposed by Krotov & Hopfield (2016) or the exponential energy function proposed by Demircigil et al. (2017). If one additionally takes logarithms of each term on the right-hand-side of equation 8 separately, then the energy function closely resembles the one proposed by Ramsauer et al. (2021). An important distinction is that we assume the model is trained on noisy observations $\mathbf{x} \sim p(\mathbf{x})$, not clean storage patterns $\{\boldsymbol{\xi}^\mu\}$.

4.3 OPTIMIZATION ALGORITHM

To optimize the nested minimax optimization problem in equation 7, we use a two-timescale stochastic gradient-based algorithm that parallels the one in section 3.3. At each iteration, we first sample an observation $\mathbf{x}$ and minimize with respect to $\mathbf{y}$:

$$\frac{d\mathbf{y}}{dt} = -\frac{1}{2}\nabla E_{\text{dense}}(A, g, \mathbf{x}, \mathbf{y}) = J_\Phi(\mathbf{y})^\top \left(\Phi(\mathbf{x}) + A\Phi(\mathbf{y}) - g\Phi(\mathbf{y}) \right), \tag{9}$$

where $J_\Phi(\mathbf{y})$ is the Jacobian of $\Phi(\mathbf{y})$ with respect to $\mathbf{y}$. Then, once $\mathbf{y}$ converges, we update A and g as follows:

$$\Delta A = \eta_A(\Phi(\mathbf{y}) \otimes \Phi(\mathbf{y}) - A), \qquad \Delta g = \eta_g(K(\mathbf{y}, \mathbf{y}) - 1). \tag{10}$$

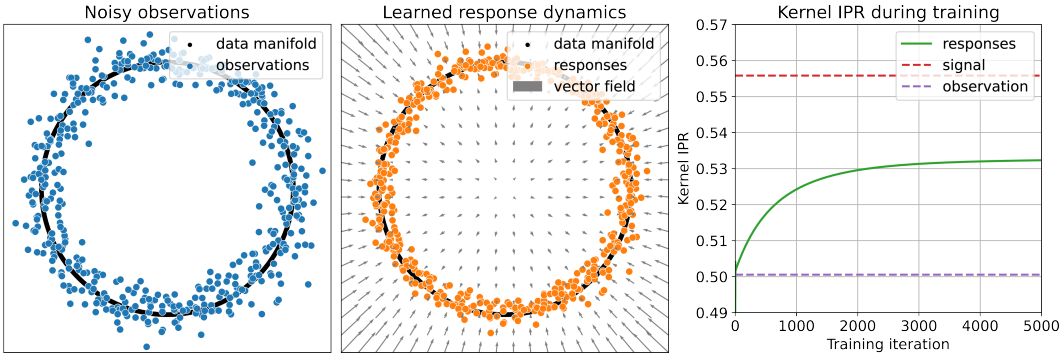


Figure 2: Dense AM model with quadratic kernel learns a ring attractor from noisy data. **Left:** the blue dots are noisy observations $\mathbf{x}$ of the unit circle. Specifically, data points $\mathbf{s}$ are sampled uniformly from the unit circle (black circle) and $\mathbf{x} = \mathbf{s} + 0.1\mathbf{n}$ where $\mathbf{n}$ is isotropic Gaussian noise. **Center:** the orange dots are learned responses of a dense AM model with quadratic kernel trained on the noisy observations $\mathbf{x}$. The vector field $f(\mathbf{y}) = \nabla_{\mathbf{y}} E_{\text{dense}}(A, g, \mathbf{x}, \mathbf{y})|_{\mathbf{x}=\mathbf{y}}$ shows the learned ring attractor dynamics. Note that the arrow lengths are proportional to the *logarithm* of the vector magnitude for visualization purposes. **Right:** The kernel IPR of the responses (green line) increases throughout training. The red and purple dashed lines are the kernel IPR of the signal (black circle) and observations (blue dots).

Figure 2 illustrates the optimized model when the inputs are noisy two-dimensional observations of a nonlinear one-dimensional ring (unit circle). Here, we use the quadratic kernel $K(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle^2$ associated with the centered quadratic feature map $\Phi(x_1, x_2) = (2x_1^2 - 1, 2x_2^2 - 1, 2x_1 x_2)$. We then optimize according to the updates in equations 9–10. After optimization, we see that the model has learned a ring attractor from the noisy observations.

5 Discussion

We provide an interpretation of AM energy minimization that complements the standard storage-capacity perspective. Under a simple normalization constraint on activity, minimizing the AM energy is equivalent to optimizing a representational objective that trades off input reconstruction against maximizing the IPR. This suggests that AM dynamics are not only useful for retrieving discrete stored patterns, but can also operate as an unsupervised mechanism for learning low-dimensional structure from noisy observations. However, we only empirically test the model on toy datasets, so it remains to be seen if it is effective on natural signals.

Our model invites comparison with common dimensionality reduction methods such as PCA and kernel PCA (Schölkopf et al., 1998). Our results also complement earlier work on learning continuous attractors in recurrent neural networks (Seung, 1997) by providing an explicit objective-based learning algorithm.

Our work is also related to other perspectives on AM models. Recent work has highlighted similarities between dense AM dynamics and denoising-based generative diffusion models across data regimes (Pham et al., 2024). Our contribution is complementary: we focus on the unsupervised setting where the model only observes noisy inputs and learns structure by encouraging low-dimensional responses, rather than training on a supervised denoising objective. In addition, Blümel et al. (2025) interpret AM learning as maximizing redundancy through the lens of partial information decomposition. Dimensionality minimization can be viewed as method for increasing redundancy between neurons, so our work is consistent with and complementary to their findings.

Our work may have biological relevance. AM networks have long been used as models of memory storage and retrieval, often in the context of hippocampal computation. Our results suggest that the same Hebbian recurrent circuits could also be used for unsupervised learning of low-dimensional structure, consistent with proposals that early cortical circuits learn low-dimensional latent structure from sensory signals. This is compatible with the idea that canonical circuit motifs recur across the cortical hierarchy (Douglas et al., 1989).

REFERENCES

Larry Abbott, Kanaka Rajan, and Haim Sompolinsky. Interactions between intrinsic and stimulus-evoked activity in recurrent neural networks. In *The Dynamic Brain: An Exploration of Neuronal Variability and Its Functional Significance*, pp. 65–82. Oxford University Press, 2010.

Shun-Ichi Amari. Learning patterns and pattern sequences by self-organizing nets of threshold elements. *IEEE Transactions on computers*, 100(11):1197–1206, 1972.

Luca Ambrogioni. In search of dispersed memories: Generative diffusion models are associative memory networks. *Entropy*, 26(5):381, 2024.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

Mark Blümel, Andreas Christian Schneider, David Alexander Ehrlich, Valentin Neuhaus, Marcel Graetz, Michael Wibral, Abdullah Makkeh, and Viola Priesemann. Associative memory learning through redundancy maximization. In *New Frontiers in Associative Memories*, 2025.

Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168(2):288–299, 2017.

Rodney J Douglas, Kevan AC Martin, and David Whitteridge. A canonical microcircuit for neocortex. *Neural Computation*, 1(4):480–488, 1989.

Lyndon R Duong, Thomas E Yerxa, Eero P Simoncelli, and David Lipshutz. Efficiently encoding noisy inputs in a recurrent network with adaptive representational dimensionality. In *Computational and Systems Neuroscience (COSYNE)*, Lisbon, March 2026.

Peiran Gao, Eric Trautmann, Byron Yu, Gopal Santhanam, Stephen Ryu, Krishna Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. *BioRxiv*, pp. 214262, 2017.

Benjamin Hoover, Hendrik Strobelt, Dmitry Krotov, Judy Hoffman, Zsolt Kira, and Duen Horng Chau. Memory in plain sight: A survey of the uncanny resemblances between diffusion models and associative memories. In *Associative Memory & Hopfield Networks*, 2023.

John J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

John J. Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, 81(10):3088–3092, 1984.

Dmitry Krotov. A new frontier for Hopfield networks. *Nature Reviews Physics*, 5(7):366–367, 2023.

Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition. *Advances in Neural Information Processing Systems*, 29, 2016.

William A. Little. The existence of persistent states in the brain. *Mathematical Biosciences*, 19 (1-2):101–120, 1974.

Cengiz Pehlevan, Anirvan M. Sengupta, and Dmitri B. Chklovskii. Why Do Similarity Matching Objectives Lead to Hebbian/Anti-Hebbian Networks? *Neural Computation*, 30(1):84–124, January 2018. doi: 10.1162/neco_a_01018.

Bao Pham, Gabriel Raya, Matteo Negri, Mohammed J Zaki, Luca Ambrogioni, and Dmitry Krotov. Memorization to generalization: The emergence of diffusion models from associative memory. In *NeurIPS Workshop on Scientific Methods for Understanding Deep Learning*, 2024.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David Kreil, Michael K Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021.

Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. *Advances in Neural Information Processing Systems*, 36:66377–66389, 2023.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

H Sebastian Seung. Learning continuous attractors in recurrent networks. *Advances in Neural Information Processing Systems*, 10, 1997.

Matthew Smart, Alberto Bietti, and Anirvan M Sengupta. In-context denoising with one-layer transformers: connections between attention and associative memory retrieval. In *New Frontiers in Associative Memories*, 2025.

A LEGENDRE TRANSFORM FOR KERNEL IPR

Under the constraint $\mathbb{E}_{\mathbf{y}\sim p(\mathbf{y})}[K(\mathbf{y},\mathbf{y})] = 1$, we can rewrite equation 5 as

$$D_{\text{IPR},K} = \mathbb{E}_{(\mathbf{y},\mathbf{y}')\sim p(\mathbf{y})\times p(\mathbf{y}')}\left[K^2(\mathbf{y},\mathbf{y}')\right].$$

Since K is a symmetric positive definite kernel, there exists a Hilbert space $\mathcal{H}$ with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and a feature map $\Phi : \mathbb{R}^N \mapsto \mathcal{H}$ such that $K(\mathbf{x},\mathbf{y}) = \langle\Phi(\mathbf{x}),\Phi(\mathbf{y})\rangle_{\mathcal{H}}$ (Aronszajn, 1950). Let $\mathcal{L}(\mathcal{H})$ denote the space of bounded linear operators on $\mathcal{H}$ and let $\|\cdot\|_{\text{HS}}$ denote the Hilbert-Schmidt norm on $\mathcal{L}(\mathcal{H})$. Using a Legendre transform, we can write

$$\begin{aligned} D_{\text{IPR},K} &= \mathbb{E}\left[\langle\Phi(\mathbf{y}),\Phi(\mathbf{y}')\rangle^2\right] \\ &= \mathbb{E}\left[\langle(\Phi(\mathbf{y}')\otimes\Phi(\mathbf{y}'))\,\Phi(\mathbf{y}),\Phi(\mathbf{y})\rangle\right] \\ &= \max_A \left\{2\mathbb{E}_{\mathbf{y}\sim p(\mathbf{y})}\left[\langle A\Phi(\mathbf{y}),\Phi(\mathbf{y})\rangle\right] - \|A\|_{\text{HS}}^2\right\}, \end{aligned}$$

where $\Phi(\mathbf{y}')\otimes\Phi(\mathbf{y}')$ denotes the operator in $\mathcal{L}(\mathcal{H})$ that maps $\Phi(\mathbf{y})$ to $\langle\Phi(\mathbf{y}),\Phi(\mathbf{y}')\rangle\Phi(\mathbf{y}')$.