Perspective: Summary Statistics of Learning*

Jacob A. Zavatone-Veth[1,2], **Blake Bordelon**[3], and **Cengiz Pehlevan**[1,4,5]
[1]Center for Brain Science, [2]Society of Fellows,
[3]Center for Mathematical Sciences and Applications,
[4]John A. Paulson School of Engineering and Applied Sciences,
[5]Kempner Institute for the Study of Natural and Artificial Intelligence,
Harvard University, Cambridge, MA, USA
{jzavatoneveth@fas, blake_bordelon@g, cpehlevan@seas}.harvard.edu

Editors: Marco Fumero, Clementine Domine, Zorah Lähner, Irene Cannistraci, Bo Zhao, Alex Williams

Abstract

How can we make sense of large-scale recordings of neural activity across learning? Theories of neural network learning with their origins in statistical physics offer a potential answer: for a given task, there are often a small set of summary statistics that are sufficient to predict performance as the network learns. Here, we review recent advances in how summary statistics can be used to build theoretical understanding of neural network learning. We then argue for how this perspective can inform the analysis of neural data, enabling better understanding of learning in biological and artificial neural networks.

1 Introduction

Experience reshapes neural population activity, molding an animal's representations of the world as it learns to perform new tasks. Thanks to advances in experimental technologies, it is just now becoming possible to measure changes in the activity of large neural populations across the course of learning [2–8]. However, with this new capability comes the challenge of identifying which features of high-dimensional activity patterns are meaningful for understanding learning. While analyses of representations have begun how to elucidate how learning reshapes the structure of activity, it is not in general clear whether these measurements are sufficient to understand how representational changes relate to behavior [4, 9–11].

In this Perspective, we propose that the principled identification of **summary statistics of learning** offers a possible path forward. This framework is grounded in theories of the statistical physics of learning in neural networks, which show that low-dimensional summary statistics are often sufficient to predict task performance over the course of learning [12–14]. We argue that thinking systematically about summary statistics gives new insight into what existing approaches of quantifying neural representations reveal about learning, and allows identification of what additional measurements would be required to constrain models of plasticity. We emphasize that the goal of this Perspective is not to advocate for the use of a particular set of summary statistics, but rather to explain the general philosophy of this approach to understanding learning in high dimensions.

*This extended abstract is an abbreviated version of our long-form perspective [1].

Proceedings of the III edition of the Workshop on Unifying Representations in Neural Models (UniReps 2025).

2 What is a summary statistic?

We posit that summary statistics of learning must satisfy two minimal desiderata: First, **they must be low-dimensional.** That is, their dimension is low relative to the number of neurons in the network of interest. Indeed, most summary statistics we will encounter are determined by averages over the population of neurons. Second, **they must be sufficient to predict behavior across learning.** From a theoretical standpoint, there should exist a closed set of equations describing the evolution of the summary statistics that predict the network's performance. Summary statistics satisfying these two desiderata are often highly interpretable thanks to their clear relationship to the network architecture and learning task (Section 3). However, the summary statistics relevant for predicting performance may not be sufficient to predict all statistical properties of population activity (Section 4).

3 Summary statistics in theories of neural network learning

We now review how summary statistics emerge naturally in theoretical analyses of neural network learning. Out of many theoretical results, we focus on two example settings: here we discuss batch learning in wide and deep networks [12–25], and in Appendix A we discuss online learning from high-dimensional data in shallow networks. These model problems illustrate how relevant summary statistics may be identified given a task, network architecture, and learning rule.

Consider a deep fully-connected network with input $\mathbf{x} \in \mathbb{R}^D$, at training time t:

$$f(\mathbf{x}, t) = \frac{1}{\gamma\sqrt{N}} \sum_{i=1}^{N} w_i(t)\phi(h_i^{(L)}(\mathbf{x}, t)),$$
$$h_i^{(\ell+1)}(\mathbf{x}, t) = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} W_{ij}^{(\ell)}(t)\phi(h_j^{(\ell)}(\mathbf{x}, t)), \quad \ell \in \{1, \dots, L+1\},$$
$$h_i^{(1)}(\mathbf{x}, t) = \frac{1}{\sqrt{D}} \sum_{j=1}^{D} W_{ij}^{(0)}(t)x_j.$$

Suppose we use gradient flow to minimize the average error on a fixed set of training examples, and consider a regime where the hidden layer width N is small relative to the input dimension D (Figure 1a). What are the relevant summary statistics? Applying the chain rule, one finds that

$$\frac{d}{dt} f(\mathbf{x}, t) = -\mathbb{E}_{\mathbf{x}'} \sum_{\ell} G^{(\ell+1)}(\mathbf{x}, \mathbf{x}', t, t)\Phi^{(\ell)}(\mathbf{x}, \mathbf{x}', t, t)\frac{\partial \mathcal{L}}{\partial f(\mathbf{x}', t)},$$

where $\mathcal{L}$ is the loss function and $\mathbb{E}_{\mathbf{x}'}$ denotes expectation over the training dataset [20, 26, 27]. Here,

$$\Phi^{(\ell)}(\mathbf{x}, \mathbf{x}', t, t') = \frac{1}{N} \sum_{i=1}^{N} \phi(h_i^{(\ell)}(\mathbf{x}, t))\phi(h_i^{(\ell)}(\mathbf{x}', t'))$$

are **representational similarity matrices**, and

$$G^{(\ell)}(\mathbf{x}, \mathbf{x}', t, t') = \frac{1}{N} \sum_{i=1}^{N} g_i^{(\ell)}(\mathbf{x}, t)g_i^{(\ell)}(\mathbf{x}', t'), \quad g_i^{(\ell)}(\mathbf{x}, t) \equiv \gamma\sqrt{N}\frac{\partial f(\mathbf{x}, t)}{\partial h_i^{(\ell)}(\mathbf{x}, t)},$$

are **gradient similarity matrices**, which at each layer compare the hidden states and the gradient signals for each pair of data points and each pair of training times. Thus, as these matrices determine the dynamics of f, they are suitable summary statistics if they are low-dimensional relative to the set of synaptic weights, and if we can write down a closed set of equations for their dynamics.

First, it is easy to see that the criterion of dimensionality reduction requires that the number of training examples P is much less than the network width N, as the number of similarity matrix elements and the number of synaptic weights are of order P^2 and N^2, respectively. Second, it turns out that one can close the equations for $\Phi^{(\ell)}$ and $G^{(\ell)}$ provided that the width is large and that the synaptic weights start from an uninformed initial condition (*i.e.*, Gaussian random matrices) [20, 26–28]. Depending on how weights and learning rates are scaled, one can obtain different types of large-width ($N \to \infty$) limits (Figure 1b). In the *lazy / kernel* limit where γ is constant, these representational similarity matrices are static over the course of learning [26, 27]. If instead $\gamma \propto \sqrt{N}$), these objects evolve in a task-dependent manner even as $N \to \infty$ (Figure 1c) [20, 28].

A significant line of recent work in neuroscience aims to quantify neural representations and compare them across networks through analysis of representational similarity matrices [10, 11, 29, 30]. Here, we see that these kernel matrices arise naturally as summary statistics of forward signal propagation in wide and deep neural networks (Figure 1c-d). At the same time, those results show that tracking only feature kernels is *not* in general sufficient to predict performance over the course of learning. One needs access also to coarse-grained information about the plasticity rule in the form of gradient kernels, and to information about the network outputs (for instance $\partial\mathcal{L}/\partial f$). More theoretical work

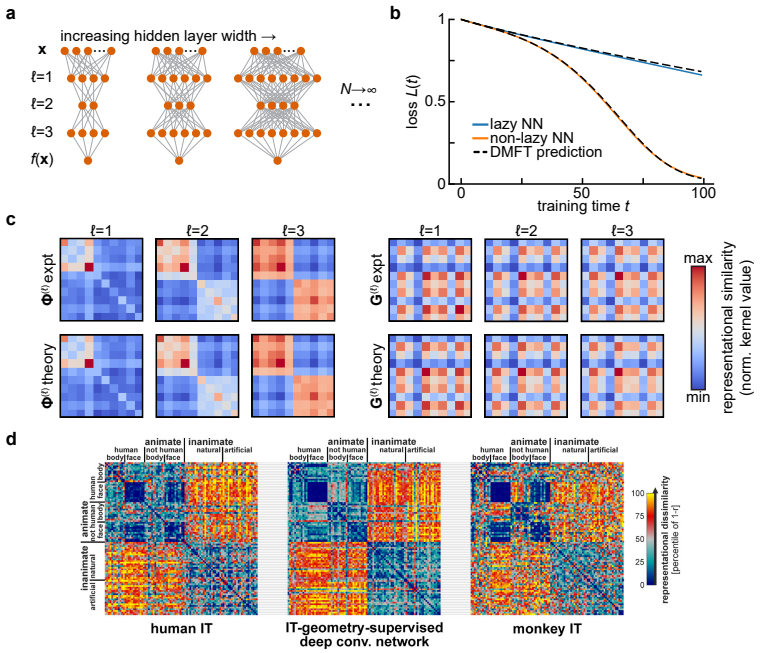

Figure 1: **Representational similarity kernels in wide neural network models and in the brain.** **a**. Diagram of the infinite-width limit of a deep feedforward neural network. For a fixed input and output dimension, one considers a sequence of networks of increasing hidden layer widths. **b**. Predicting the performance of width-2500 fully-connected networks with three hidden layers and $\tanh$ activations over training using the dynamical mean-field theory described in Section 3. Networks are trained on a synthetic binary classification dataset of 10 examples, with 5 examples assigned each class at random. This leads to block structure in the final representations. Adapted from Bordelon and Pehlevan [20]. **c**. The summary statistics in the dynamical mean field theory for the network in **b** are representational similarity kernels ($\mathbf{\Phi}^{(\ell)}$; *left*) and gradient similarity kernels ($\mathbf{G}^{\ell}$; *right*) for each layer. The top row shows kernels estimated from gradient descent training, and the bottom row the theoretical predictions. All kernels are shown at the end of training ($t = 100$). Adapted from Bordelon and Pehlevan [20]. **d**. Comparing representational similarity kernels across models and brains. Here, similarity is measured using the Pearson correlation r, and the *dissimilarity* $1 - r$ is plotted as a heatmap. Kernels resulting from fMRI measurements of human inferior temporal (IT) cortex (*left*) and electrophysiological measurements of macaque monkey IT cortex (*right*) are compared with the kernel for features from a deep convolutional neural network after optimal re-weighting to match human IT (*center*). Adapted from Figure 10 of Khaligh-Razavi and Kriegeskorte [31] with permission from N. Kriegeskorte under a CC-BY License.

is required to determine how to reliably estimate these gradient kernels from data, thereby providing a means to gain coarse-grained information about the underlying plasticity rule.

4 Implications for neural measurements

The two example settings detailed in Section 3 show how the relevant summary statistics of learning depend on network architecture and learning rule. Theoretical studies are just beginning to map out the full space of possible summary statistics for different network architectures [13–25]. Though details of the relevant summary statistics vary depending on the scaling regime and task—as illustrated by the examples above, where network width, training dataset size, and learning rule change the relevant statistics and their effective dynamics—they share broad structural principles. Thanks to these common structural features, these varied theories of summary statistics have common implications for the analysis and interpretation of neuroscience experiments.

4.1 Benign subsampling

The summary statistics encountered in Section 3 are robust to subsampling thanks to their basic nature as averages over the population of neurons. These statistical theories in fact post a far stronger notion of benign subsampling: they result in neurons that are statistically exchangeable. This is highly advantageous from the perspective of long-term recordings of neural activity, as reliable measurement of summary statistics does not require one to track the exact same neurons over time. Instead, it suffices to measure a sufficiently large subpopulation on any given day. This obviates many of the challenges presented by tracking neurons over multiple recording sessions [2]. Moreover, the variability and bias introduced by estimating summary statistics from a limited subset of relevant neurons can be characterized systematically [32–34]. Taken together, these properties mean that summary statistics are relatively easy to estimate given limited neural measurements, provided that exchangability is not too strongly violated [35]. There are limits, however, to how far one can subsample. For instance, representational similarity kernels are more affected by small, coordinated changes in the tuning of many neurons than large changes in single-neuron tuning (Figure 2) [4]. Determining the minimum number of neurons one must record in order to predict generalization dynamics across learning will be an important subject for future theoretical work [4, 35].

4.2 Invariances and representational drift

Though by our definition the summary statistics mentioned in Section 3 are sufficient to predict the network's performance, they are not sufficient statistics for all properties of the neural code. In particular, in part because they arise from theories in which neurons become exchangable, they have many invariances. These invariances mean that individual tuning curves can change substantially without altering the population-level computation [4]. For instance, the representational similarity kernels are invariant under rotation of the neural code at each layer. Similarly, overlaps with task-relevant directions are invariant to changes in the null space of those low-dimensional projections. These invariances mean that focusing on summary statistics of learning sets a particular aperture on what aspects of representations one can assay.

At the same time, the invariances of summary statistics have important consequences for functional robustness. In particular, they are closely related to theories of representational drift, the seemingly puzzling phenomenon of continuing changes in neural representations of task-relevant variables despite stable behavioral performance [2, 36]. Many models of drift explicitly propose that representational changes are structured in such a way that certain summary statistics are preserved (Figure 2a) [2, 37, 38]. Identifying the invariances of the summary statistics sufficient to determine task performance can allow for a systematic characterization of what forms of drift can be accommodated by a given network. Conversely, identifying the invariances of a representation once task performance stabilizes might suggest which summary statistics are relevant for the learning problem at hand.

4.3 Universality

An important lesson from the theory of high-dimensional statistics is that of *universality*: certain coarse-grained statistics are asymptotically insensitive to the details of the distribution. The most prominent example is the central limit theorem: the distribution of the sample mean of independent random variables tends to a Gaussian as the number of samples becomes large. A broader class of universality principles arise in random matrix theory: the distribution of eigenvalues and eigenvectors of a random matrix often become insensitive to details of the distribution of the elements as the matrix becomes large. Most famously, the Marčenko-Pastur theorem specifies that the singular values of a matrix with independent elements have a distribution that depends only on the mean and variance of the elements [39]. In learning problems, universality manifests through insensitivity of the model performance to details of the distributions of parameters or of features [40, 41].

From the perspective of summary statistics, statistical universality can allow simple theories to make informative macroscopic predictions even if they do not capture detailed properties of single neurons. For instance, the mean-field description of the learning dynamics of wide neural networks introduced in Section 3 are universal in that they depend on the initial distribution of hidden layer weights only through its mean and variance, even though the details of that distribution will affect the distribution of weights throughout training (Figure 2b-d) [42, 43]. Like the invariances to transformations of the neural population code mentioned before, this is nonetheless a double-edged sword: these universality properties mean that focusing on predicting performance commits one to coarse-graining away certain

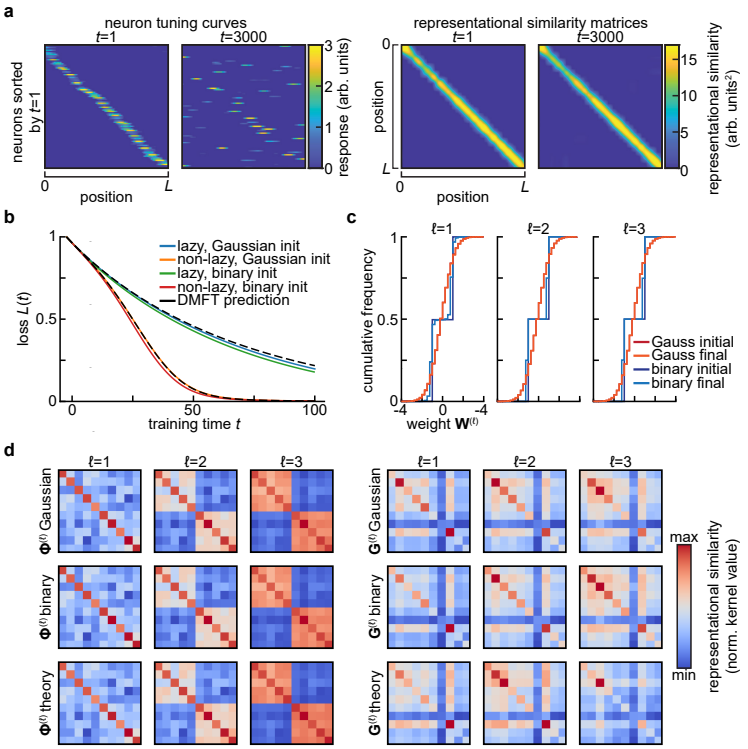

Figure 2: **Invariance and universality in summary statistics.**
a. Stable summary statistics despite drifting single-neuron responses. In Qin et al. [38]'s model of representational drift, single neurons are strongly tuned to a spatial variable, yet their tuning changes dramatically over time (*left*). Despite this drift, the similarity of the population representations of different spatial positions remains nearly constant (*right*). Adapted from Figure 5e of Qin et al. [38]. **b**. Universality of summary statistics in wide and deep networks with respect to the distribution of initial weights. Setting is as in Figure 1b-c, but also including a network for which the weights are initially drawn from $\{-1, +1\}$ with equal probability. Here, $N = 2000$, and a different realization of the random task is sampled relative to Figure 1b-c, so the loss curves are not identical. **c**. Cumulative distribution of weights at the start (*initial*) and end (*final*) of training for the networks shown in (**b**). Note that the small change in the weight distributions for the Gaussian-initialized networks is not visible at this resolution, and that one expects the size of weight changes to scale with $1/\sqrt{N}$ [20]. **d**. Feature and gradient kernels at the end of training for the networks in **b**. No substantial differences are visible between networks initialized with different weight distributions.

microscopic aspects of neural activity. Though these features are not required to predict macroscopic behavior, they may be important for understanding biological mechanisms.

5 Discussion

As reviewed here, the core insight of the statistical mechanics of learning in neural networks is the existence of low-dimensional summary statistics sufficient to predict behavioral performance. We now conclude by discussing future directions for theoretical inquiry. The models reviewed here are composed of exchangable neurons, which simplifies the relevant summary statistics and renders them particularly robust to subsampling. However, the brain has rich structure that can affect which summary statistics are sufficient to track learning and how those summary statistics may be measured. Biological neural networks are embedded in space, and their connectivity and selectivity is shaped by spatial structure [44–46]. Notably, many sensory areas are topographically organized: neurons with similar response properties are spatially proximal [47, 48]. Moreover, neurons can be classified into genetically-identifiable cell types [49], which may play distinct functional roles during learning [3, 50]. Future theoretical work must contend with these biological complexities in order to determine the relevant summary statistics of learning subject to these constraints.

Acknowledgments and Disclosure of Funding

J.A.Z.-V. is supported by the Office of the Director of the National Institutes of Health under Award Number DP5OD037354. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. JAZV is further supported by a Junior Fellowship from the Harvard Society of Fellows. B.B. is supported by a Google PhD Fellowship. C.P. is supported by NSF grant DMS-2134157, NSF CAREER Award IIS-2239780, DARPA grant DIAL-FP-038, a Sloan Research Fellowship, and The William F. Milton Fund from Harvard University. This work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence.

References

[1] Jacob A. Zavatone-Veth, Blake Bordelon, and Cengiz Pehlevan. Summary statistics of learning link changing neural representations to behavior. *Frontiers in Neural Circuits*, 19, 2025. ISSN 1662-5110. doi: 10.3389/fncir.2025.1618351. URL https://www.frontiersin.org/journals/neural-circuits/articles/10.3389/fncir.2025.1618351.

[2] Paul Masset, Shanshan Qin, and Jacob A Zavatone-Veth. Drifting neuronal representations: Bug or feature? *Biological Cybernetics*, pages 1–14, 2022. doi: doi.org/10.1007/s00422-021-00916-3.

[3] Andrew J.P. Fink, Samuel P. Muscinelli, Shuqi Wang, Marcus I. Hogan, Daniel F. English, Richard Axel, Ashok Litwin-Kumar, and Carl E. Schoonover. Experience-dependent reorganization of inhibitory neuron synaptic connectivity. *bioRxiv*, 2025. doi: 10.1101/2025.01.16.633450. URL https://www.biorxiv.org/content/early/2025/01/16/2025.01.16.633450.

[4] Nikolaus Kriegeskorte and Xue-Xin Wei. Neural tuning and representational geometry. *Nature Reviews Neuroscience*, 22(11):703–718, Nov 2021. ISSN 1471-0048. doi: 10.1038/s41583-021-00502-3. URL https://doi.org/10.1038/s41583-021-00502-3.

[5] Nicholas A. Steinmetz, Cagatay Aydin, Anna Lebedeva, Michael Okun, Marius Pachitariu, Marius Bauza, Maxime Beau, Jai Bhagat, Claudia Böhm, Martijn Broux, Susu Chen, Jennifer Colonell, Richard J. Gardner, Bill Karsh, Fabian Kloosterman, Dimitar Kostadinov, Carolina Mora-Lopez, John O'Callaghan, Junchol Park, Jan Putzeys, Britton Sauerbrei, Rik J. J. van Daal, Abraham Z. Vollan, Shiwei Wang, Marleen Welkenhuysen, Zhiwen Ye, Joshua T. Dudman, Barundeb Dutta, Adam W. Hantman, Kenneth D. Harris, Albert K. Lee, Edvard I. Moser, John O'Keefe, Alfonso Renart, Karel Svoboda, Michael Häusser, Sebastian Haesler, Matteo Carandini, and Timothy D. Harris. Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, 372(6539):eabf4588, 2021. doi: 10.1126/science.abf4588. URL https://www.science.org/doi/abs/10.1126/science.abf4588.

[6] Lin Zhong, Scott Baptista, Rachel Gattoni, Jon Arnold, Daniel Flickinger, Carsen Stringer, and Marius Pachitariu. Unsupervised pretraining in biological neural networks. *Nature*, Jun 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09180-y. URL https://doi.org/10.1038/s41586-025-09180-y.

[7] Weinan Sun, Johan Winnubst, Maanasa Natrajan, Chongxi Lai, Koichiro Kajikawa, Arco Bast, Michalis Michaelos, Rachel Gattoni, Carsen Stringer, Daniel Flickinger, James E. Fitzgerald, and Nelson Spruston. Learning produces an orthogonalized state machine in the hippocampus. *Nature*, 640(8057):165–175, Apr 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08548-w. URL https://doi.org/10.1038/s41586-024-08548-w.

[8] Sachin P. Vaidya, Guanchun Li, Raymond A. Chitwood, Yiding Li, and Jeffrey C. Magee. Formation of an expanding memory representation in the hippocampus. *Nature Neuroscience*, 28(7):1510–1518, Jul 2025. ISSN 1546-1726. doi: 10.1038/s41593-025-01986-3. URL https://doi.org/10.1038/s41593-025-01986-3.

[9] John W. Krakauer, Asif A. Ghazanfar, Alex Gomez-Marin, Malcolm A. MacIver, and David Poeppel. Neuroscience needs behavior: Correcting a reductionist bias. *Neuron*, 93(3):480–490, 2017. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2016.12.041. URL https://www.sciencedirect.com/science/article/pii/S0896627316310406.

[10] Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Christopher J. Cueva, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nathan Cloos, Nikolaus Kriegeskorte, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith,

Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment. *arXiv*, 2024. URL https://arxiv.org/abs/2310.13018.

[11] Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, Volume 2 - 2008, 2008. ISSN 1662-5137. doi: 10.3389/neuro.06.004.2008. URL https://www.frontiersin.org/journals/systems-neuroscience/articles/10.3389/neuro.06.004.2008.

[12] Timothy L. H. Watkin, Albrecht Rau, and Michael Biehl. The statistical mechanics of learning a rule. *Rev. Mod. Phys.*, 65:499–556, Apr 1993. doi: 10.1103/RevModPhys.65.499. URL https://link.aps.org/doi/10.1103/RevModPhys.65.499.

[13] Andreas Engel and Christian van den Broeck. *Statistical Mechanics of Learning*. Cambridge University Press, 2001. doi: https://doi.org/10.1017/CBO9781139164542.

[14] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016. doi: 10.1080/00018732.2016.1211393. URL https://doi.org/10.1080/00018732.2016.1211393.

[15] Gérard Ben Arous, Reza Gheissari, Jiaoyang Huang, and Aukosh Jagannath. High-dimensional SGD aligns with emerging outlier eigenspaces. *arXiv*, 2023. URL https://arxiv.org/abs/2310.03010.

[16] Sebastian Goldt, Madhu Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *Advances in neural information processing systems*, 32, 2019.

[17] David Saad and Sara A Solla. On-line learning in soft committee machines. *Physical Review E*, 52(4): 4225, 1995.

[18] Hugo Cui, Florent Krzakala, and Lenka Zdeborova. Bayes-optimal learning of deep random networks of extensive-width. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 6468–6521. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/cui23b.html.

[19] Jacob A. Zavatone-Veth and Cengiz Pehlevan. Depth induces scale-averaging in overparameterized linear Bayesian neural networks. In *Asilomar Conference on Signals, Systems, and Computers*, volume 55, 2021. doi: 10.1109/IEEECONF53345.2021.9723137.

[20] Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2023(11):114009, 2023.

[21] Jacob A Zavatone-Veth, William L Tong, and Cengiz Pehlevan. Contrasting random and learned features in deep Bayesian linear regression. *Physical Review E*, 105(6):064118, 2022.

[22] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

[23] Blake Bordelon, Jordan Cotler, Cengiz Pehlevan, and Jacob A. Zavatone-Veth. Dynamically learning to integrate in recurrent neural networks. *arXiv*, 2025. URL https://arxiv.org/abs/2503.18754.

[24] Luca Arnaboldi, Ludovic Stephan, Florent Krzakala, and Bruno Loureiro. From high-dimensional & mean-field dynamics to dimensionless ODEs: A unifying approach to SGD in two-layers networks. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 1199–1227. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/arnaboldi23a.html.

[25] Alexander van Meegen and Haim Sompolinsky. Coding schemes in neural networks learning classification tasks. *Nature Communications*, 16(1):3354, Apr 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-58276-6. URL https://doi.org/10.1038/s41467-025-58276-6.

[26] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[27] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/0d1a9651497a38d8b1c3871c84528bd4-Paper.pdf.

[28] Greg Yang and Edward J Hu. Tensor Programs IV: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pages 11727–11737. PMLR, 2021.

[29] Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4738–4750. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/252a3dbaeb32e7690242ad3b556e626b-Paper.pdf.

[30] Alex H. Williams. Equivalence between representational similarity analysis, centered kernel alignment, and canonical correlations analysis. *bioRxiv*, 2024. doi: 10.1101/2024.10.23.619871. URL https://www.biorxiv.org/content/early/2024/10/24/2024.10.23.619871.

[31] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLOS Computational Biology*, 10:1–29, 11 2014. doi: 10.1371/journal.pcbi.1003915. URL https://doi.org/10.1371/journal.pcbi.1003915.

[32] Hyunmo Kang, Abdulkadir Canatar, and SueYeon Chung. Spectral analysis of representational similarity with limited neurons. *arXiv*, 2025. URL https://arxiv.org/abs/2502.19648.

[33] Blake Bordelon and Cengiz Pehlevan. Dynamics of finite width kernel and prediction fluctuations in mean field neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2024(10):104021, 2024. doi: 10.1088/1742-5468/ad642b. URL https://dx.doi.org/10.1088/1742-5468/ad642b.

[34] Jacob A Zavatone-Veth, Abdulkadir Canatar, Benjamin S Ruben, and Cengiz Pehlevan. Asymptotics of representation learning in finite Bayesian neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11):114008, 2022. doi: 10.1088/1742-5468/ac98a6. URL https://dx.doi.org/10.1088/1742-5468/ac98a6.

[35] Peiran Gao, Eric Trautmann, Byron Yu, Gopal Santhanam, Stephen Ryu, Krishna Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv*, 2017. doi: 10.1101/214262. URL https://www.biorxiv.org/content/early/2017/11/12/214262.

[36] Michael E Rule, Timothy O'Leary, and Christopher D Harvey. Causes and consequences of representational drift. *Current Opinion in Neurobiology*, 58:141–147, 2019. ISSN 0959-4388. doi: https://doi.org/10.1016/j.conb.2019.08.005. URL https://www.sciencedirect.com/science/article/pii/S0959438819300303. Computational Neuroscience.

[37] Farhad Pashakhanloo and Alexei Koulakov. Stochastic gradient descent-induced drift of representation in a two-layer neural network. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 27401–27419. PMLR, 2023. URL https://proceedings.mlr.press/v202/pashakhanloo23a.html.

[38] Shanshan Qin, Shiva Farashahi, David Lipshutz, Anirvan M. Sengupta, Dmitri B. Chklovskii, and Cengiz Pehlevan. Coordinated drift of receptive fields in Hebbian/anti-Hebbian network models during noisy representation learning. *Nature Neuroscience*, 26(2):339–349, 2023. ISSN 1546-1726. doi: 10.1038/s41593-022-01225-z. URL https://doi.org/10.1038/s41593-022-01225-z.

[39] Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

[40] Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2022.

[41] Theodor Misiakiewicz and Basil Saeed. A non-asymptotic theory of kernel ridge regression: deterministic equivalents, test error, and GCV estimator. *arXiv preprint arXiv:2403.08938*, 2024.

[42] Eugene Golikov and Greg Yang. Non-Gaussian Tensor Programs. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 21521–21533. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/8707924df5e207fa496f729f49069446-Paper-Conference.pdf.

[43] Christopher Williams. Computing with infinite networks. In M.C. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996. URL https://proceedings.neurips.cc/paper_files/paper/1996/file/ae5e3ce40e0404a45ecacaaf05e5f735-Paper.pdf.

[44] Mikail Khona, Sarthak Chandra, and Ila Fiete. Global modules robustly emerge from local interactions and smooth gradients. *Nature*, 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08541-3. URL https://doi.org/10.1038/s41586-024-08541-3.

[45] Dmitri B. Chklovskii, Thomas Schikorski, and Charles F. Stevens. Wiring optimization in cortical circuits. *Neuron*, 34(3):341–347, 2002. ISSN 0896-6273. doi: 10.1016/S0896-6273(02)00679-7. URL https://doi.org/10.1016/S0896-6273(02)00679-7.

[46] Jennifer Stiso and Dani S. Bassett. Spatial embedding imposes constraints on neuronal network architectures. *Trends in Cognitive Sciences*, 22(12):1127–1142, 2018. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2018.09.007. URL https://www.sciencedirect.com/science/article/pii/S1364661318302250.

[47] Karl Kandler, Amanda Clause, and Jihyun Noh. Tonotopic reorganization of developing auditory brainstem circuits. *Nature Neuroscience*, 12(6):711–717, 2009. doi: https://doi.org/10.1038/nn.2332.

[48] Venkatesh N. Murthy. Olfactory maps in the brain. *Annual Review of Neuroscience*, 34(1): 233–258, 2011. doi: 10.1146/annurev-neuro-061010-113738. URL https://doi.org/10.1146/annurev-neuro-061010-113738.

[49] Meng Zhang, Xingjie Pan, Won Jung, Aaron R. Halpern, Stephen W. Eichhorn, Zhiyun Lei, Limor Cohen, Kimberly A. Smith, Bosiljka Tasic, Zizhen Yao, Hongkui Zeng, and Xiaowei Zhuang. Molecularly defined and spatially resolved cell atlas of the whole mouse brain. *Nature*, 624(7991):343–354, 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06808-9. URL https://doi.org/10.1038/s41586-023-06808-9.

[50] Junya Hirokawa, Alexander Vaughan, Paul Masset, Torben Ott, and Adam Kepecs. Frontal cortex neuron types categorically encode single decision variables. *Nature*, 576(7787):446–451, 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1816-9. URL https://doi.org/10.1038/s41586-019-1816-9.

[51] Gérard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for SGD: Effective dynamics and critical scaling. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 25349–25362. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a224ff18cc99a71751aa2b79118604da-Paper-Conference.pdf.

[52] Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10:041044, Dec 2020. doi: 10.1103/PhysRevX.10.041044. URL https://link.aps.org/doi/10.1103/PhysRevX.10.041044.

[53] M Biehl and H Schwarze. Learning by on-line gradient descent. *Journal of Physics A: Mathematical and General*, 28(3):643, feb 1995. doi: 10.1088/0305-4470/28/3/018. URL https://dx.doi.org/10.1088/0305-4470/28/3/018.

[54] Francesco Mori, Stefano Sarao Mannelli, and Francesca Mignacco. Optimal protocols for continual learning via statistical physics and control theory. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=rhhQjGjO9A.

[55] Francesca Mignacco and Francesco Mori. A statistical physics framework for optimal learning. *arXiv*, 2025. doi: 10.48550/arXiv.2507.07907. URL https://arxiv.org/abs/2507.07907.

[56] Andrea Montanari and Pierfrancesco Urbani. Dynamical decoupling of generalization and overfitting in large two-layer networks. *arXiv preprint arXiv:2502.21269*, 2025.

[57] Kazuyuki Hara, Kentaro Katahira, Kazuo Okanoya, and Masato Okada. Statistical mechanics of on-line node-perturbation learning. *IPSJ Online Transactions*, 4:23–32, 2011. doi: 10.2197/ipsjtrans.4.23.

[58] Kazuyuki Hara, Kentaro Katahira, Kazuo Okanoya, and Masato Okada. Statistical mechanics of node-perturbation learning for nonlinear perceptron. *Journal of the Physical Society of Japan*, 82(5):054001, 2013. doi: 10.7566/JPSJ.82.054001.

A Online learning in shallow neural networks with high dimensional data

Classical models of online gradient descent learning in high dimensions can be often be summarized with simple summary statistics [12, 13, 16, 17, 24, 51–53]. In this section, we discuss how the generalization performance of perceptrons and shallow (two-layer) neural networks trained on large quantities of high dimensional data can be summarized by simple weight alignment measures. Most simply, the perceptron model $f(\mathbf{x}) = \sigma\left(\frac{1}{\sqrt{D}}\mathbf{w}\cdot\mathbf{x}\right)$ seeks to learn a weight vector $\mathbf{w}\in\mathbb{R}^D$ which correctly classifies a finite set of randomly sampled training input-output pairs $(\mathbf{x}_\mu, y_\mu)$. If the inputs are random, $\mathbf{x}_\mu \sim \mathcal{N}(0,\mathbf{I}_D)$, and the targets $y_\mu = y(\mathbf{x}_\mu)$ are generated by a **teacher network** $y(\mathbf{x}) = \sigma\left(\frac{1}{\sqrt{D}}\mathbf{w}_\star\cdot\mathbf{x}\right)$, then the generalization performance (performance of the model on new *unseen data*, $\mathbb{E}_{\mathbf{x}}[(f(\mathbf{x})-y(\mathbf{x}))^2]$) is completely determined by the overlap of $\mathbf{w}$ with itself and with the target direction $\mathbf{w}_\star$

$$Q = \frac{1}{D}\mathbf{w}\cdot\mathbf{w}\ ,\ R = \frac{1}{D}\mathbf{w}\cdot\mathbf{w}_\star. \tag{1}$$

If the learning rate is scaled appropriately with the dimension D, the high-dimensional (large-D) limit of online stochastic gradient descent is given by a deterministic set of equations for Q and R:

$$\frac{d}{d\tau}\begin{bmatrix}Q(\tau)\\R(\tau)\end{bmatrix} = \mathbf{F}[Q(\tau),R(\tau)], \tag{2}$$

where the continuous training 'time' τ is the ratio of the number of samples seen to the dimension and $\mathbf{F}:\mathbb{R}^2\to\mathbb{R}^2$ is a nonlinear function that depends on the learning rate, the loss function, and the link function $\sigma(\cdot)$ [13, 16, 17, 24, 51]. Integrating this update equation allows one to predict the evolution of the generalization error as more training data are provided to the algorithm. Despite the infinite dimensionality of the original optimization problem, only two dimensions are necessary to capture the dynamics of generalization error.

The analysis of online perceptron learning can be extended to two layer neural networks with a small number of hidden neurons N,

$$f(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} a_i\,\phi\left(h_i(\mathbf{x})\right) \quad h_i(\mathbf{x}) = \frac{1}{\sqrt{D}}\mathbf{w}_i\cdot\mathbf{x}\ ,\ i\in\{1,...,N\}. \tag{3}$$

$$y(\mathbf{x}) = \sigma\left(h_1^\star(\mathbf{x}),...,h_K^\star(\mathbf{x})\right) \quad h_k^\star(\mathbf{x}) = \frac{1}{\sqrt{D}}\mathbf{w}_k^\star\cdot\mathbf{x}\ ,\ k\in\{1,...,K\}. \tag{4}$$

In this setting with isotropic random data, the relevant summary statistics are the readout weights $\mathbf{a}\in\mathbb{R}^N$, along with **overlap matrices** $\mathbf{Q}\in\mathbb{R}^{N\times N}$ and $\mathbf{R}\in\mathbb{R}^{N\times K}$ with entries

$$Q_{ij} = \frac{1}{D}\mathbf{w}_i\cdot\mathbf{w}_j\ ,\ R_{ik} = \frac{1}{D}\mathbf{w}_i\cdot\mathbf{w}_k^\star \tag{5}$$

For this system, we can track the gradient descent dynamics for $\mathbf{a}$, $\mathbf{Q}$, and $\mathbf{R}$ through a generalization of Equation (2) [16, 17, 52, 53]. This reduces the dimensionality of the dynamics from the $N+DN$ trainable parameters $\{a_i\}$, $\{w_j\}$ to $N+N^2+NK$ summary statistics, which is significant when $D\gg N+K$. This reduction enables the application of analyses that cannot scale to high dimensions, for instance control-theoretic methods to study optimal learning hyperparameters and curricula [54, 55]. Recent works have also begun to study approximations to these summary statistics when the network width N is also large, as further dimensionality reduction if possible when $\mathbf{Q}$ and $\mathbf{R}$ have stereotyped structures [24, 56].

Under what conditions is this reduction possible? Fundamentally, the summary statistics $\mathbf{a}$, $\mathbf{Q}$, and $\mathbf{R}$ are sufficient to determine the network's performance so long as the preactivations h_i and $h_k^\star$ are approximately Gaussian. Thus, one can relax the assumption that the inputs $\mathbf{x}$ are exactly Gaussian so long as a central limit theorem applies to h_i and $h_k^\star$ [16, 52]. Moreover, one can allow for correlations between the different input dimensions so long as h_i and $h_k^\star$ remain Gaussian. If $\mathbb{E}[\mathbf{x}\mathbf{x}^\top]=\mathbf{\Sigma}$, with a modification of the definition of the overlaps to $Q_{ij}=\frac{1}{D}\mathbf{w}_i\cdot\mathbf{\Sigma}\mathbf{w}_j$ and $R_{ik}=\frac{1}{D}\mathbf{w}_i\cdot\mathbf{\Sigma}\mathbf{w}_k^\star$ a similar reduction applies [24]. One can even consider extensions to plasticity rules other than stochastic gradient descent. For example, online node perturbation leads to a different effective dynamics for the same set of summary statistics [57, 58].

How could the overlaps $\mathbf{Q}$ and $\mathbf{R}$ be accessed from measurements of neural activity? And, in the absence of detailed knowledge of a teacher network, how could one identify the relevant overlaps? Under the simple structural assumptions of these models, one could estimate the overlaps from

covariances of network activity across stimuli, *i.e.*, with isotropic inputs one has $\mathbb{E}_{\mathbf{x}}[h_i h_k^\star] = R_{ik}$ and $\mathbb{E}_{\mathbf{x}}[h_i h_j] = Q_{ij}$. Moreover, one can in some cases detect this underlying low-dimensional structure by examining the principal components of the learning trajectory [15]. However, more theoretical work is required in this vein.