Dynamical Mean Field Theory, Random Matrices and Learning in High Dimensions *

Blake Bordelon
Harvard CMSA
UT Austin Oden Institute and Department of Neuroscience

June 24, 2026

Abstract

In this tutorial, I will introduce Dynamical Mean Field Theory (DMFT), a powerful framework that enables exact asymptotic descriptions of high dimensional disordered dynamical systems. First, I will introduce a few classic historical applications of DMFT in statistical physics and computational neuroscience. Second, I will examine linear dynamical systems and describe connections between DMFT correlation and response functions and objects that arise naturally in random matrix theory. Both cavity and Martin-Siggia-Rose path integral formalisms will be introduced and worked out in simple cases. One relevant application of this setting that I will describe is test and train loss dynamics for gradient flow in a random feature model. Lastly, after discussing how these ideas relate to infinite width feature learning neural networks trained from random initialization, I will introduce a recently developed method for computing spectral outliers for spiked matrix ensembles where spikes depend on the random initial bulk. This formalism generalizes the classical BBP phase transition to a setting that can describe weights in trained wide neural networks.

Contents

*These notes were developed for the ProbAI Theory of Scaling Laws Workshop 2026 at University of Warwick.

1 Motivation: Why High Dimensional Dynamics?

Scaling up deep learning systems is leading to significant advances in artificial intelligence [1, 2]. Performance of models in a variety of domains tends to increase regularly and predictably with increases in model size and training horizon (data). However, scaling up model size or training horizon naively can present potential challenges including instability in the learning dynamics (divergences in hidden features or model outputs) or suboptimal limiting behavior (degeneration to a kernel method) [3, 4]. This motivates a scientific understanding of (1) how to scale up towards well defined and stable limits and (2) what scaling laws you obtain when you adopt these rules.

Disordered Dynamics in High Dimensions The learning dynamics of large machine learning systems can be viewed as high dimensional dynamical systems that depend on many sources of disorder. These can generally include

- Random initialization of parameters [5, 6, 7, 8, 3, 9, 10]
- Randomly sampled data [11, 12, 13, 14]
- Stochastic gradient noise from randomly sampled minibatches [15, 16]

How can we develop analytical tools that can enable us to make sense of such systems? In this tutorial we will explore **dynamical mean field theory** (DMFT) methods as a flexible method for tracking high dimensional dynamical systems.

One Line Slogan A useful slogan that summarizes this approach is

High-dimensional random dynamics $\longrightarrow$ Single-site stochastic process with self-consistent noise and memory.

The cost of going from a high dimensional (often Markovian) random system, to a single site dynamical system is the emergence of non-Markovian effects in the reduced system that account for all the couplings [17].

Where we are going

1. Brief potted history of DMFT
2. Warmup Wigner/GOE example
3. Use of the method on simple learning problems: linear + random feature regression.
4. Training dynamics of infinite width neural networks
5. BBP for μP: weight spectral outlier dynamics with a two-level DMFT

1.1 Some History: Spin Glasses and Random Recurrent Networks

The method arose in the statistical mechanics of disordered systems, especially spin glasses. In a long range spin glass, each degree of freedom (spin) s_i interacts with all others through random couplings. These systems exhibit non-trivial energy landscapes and optimization over the spins s can be challenging.

p-Spin Model One popular classical model is the *spherical p-spin model* with relevant citations

1. Crisanti, Horner & Sommers 1993 [18]
2. Cugliandolo & Kurchan, 1993 [19]

which has Hamiltonian $H(\boldsymbol{s})$ which depends on soft-spins $\boldsymbol{s} \in \mathbb{R}^N$

$$H(\boldsymbol{s}) = -\frac{1}{\sqrt{p!}\ N^{(p-1)/2}} \sum_{i_1 \dots i_p = 1}^{N} J_{i_1 \dots i_p} s_{i_1} \dots s_{i_p} \quad J_{i_1 \dots i_p} \sim \mathcal{N}(0,1) \quad \frac{1}{N}\sum_{i=1}^{N} s_i^2 = 1 \tag{1}$$

where the tensor $\boldsymbol{J}$ is completely symmetric random tensor with $J_{\boldsymbol{\pi}(\boldsymbol{i})} = J_{\boldsymbol{i}}$ for any permutation $\boldsymbol{\pi}$ over the indices $\boldsymbol{i}$ and for any unordered tuple $i_1 < i_2 ... < i_p$ we have $J_{i_1 \dots i_p} \sim \mathcal{N}(0,1)$. The constraint $\frac{1}{N}\sum_{i=1}^{N} s_i^2 = 1$ ensures that the spins $\boldsymbol{s}$ are confined to a sphere of radius $\sqrt{N}$ (without this constraint the spins would blow up[1]).

1.1.1 Equilibrium Approach

A common approach to study these sytems is to analyze their equilibrium behaviors by examining properties of the associated Gibbs measure over states $\boldsymbol{s}$

$$\mu(\boldsymbol{s}, \boldsymbol{J}) = \frac{1}{Z(\boldsymbol{J})} \exp\left(-\beta H(\boldsymbol{s}, \boldsymbol{J})\right) \ , \ Z(\boldsymbol{J}) = \int_{\mathcal{S}^{N-1}} d\boldsymbol{s} \exp\left(-\beta H(\boldsymbol{s}, \boldsymbol{J})\right) \tag{2}$$

where $1/\beta$ is the temperature. As the Hamiltonian H depends on $\boldsymbol{J}$ it is actually itself a *random variable*, as is the partition function $Z(\boldsymbol{J})$. Of central interest are computing certain averages of observables $\mathcal{O}(\boldsymbol{s})$ of the spin state $\boldsymbol{s}$ when sampling from the Gibbs measure and also averaging over the random couplings $\boldsymbol{J}$

$$\langle \mathcal{O} \rangle = \left\langle \frac{1}{Z(\boldsymbol{J})} \int_{\mathcal{S}^{N-1}} d\boldsymbol{s} \exp\left(-\beta H(\boldsymbol{s}, \boldsymbol{J})\right)\ \mathcal{O}(\boldsymbol{s}) \right\rangle_{\boldsymbol{J}} \tag{3}$$

[1] Classically the spins are binary $\boldsymbol{s} \in \{\pm 1\}^N$, which means that such blowups are impossible. The spherical variant is easier to analyze than the binary one but still has the correct phenomenology and lessons for our purposes.

where $\langle \cdot \rangle_{\boldsymbol{J}}$ denotes an average over the random couplings between spins $J_{i_1 \ldots i_p}$. This average is very challenging to perform in general due to the dependence of $\boldsymbol{J}$ in the normalization factor $Z(\boldsymbol{J})$. A common (but nonrigorous strategy) that physicists employ to deal with such averages is the *replica trick* where the free energy $\langle \ln Z(\boldsymbol{J}) \rangle$ (from which, average observables $\langle \mathcal{O} \rangle$ can be computed) is treated as a limit $\lim_{n \to 0} \frac{1}{n} \ln \langle Z^n \rangle$. One performs such calculations at integer n, which can be interpreted as n copies/replicas of the original system. After averaging over $\boldsymbol{J}$, one must figure out some kind of *replica ansatz* that enables a limit $n \to 0$. Figuring out the appropriate ansatz is situation dependent.

Mathematical and Physical Concerns about Equilibrium Approach In general, the replica approach to study equilibrated systems makes mathematicians nervous for a variety of reasons relating to questions about whether integer moments $\langle Z^n \rangle$ enable a unique analytical continuation to $n \to 0$, how to justify a particular choice of replica ansatz, and issues about taking a saddle point as $N \to \infty$ before taking $n \to 0$ [20, 21]. It should be noted that there are rigorous ways to study these systems as well that avoid these issues. A physicist could be comfortable with these issues, but may wonder whether/how quickly low energy configurations of the Gibbs measure are **reachable** in reasonable times by dynamics in the $N \to \infty$ limit [19, 18]. Both of these concerns motivated another approach. Indeed for a certain range of temperatures β, the p-spin model for $p \geq 3$ can develop a plateau $\lim_{\tau \to \infty} \lim_{t_w \to \infty} C(t_w, t_w + \tau) = q_{\text{DMFT}}$ which differs from the equilibrium cross-replica overlap $q_{\text{equilibrium}} = 0$, reflecting a failure of the dynamics to reach equilibrium in $\mathcal{O}(1)$ times [19]. This is an issue relating to the order of the $N \to \infty$ and $t \to \infty$ limits.

1.1.2 Dynamical Approach

An alternative approach to studying the properties of the Gibbs measure is to analyze Langevin trajectories over states $s_i(t)$[2]

$$\frac{d}{dt} s_i(t) = -\lambda(t) s_i(t) - \frac{\partial}{\partial s_i} H(\boldsymbol{s}(t), \boldsymbol{J}) + \sqrt{2\beta^{-1}} \epsilon_i(t), \tag{4}$$

where the function $\lambda(t)$ is chosen to enforce the spherical constraint $\frac{1}{N} \sum_{i=1}^{N} s_i(t)^2 = 1$. The thermal noise $\epsilon_i(t)$ is a white noise process with statistics $\langle \epsilon_i(t) \epsilon_j(t') \rangle = \delta_{ij} \delta(t - t')$. If these dynamics are run for $t \to \infty$, the final states $\boldsymbol{s}$ should represent samples from the Gibbs measure at temperature β^{-1}. Because $H(\boldsymbol{s}, \boldsymbol{J})$ is a random variable (due to $\boldsymbol{J}$), the states $\boldsymbol{s}(t)$ are also random variables. However, unlike the equilibrium approach, it is very easy to perform the average over the random tensor $\boldsymbol{J}$ by tracking the transitions over states $s_i(t)$ [3]. The result is a **dynamical mean field theory** which can be expressed succinctly as a single site process with correlation and response functions.

DMFT Equations The result of averaging over $\boldsymbol{J}$ generates a simpler *single-site process* which represents the dynamics of a **typical spin** in the system

$$\frac{d}{dt} s(t) = -\lambda(t) s(t) + \xi(t) + p(p-1) \int_0^t dt' R(t, t') C(t, t')^{p-2} s(t') + \sqrt{2\beta^{-1}} \epsilon(t) \tag{5}$$

[2] Such dynamical Langevin trajectories should sample the Gibbs measure if $t \to \infty$ at fixed N. However, DMFT will describe $N \to \infty$ first, and these limits may not commute in the general case.

[3] By construction, tracking the dynamics from an initial state or distribution of states preserves the normalization of the probability distribution for s since $p(s_{t+dt}) = \int ds_t p(s_{t+dt} | s_t) p(s_t) = \int ds_t \delta(s_{t+dt} - s_t - dt\,[...]) p(s_t)$. The equilibrium approach has a normalization factor $Z(\boldsymbol{J})$ which must be dealt with.

where C is the **correlation function** and R is the **response function** defined as

$$C(t,t') = \left\langle s(t)s(t') \right\rangle \ , \ R(t,t') = \left\langle \frac{\delta s(t)}{\delta \xi(t')} \right\rangle \tag{6}$$

where $\langle\rangle$ represents an average over the thermal noise $\epsilon(t)$ and the zero-mean **colored noise process** $\xi(t)$ which has covariance

$$\left\langle \xi(t)\xi(t') \right\rangle = p\ C(t,t')^{p-1}. \tag{7}$$

Intuitively, the correlation function represents the two-time correlation of the spins and the response function represents the effect of the spin at time t due to a perturbation to the system at an earlier time t'. For random draws of $\epsilon(t)$ and $\xi(t)$, the $s(t)$ variable forms a distribution which **matches the distribution of spins within the infinite size system**.

Problem 1: Deriving p-spin DMFT (Return to this after the Tutorial)

Derive the DMFT equations for the soft spherical p-spin model using the cavity and MSR path integral techniques after you learn them from the later sections of this note.
Hint: when doing the MSR approach it may be useful to express $J_{i_1 \ldots i_p}$ as a sum over all possible permutations of a random asymmetric tensor $A_{i_1 \ldots i_p} \sim \mathcal{N}(0,1)$ with iid Gaussian entries.

In this DMFT, the resulting single site variables $s(t)$ are Gaussian, so using some Gaussian identities, one can generate closed integro-differential equations for these two variables

$$\partial_t C(t,t') = F_1[C,R] \ , \ \partial_t R(t,t') = F_2[C,R] \tag{8}$$

where F_1 and F_2 are functionals of C and R.

Problem 2: Closure of C, R Equations (Return to this after the Tutorial)

Work out the full expressions for F_1 and F_2. Hint: Novikov/Stein's Lemma to compute $\langle \xi(t)s(t') \rangle$.

1.1.3 Pros and Cons with the Dynamical Approach

The pros of the dynamical approach are

- As a substitute for replica method (no need for a replica ansatz to take $n \to 0$) [21].
- Reveals nontrivial phenomena that may be invisible to equilibrium / statics approaches (out-of-equilibrium effects, aging, violations of the fluctuation-dissipation relation, etc) [19].
- Perhaps most importantly, applies to dynamics that **do not have an energy function** [22].

The last of these points is important, this means DMFT can apply to other kinds of models such as systems with asymmetric couplings (like recurrent neural networks). In the case of learning dynamics, it means that DMFT can potentially apply to **other learning rules than exact / noisy gradient descent** [23].

Cons? The main downside/con of the DMFT approach is that the resulting single site equations are often challenging to characterize. To obtain analytical descriptions of the late time behaviors of these spin glasses, one often needs to assert either a *time-translation invariant ansatz* or an *aging ansatz* which plays a similar role to a replica symmetry or replica symmetry breaking ansatz. These can be verified / checked self-consistently and against the full numerical solutions (solutions for the full two-time correlation and response functions).

1.2 What if there is no Energy Function? The Random RNN Model

As mentioned above, DMFT can still work even if the dynamics do not correspond to gradient descent on an energy function. A very important model in computational neuroscience, is essentially an *asymmetric* random recurrent neural network where the RNN dynamics are

$$\frac{d}{dt}h_i(t) = -h_i(t) + \frac{g}{\sqrt{N}}\sum_{j=1}^{N} J_{ij}\phi(h_j(t)), \qquad J_{ij} \sim \mathcal{N}(0,1) \tag{9}$$

Key references for DMFT of this model

- Sompolinsky Crisanti & Sommers (SCS 1988) [5]
- Crisanti & Sompolinsky 2018 [24]
- Helias & Dahmen 2019 [25]

As the couplings J_{ij} are random and asymmetric, this theory is actually *easier* than the symmetric p-spin. In fact the heuristic of approximating

$$\xi_i(t) = \frac{g}{\sqrt{N}}\sum_{j=1}^{N} J_{ij}\phi(h_j(t)) \sim \text{Gaussian Process with independent neurons} \tag{10}$$

results in the correct description in this model.

DMFT for the Random RNN Model As in the previous model, in the $N \to \infty$ limit, all neurons become iid random variables drawn from a single-site measure. We can work out a single site description for the neuron activity $h(t)$ as a simple linear filter of a colored noise process $\xi(t)$

$$(1+\partial_t)h(t) = \xi(t)\ ,\ \left\langle \xi(t)\xi(t')\right\rangle = g^2\left\langle \phi(h(t))\phi(h(t'))\right\rangle \equiv g^2 C_\phi(t,t'). \tag{11}$$

The noise $\xi(t)$ has two-time correlation that is given by $g^2 C_\phi(t,t')$, which itself is computed as an average over the h distribution. This reveals a self-consistency structure

1. If we knew $C_\phi(t,t')$ we could characterize $\xi(t)$ and compute the distribution of $\{h(t)\}_{t\in\mathbb{R}}$.
2. If we knew the distribution of $\{h(t)\}_{t\in\mathbb{R}}$, we could perform the average $C_\phi(t,t') = \langle\phi(h(t))\phi(h(t'))\rangle$.

This is a common structure that we will commonly see in later sections. In this DMFT, we can also close the equations at the level of correlation functions.

Problem 3: Closure of DMFT Correlation Function (Easy)

Use the single site equation to derive the differential equation for $C_h(t,t') = \langle h(t)h(t')\rangle$

$$(1+\partial_t)(1+\partial_{t'})C_h(t,t') = g^2 C_\phi(t,t') = g^2 F(C_h(t,t), C_h(t',t'), C_h(t,t'))$$
$$F(c_1,c_2,c_3) = \left\langle \phi(h)\phi(h')\right\rangle_{h,h'\sim\mathcal{N}(0,\mathbf{\Sigma})} \ , \ \mathbf{\Sigma} = \begin{bmatrix} c_1 & c_3 \\ c_3 & c_2 \end{bmatrix}$$

where $\langle\rangle_{h,h'\sim\mathcal{N}(0,\mathbf{\Sigma})}$ represents the bivariate Gaussian measure with covariance $\mathbf{\Sigma} \in \mathbb{R}^{2\times 2}$.

TTI Ansatz After an initial transient period, we could imagine that the dynamics eventually forget about the initial conditions and converge to a history independent dynamics. In such a regime we could seek a **time-translation-invariant** ansatz

$$\text{TTI Ansatz:} \quad C_h(t,t') \simeq c(\tau) \ , \ \tau = t - t' \tag{12}$$

In this regime, the correlation function only depends on the time lag $\tau = t - t'$.

Problem 4: Newton's Law for TTI Dynamics (Easy)

Show that under the TTI ansatz the $c(\tau)$ correlation function satisfies the following second order differential equation

$$\partial_\tau^2 c(\tau) = c(\tau) - g^2 F(c_0, c_0, c(\tau)) = c(\tau) - g^2 F(c_0, c_0, c(\tau)) \tag{13}$$

Argue that this can be viewed as **Newton's Law** $\partial_\tau^2 c(\tau) = -\partial_c V_{c_0}(c)$ with a potential energy function $V_{c_0}(c)$ (parameterized by c_0) defined as

$$V_{c_0}(c) = -\frac{1}{2}c^2 + g^2 \left\langle \psi(h)\psi(h')\right\rangle_{h,h'\sim\mathcal{N}(0,\mathbf{\Sigma})} - g^2 \left[\left\langle \psi(h)\right\rangle_{h\sim\mathcal{N}(0,c_0)}\right]^2 \ , \ \mathbf{\Sigma} = \begin{bmatrix} c_0 & c \\ c & c_0 \end{bmatrix} \tag{14}$$

where $\psi(h) = \int_0^h dh' \phi(h')$ is the anti-derivative of the activation function ϕ. The last term is a constant that is chosen so that $V_{c_0}(0) = 0$. Show that Newton's Law implies the conservation of $\mathcal{E}$

$$\mathcal{E} = \frac{1}{2}\left(\partial_\tau c(\tau)\right)^2 + V_{c_0}(c) \ , \ \frac{d}{d\tau}\mathcal{E} = 0 \tag{15}$$

Hint: look up Price's theorem.

Interestingly, there are potentially many different $V_{c_0}(c)$ functions that are all parameterized by the value of $c_0 = c(0)$. Different types of solutions can thus be sought (time-independent solutions like $c(\tau) = c_0$, time decaying solutions where $\partial_\tau c(\tau) < 0$ for $\tau > 0$, periodic solutions, etc). It turns out that not all of these are stable depending on the value of g. In a minute, we will look at the *decaying solution*, which can be shown to be stable for $g > 1$.

Problem 5: Shape of the Potential (Medium)

Show that all higher derivatives (greater than two) of $V_{c_0}(c)$ are non-negative for $c \geq 0$

$$\forall n \geq 3 \ , \ \frac{\partial^n}{\partial c^n} V(c) \geq 0 \tag{16}$$

This implies that the second derivative is increasing monotonically. Argue that this implies that $V''_{c_0}(c)$ can equal zero at most at one point along the solution orbit. Now assume that $\phi(h)$ is an odd function (so that ψ is even). Argue that all odd derivatives of the potential vanish at $c = 0$

$$\frac{\partial^n}{\partial c^n} V_{c_0}(c)|_{c=0} = 0 \ , \ \forall n \in \{1, 3, 5, ...\} \tag{17}$$

If $V''_{c_0}(0) > 0$ then there is a single well centered at $c = 0$. If $V''_{c_0}(0) < 0$ then this means that the potential has a double well centered around some other values $\pm c_{\text{min}}$. Show that for odd functions ϕ, the condition $V''_{c_0}(0) > 0$ is equivalent to

$$g^2 \left[\left\langle \dot{\phi}(h) \right\rangle_{h \sim \mathcal{N}(0, c_0)} \right]^2 < 1 \tag{18}$$

Hint: Price's theorem is still your friend.

Now we have a sense of the structure of the potential. It is feasibly possible that $V_{c_0}(c)$ can have a double well potential that could generate decaying dynamics for $c(\tau)$. Let's seek a decaying solution.

Problem 6: Decaying Solution (Medium)

Assume a decaying solution so that $c(\infty) = 0$ and $\partial_\tau c(\tau)|_{\tau=\infty} = 0$. By the conservation law (and our choice of $V_{c_0}(0) = 0$) argue that this implies a zero energy condition $\mathcal{E} = 0$. Due to the conservation law and the even-symmetry condition $\partial_\tau c(\tau)|_{\tau=0} = 0$, show that this fixes the value of c_0

$$\mathcal{E} = \frac{1}{2}\dot{c}(\infty)^2 + V_{c_0}(c(\infty)) = 0 = V_{c_0}(c_0) = -\frac{1}{2}c_0^2 + g^2 \left\langle \psi(h)^2 \right\rangle - g^2 \left[\langle \psi(h) \rangle \right]^2 \tag{19}$$

where $\langle\rangle$ represents an average over the marginal $h \sim \mathcal{N}(0, c_0)$. With this self-consistent scalar solution c_0, we now have completely characterized the shape of the potential $V_{c_0}(c)$ and we can integrate forward from the initial condition $c(0) = c_0$ and $\partial_\tau c(\tau)|_{\tau=0} = 0$ to $\tau \to +\infty$ forward with the equation of motion

$$\partial_\tau^2 c(\tau) = -V'_{c_0}(c). \tag{20}$$

For more information on the stability of this solution for $g > 1$ and the Lyapunov exponent, please review the classic work.

1.3 Signal Propagation in Randomly Initialized Neural Networks

Often in deep learning theory we are concerned with propagation of signals through feedforward neural network architectures where each layer's weight matrix is independent and random (instead of recurrent networks where weights are shared across steps). In a feedforward multi-layer perceptron, consider mapping input data $\boldsymbol{x}_\mu \in \mathbb{R}^D$ through hidden states $\boldsymbol{h}^\ell \in \mathbb{R}^N$ which have the following

forward pass dynamics through layers ℓ

$$\boldsymbol{h}_\mu^{\ell+1} = \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell \phi(\boldsymbol{h}_\mu^\ell) + \boldsymbol{b}^\ell \ , \ \boldsymbol{h}_\mu^1 = \frac{1}{\sqrt{D}} \boldsymbol{W}^0 \boldsymbol{x}_\mu \tag{21}$$

where the weight matrices $W_{ij}^\ell \sim \mathcal{N}(0, \sigma_W^2)$ have iid normal entries and the bias vectors $b_i^\ell \sim \mathcal{N}(0, \sigma_b^2)$ are also random with iid entries.
Key references are

- Neal et al 1998 [26]
- Poole et al., 2016 [27]
- Schoenholz et al 2016 [28]
- Lee et al 2018 [6]
- Jacot et al 2019 [7]

As in the previous sections, the $N \to \infty$ results in random neuron activity h_μ^ℓ that is iid across neurons but correlated across inputs

$$N \to \infty \implies \{h_\mu^\ell\} \sim \text{Gaussian with Correlations across Data Points} \tag{22}$$

Below we can work out the formula for the correlations $C_{\mu\nu}^\ell = \left\langle h_\mu^\ell h_\nu^\ell \right\rangle$ with base case $C_{\mu\nu}^1 = \frac{1}{D} \boldsymbol{x}_\mu \cdot \boldsymbol{x}_\nu$.

Problem 7: Signal Propagation Recursion (Easy)

Show that the correlations $C_{\mu\nu}^\ell$ satisfy the recursion over hidden layers ℓ

$$C_{\mu\nu}^{\ell+1} = \sigma_W^2 \left\langle \phi(h)\phi(h') \right\rangle_{h,h' \sim \mathcal{N}(0, \boldsymbol{\Sigma}(\boldsymbol{C}^\ell))} + \sigma_b^2 \quad , \quad \boldsymbol{\Sigma}(\boldsymbol{C}^\ell) = \begin{bmatrix} C_{\mu\mu}^\ell & C_{\mu\nu}^\ell \\ C_{\mu\nu}^\ell & C_{\nu\nu}^\ell \end{bmatrix} \tag{23}$$

Work out the expression under the assumption that $C_{\mu\mu}^1 = C_{\nu\nu}^1 = 1$ and $C_{\mu\nu}^1 = \rho \geq 0$. Argue that the correlation at each layer is parameterized by $C_{\mu\mu}^\ell = C_{\nu\nu}^\ell = q_\ell$ and $C_{\mu\nu}^\ell = c_\ell$ where

$$q_{\ell+1} = \sigma_W^2 \int \mathcal{D}z \ \phi\left(\sqrt{q_\ell} z\right)^2 + \sigma_b^2 \tag{24}$$

$$c_{\ell+1} = \sigma_W^2 \int \mathcal{D}z \left[\int \mathcal{D}z' \phi\left(\sqrt{q_\ell - c_\ell} \ z' + \sqrt{c_\ell} \ z \right) \right]^2 + \sigma_b^2 \tag{25}$$

where $\mathcal{D}z = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$ is the standard Gaussian measure. Argue that, unlike the random RNN model, this is a Markovian process $(q_{\ell+1}, c_{\ell+1}) = F[q_\ell, c_\ell]$ over layers rather than a self-consistent equation for the correlation. Argue that this is primarily different from the random RNN model because **weights are independent across layers in the MLP but are shared across timesteps in the RNN**.

This recursion is the starting point for many results in deep learning theory including the neural network Gaussian process theory and neural tangent kernel.

2 Simple/Minimal Models of DMFT to Learn the Method

Let $\boldsymbol{h}(t) \in \mathbb{R}^N$ denote the state of a high-dimensional dynamical system like some of the examples of previous sections. Two central DMFT order parameters are the correlation and response functions. These examples build on many of the examples from the review article [17].

Definition 2.1 (Correlation)**.** The empirical two-time correlation is

$$C(t,t') = \frac{1}{N}\boldsymbol{h}(t)\cdot\boldsymbol{h}(t'). \tag{26}$$

This function measures the correlation of the state at time t to the state at time t'. We have already seen an example of this in the SCS random RNN model above.

Definition 2.2 (Response)**.** Introduce a source perturbation $\boldsymbol{j}(t) \in \mathbb{R}^N$ in the dynamics. The normalized linear response is

$$R(t,t') = \frac{1}{N}\mathrm{Tr}\frac{\delta\boldsymbol{h}(t)}{\delta\boldsymbol{j}(t')^\top}. \tag{27}$$

It measures the average effect at time t of an infinitesimal perturbation at an earlier time t'.

The response function is typically causal: $R(t,t') = 0$ for $t < t'$. At equilibrium, the correlation and response function are related through the fluctuation dissipation theorem, but this can be violated when out of equilibrium.

3 Linear systems and the random matrix dictionary

Key references

- Bordelon & Pehlevan 2026 [17]

The simplest place where DMFT meets random matrix theory is a linear system

$$\frac{d}{dt}\boldsymbol{h}(t) = -\boldsymbol{M}\boldsymbol{h}(t) + \boldsymbol{j}(t), \tag{28}$$

where $\boldsymbol{M} \in \mathbb{R}^{N\times N}$ is a random matrix. If $\boldsymbol{M}$ is fixed, the response is exactly

$$R(t,t') = \frac{1}{N}\mathrm{Tr}\exp[-\boldsymbol{M}(t-t')]\,\Theta(t-t'), \tag{29}$$

where Θ is the Heaviside step function ($\Theta(z) = 0$ for $z \le 0$ and $\Theta(z) = 1$ for $z > 0$). This function is time-translation invariant since $R(t,t') = R(\tau)$ with $\tau = t - t'$. The Fourier/Laplace transform is

$$\mathcal{R}(\omega) = \int_{-\infty}^{\infty} \mathrm{d}\tau\, e^{-\mathrm{i}\omega\tau} R(\tau) = \frac{1}{N}\mathrm{Tr}(i\omega + \boldsymbol{M})^{-1}. \tag{30}$$

Thus the DMFT response is the normalized resolvent of $\boldsymbol{M}$ evaluated on the imaginary axis. If M has real eigenvalues with empirical density $\rho(\lambda)$, then

$$R(\tau) = \int \mathrm{d}\lambda\,\rho(\lambda)e^{-\lambda\tau}, \qquad \mathcal{R}(\omega) = \int \mathrm{d}\lambda\,\frac{\rho(\lambda)}{i\omega+\lambda}. \tag{31}$$

The density can be recovered by the Stieltjes inversion formula

$$\rho(\lambda) = \lim_{\varepsilon\downarrow 0}\frac{1}{\pi}\,\mathrm{Im}\,\mathcal{R}(\omega)\big|_{i\omega=-\lambda-i\varepsilon}. \tag{32}$$

Problem 8: Correlation and Response Relationship for Hermitian System (Easy)

If $\boldsymbol{M} = \boldsymbol{M}^\top$, then for the homogeneous dynamics $\boldsymbol{h}(t) = e^{-\boldsymbol{M}t}\boldsymbol{h}(0)$ with isotropic initialization $\boldsymbol{h}_0 \sim \mathcal{N}(0, \boldsymbol{I})$ then,

$$C(t,t') = \frac{1}{N}\mathrm{Tr} e^{-\boldsymbol{M}(t+t')} = R(t+t'). \tag{33}$$

Thus the correlation is determined by the same spectral measure $\rho(\lambda)$ as the response.

We will see some simple examples of these linear systems where we can compute C and R from the DMFT equations and compare to standard random matrix theory. We will also describe situations where the correlation function can have dramatically different behaviors than the response.

4 Warm-up: GOE dynamics and the semicircle law

Consider the linear dynamics with $\boldsymbol{M} = \frac{1}{\sqrt{N}}\boldsymbol{A}$ where $A_{ij} \sim \mathcal{N}(0,1)$ and a symmetry requirement $\boldsymbol{A} = \boldsymbol{A}^\top$, known as a GOE/Wigner matrix.

$$\frac{d}{dt}\boldsymbol{h}(t) = -\frac{1}{\sqrt{N}}\boldsymbol{A}\boldsymbol{h}(t) + \boldsymbol{j}(t) \tag{34}$$

We will first derive the effective single-site process using the cavity method.

4.1 Cavity derivation

Add a new coordinate $h_0(t)$ coupled to the original N variables by a random vector $\boldsymbol{a}_0 \in \mathbb{R}^N$. The new coordinate obeys

$$\frac{d}{dt}h_0(t) = -\frac{1}{\sqrt{N}}\boldsymbol{a}_0 \cdot \tilde{\boldsymbol{h}}(t) + j_0(t), \tag{35}$$

where $\tilde{\boldsymbol{h}}(t)$ denotes the perturbed trajectory of the original system after adding the new site. The perturbation to the old variables is small, so to leading order

$$\tilde{\boldsymbol{h}}(t) = \boldsymbol{h}(t) - \frac{1}{\sqrt{N}}\int^t \mathrm{d}t' \, \frac{\partial \boldsymbol{h}(t)}{\partial \boldsymbol{j}(t')^\top}\boldsymbol{a}_0 h_0(t') + \text{higher order (small) terms} \tag{36}$$

Substituting into the dynamics for h_0 gives two terms:

$$\frac{d}{dt}h_0(t) = -\frac{1}{\sqrt{N}}\boldsymbol{a}_0 \cdot \boldsymbol{h}(t) + \frac{1}{N}\int^t \mathrm{d}t' \, \boldsymbol{a}_0^\top \frac{\partial \boldsymbol{h}(t)}{\partial \boldsymbol{j}(t')^\top}\boldsymbol{a}_0 h_0(t') + j_0(t). \tag{37}$$

The first term is a Gaussian process by the central limit theorem since we note that $\boldsymbol{a}_0$ is now independent of $\boldsymbol{h}(t)$

$$u_0(t) = -\frac{1}{\sqrt{N}}\boldsymbol{a}_0 \cdot \boldsymbol{h}(t), \qquad \big\langle u_0(t)u_0(t')\big\rangle = \frac{1}{N}\boldsymbol{h}(t)\cdot\boldsymbol{h}(t') = C(t,t'). \tag{38}$$

The second term concentrates around its average by a law of large numbers effect

$$\frac{1}{N}\boldsymbol{a}_0^\top \frac{\partial \boldsymbol{h}(t)}{\partial \boldsymbol{j}(t')^\top}\boldsymbol{a}_0 \sim R(t,t') + \text{Fluctuations of Size } \frac{1}{\sqrt{N}}. \tag{39}$$

Therefore, in the $N \to \infty$ limit, a typical coordinate $h_0(t)$ follows its own separate dynamics which are decoupled from other sites

$$\frac{d}{dt}h_0(t) = u_0(t) + \int^t \mathrm{d}t'\, R(t,t')h_0(t') + j_0(t), \qquad u_0 \sim \mathrm{GP}(0, C). \tag{40}$$

From this stochastic process, the response and correlation can be computed directly.

Problem 9: Response Dynamics (Easy)

Differentiate (40) with respect to the source. Show that the response function obeys the integro-differential equation

$$\partial_t R(t,t') = \delta(t-t') + \int \mathrm{d}s\, R(t,s)R(s,t'). \tag{41}$$

Assuming time-translation invariance, take a Fourier transform $\mathcal{R}(\omega) = \int d\tau e^{-i\omega\tau} R(\tau)$ to find

$$i\omega\mathcal{R}(\omega) = 1 + \mathcal{R}(\omega)^2. \tag{42}$$

We are interested in the branch that goes as $1/(i\omega)$ as $\omega \to \infty$. Show that this is the solution

$$\mathcal{R}(\omega) = \frac{1}{2}\left[i\omega - \sqrt{(i\omega)^2 - 4}\right]. \tag{43}$$

Using Stieltjes inversion (be careful about which branch of the square root with the $i\epsilon$) yields the semicircle law

$$\rho(\lambda) = \frac{1}{2\pi}\sqrt{[4-\lambda^2]_+}, \tag{44}$$

where $[z]_+ = \max(0, z)$ selects only the region where $z \geq 0$.

5 Path-integral / Martin-Siggia-Rose formulation

The cavity derivation is intuitive. The path-integral formalism is more systematic, scales better to complicated models, and enables finite size corrections. We briefly outline the Martin–Siggia–Rose construction.

We start by defining the generating functional

$$Z[\boldsymbol{\zeta}] = \left\langle \exp\left(\int \mathrm{d}t\ \boldsymbol{\zeta}(t)\cdot\boldsymbol{h}(t)\right)\right\rangle. \tag{45}$$

from which moments can be computed as derivatives near $\boldsymbol{\zeta} = 0$. This function satisfies $Z[\mathbf{0}] = 1$ by construction. The average $\langle\cdot\rangle$ is over the disorder, in this case the matrix $\boldsymbol{M}$. To enforce the dynamics (28), we insert Dirac-delta functions which represent the transition density which maps the distribution of $\boldsymbol{h}(t)$ to the distribution of $\boldsymbol{h}(t+dt)$[4]. Abstractly, we make use of the Fourier transform of the Dirac Delta $\delta(z) = \int \frac{d\hat{z}}{2\pi}\exp(i\hat{z}z)$. Formally we define the dynamics in discrete increments of step size dt first and later take $dt \to 0$

$$\begin{aligned}&\delta\left(\boldsymbol{h}(t+\mathrm{d}t) - \boldsymbol{h}(t) + \mathrm{d}t\boldsymbol{M}\boldsymbol{h}(t) - dt\boldsymbol{j}(t)\right)\\ &= \int \frac{d\hat{\boldsymbol{h}}(t)}{2\pi}\exp\left[i\ \hat{\boldsymbol{h}}(t)\cdot\left(\boldsymbol{h}(t+\mathrm{d}t) - \boldsymbol{h}(t) + \mathrm{d}t\boldsymbol{M}\boldsymbol{h}(t) - \mathrm{d}t\boldsymbol{j}(t)\right)\right]\end{aligned} \tag{46}$$

[4]This is known as the Ito convention, rather than the Stratanovich convention.

where we introduced a conjugate field $\hat{\boldsymbol{h}}(t)$. Repeating this for every time $t \in \mathbb{R}$ and taking $\mathrm{d}t \to 0$, the original moment generating function can be written as a path integral

$$Z = \int D\boldsymbol{h} D\hat{\boldsymbol{h}} \left\langle \exp\left[i \int \mathrm{d}t\, \hat{\boldsymbol{h}}(t) \cdot (\partial_t \boldsymbol{h}(t) + \boldsymbol{M}\boldsymbol{h}(t) - \boldsymbol{j}(t)) + \int dt \boldsymbol{h}(t) \cdot \boldsymbol{\zeta}(t) \right] \right\rangle_{\boldsymbol{M}}. \tag{47}$$

After averaging over the random matrix, the action can be expressed in terms of collective fields such as

$$C(t,t') = \frac{1}{N}\boldsymbol{h}(t) \cdot \boldsymbol{h}(t'), \qquad R(t,t') = -\frac{i}{N}\boldsymbol{h}(t) \cdot \hat{\boldsymbol{h}}(t'). \tag{48}$$

We note that, under the path integral, the $\hat{\boldsymbol{h}}$ field can be interpreted as a derivative with respect to the source $\boldsymbol{j}(t')$. Let $\mathcal{O}[\boldsymbol{h}]$ represent any scalar observable of the states $\{\boldsymbol{h}(t)\}_{t\in\mathbb{R}}$

$$\left\langle \frac{\partial}{\partial \boldsymbol{j}(t')} \mathcal{O}[\boldsymbol{h}] \right\rangle = \int D\boldsymbol{h} D\hat{\boldsymbol{h}} \left\langle \exp\left[i \int \mathrm{d}t\, \hat{\boldsymbol{h}}(t) \cdot (\partial_t \boldsymbol{h}(t) + \boldsymbol{M}\boldsymbol{h}(t) - \boldsymbol{j}(t)) + \int dt \boldsymbol{h}(t) \cdot \boldsymbol{\zeta}(t) \right] \right\rangle_{\boldsymbol{M}} \times \mathcal{O}[\boldsymbol{h}] \quad \left(-i\hat{\boldsymbol{h}}(t') \right) \tag{49}$$

Thus factors for objects appearing under the path integral we can make the correspondece

$$-i\hat{\boldsymbol{h}}(t) \to \frac{\partial}{\partial \boldsymbol{j}(t)} \tag{50}$$

For this reason, the objects $\hat{\boldsymbol{h}}$ are sometimes referred to as **response fields**.

Problem 10: DMFT Action (Hard)

To enforce the definitions of the order parameters C and R under the path integral, we use the same Dirac Delta trick $\delta(NC(t,t') - \boldsymbol{h}(t)\cdot\boldsymbol{h}(t'))$ and $\delta(NiR - \boldsymbol{h}(t)\cdot\hat{\boldsymbol{h}}(t'))$, introducing $\hat{C}$ and $\hat{R}$ as conjugate functions. After this is all said and done, show that the integrals over the N-sites decouple, yielding

$$Z = \int DCD\hat{C}DRD\hat{R} \exp\left(-N\mathcal{S}[C,\hat{C},R,\hat{R}] \right) \tag{51}$$

and provide the formula for the DMFT action $\mathcal{S} = \mathcal{O}_N(1)$ (it is ok to drop additive constants to $\mathcal{S}$). At large N, the path integral concentrates at a saddle point where $\frac{\partial \mathcal{S}}{\partial Q} = 0$ for $Q \in \{C, \hat{C}, R, \hat{R}\}$. Show that the saddle point equations reproduce the DMFT equations obtained by the cavity method. Argue that fluctuations around the saddle point can be computed with second derivatives $\frac{\partial^2 \mathcal{S}}{\partial Q \partial Q'}$ for $Q, Q' \in \{C, \hat{C}, R, \hat{R}\}$ and that these fluctuations have size $1/\sqrt{N}$.
Hint: see reference [17].

6 Linear and Random Feature Regression

We next turn to a learning problem. Let $\boldsymbol{\Psi} \in \mathbb{R}^{P\times N}$ be a random data matrix with $P = \alpha N$ data points in N dimensions. Consider the teacher-student linear model

$$y_\mu = \frac{1}{\sqrt{N}}\boldsymbol{\beta}_\star \cdot \boldsymbol{\psi}_\mu + \varepsilon_\mu, \qquad \left\langle \varepsilon_\mu^2 \right\rangle = \sigma^2, \tag{52}$$

with predictor / student model

$$f = \frac{1}{\sqrt{N}}\boldsymbol{w} \cdot \boldsymbol{\psi}. \tag{53}$$

The features $\boldsymbol{\psi} \sim \mathcal{N}(0, \boldsymbol{I})$ are drawn from an isotropic Gaussian distribution.

Problem 11: Linear Regression Setup (Easy)

Show that the test MSE loss $\mathcal{L}_{\text{test}} = \left\langle (f-y)^2 \right\rangle$ can be expressed in terms of the weight error $\boldsymbol{h}(t) = \boldsymbol{\beta}_\star - \boldsymbol{w}(t)$ as

$$\mathcal{L}_{\text{test}}(t) = \frac{1}{N}|\boldsymbol{h}(t)|^2 + \sigma^2 \tag{54}$$

Show that the training loss $\mathcal{L}_{\text{train}} = \frac{1}{P}\sum_{\mu=1}^{P}\left[f_\mu - y_\mu\right]^2$ can be expressed as

$$\mathcal{L}_{\text{train}}(t) = \frac{1}{P}|\boldsymbol{\Delta}(t)|^2 \ , \ \boldsymbol{\Delta}(t) \equiv \frac{1}{\sqrt{N}}\boldsymbol{\Psi}\boldsymbol{h}(t) + \boldsymbol{\epsilon} \in \mathbb{R}^P \tag{55}$$

Now, express the gradient flow on $\mathcal{L}_{\text{train}}$ as

$$\frac{d}{dt}\boldsymbol{h}(t) = -\frac{N}{2}\partial_{\boldsymbol{h}}\mathcal{L}_{\text{train}} = -\left(\frac{1}{P}\boldsymbol{\Psi}^\top\boldsymbol{\Psi}\right)\boldsymbol{h}(t) - \frac{\sqrt{N}}{P}\boldsymbol{\Psi}^\top\boldsymbol{\epsilon} = -\frac{\sqrt{N}}{P}\boldsymbol{\Psi}^\top\boldsymbol{\Delta}(t) \tag{56}$$

Under the assumption that $P = \alpha N$, argue that this can be expressed as a two-variable system

$$\partial_t\boldsymbol{h}(t) = -\frac{\sqrt{N}}{P}\boldsymbol{\Psi}^\top\boldsymbol{\Delta}(t) \ , \ \boldsymbol{\Delta}(t) = \frac{1}{\sqrt{N}}\boldsymbol{\Psi}\boldsymbol{h}(t) + \boldsymbol{\epsilon} \tag{57}$$

We see from the previous exercise that the dynamics are governed by a Wishart matrix $\frac{1}{P}\boldsymbol{\Psi}^\top\boldsymbol{\Psi}$.

6.1 Bipartite DMFT

By introducing the intermediate variable $\boldsymbol{\Delta}$ we were able to **linearize each expression in the random matrix**. We are now in a position to perform a two variable cavity argument [29, 17]. To do so, introduce sources $\boldsymbol{j}_h(t)$ and $\boldsymbol{j}_\Delta(t)$ to the right hand sides of the defining equations for $\boldsymbol{h}(t)$ and $\boldsymbol{\Delta}(t)$.

Problem 12: Bipartite Cavity (Medium)

Perform a cavity derivation for the typical behavior of $h(t)$ and $\Delta(t)$. To start, imagine adding a single data point $\boldsymbol{\psi}_0$, which comes with its own error $\Delta_0(t)$. Argue that this inclusion of this new data point induces a small change to the error vectors $\boldsymbol{h}(t) \to \tilde{\boldsymbol{h}}(t)$. Show that the dominant perturbation has the form

$$\tilde{\boldsymbol{h}}(t) \simeq \boldsymbol{h}(t) - \frac{\sqrt{N}}{P} \int dt' \frac{\partial \boldsymbol{h}(t)}{\partial \boldsymbol{j}(t')^\top} \boldsymbol{\psi}_0 \Delta_0(t') \tag{58}$$

Now use these perturbed

$$\Delta_0(t) = \frac{1}{\sqrt{N}} \boldsymbol{\psi}_0 \cdot \tilde{\boldsymbol{h}}(t) + \varepsilon \simeq \frac{1}{\sqrt{N}} \boldsymbol{\psi}_0 \cdot \boldsymbol{h}(t) - \frac{1}{\alpha} \int dt' \left[\frac{1}{N} \boldsymbol{\psi}_0^\top \frac{\partial \boldsymbol{h}(t)}{\partial \boldsymbol{j}_h(t)^\top} \boldsymbol{\psi}_0 \right] \Delta_0(t') + \varepsilon_0 \tag{59}$$

Argue that the first term obeys a central limit theorem with covariance $C_h(t,t') = \frac{1}{N} \boldsymbol{h}(t) \cdot \boldsymbol{h}(t')$ and the second term concentrates to $\frac{1}{\alpha} R_h(t,t')$. This gives the defining single site equation for $\Delta(t)$. Now perform the same analysis where you add a new feature dimension $h_0(t)$ and new vector across all P training points $\boldsymbol{\psi}^0 \in \mathbb{R}^P$. Perform the same analysis to derive the single site equations for $h_0(t)$ which have the form

$$\frac{d}{dt} h_0(t) \simeq -\frac{\sqrt{N}}{P} \boldsymbol{\psi}^0 \cdot \boldsymbol{\Delta}(t) - \int dt' \boldsymbol{\psi}^0 \left[\frac{1}{P} (\boldsymbol{\psi}^0)^\top \frac{\partial \boldsymbol{\Delta}(t)}{\partial \boldsymbol{j}_\Delta(t')^\top} \boldsymbol{\psi}^0 \right] h_0(t') \tag{60}$$

Argue that the first term is Gaussian and the second concentrates.

The two-sided cavity argument gives the effective process

$$\dot{h}(t) = u_h(t) - \int^t \mathrm{d}t' \, R_\Delta(t,t') h(t'), \tag{61}$$

$$\Delta(t) = u_\Delta(t) - \frac{1}{\alpha} \int^t \mathrm{d}t' \, R_h(t,t') \Delta(t') + \varepsilon, \tag{62}$$

with Gaussian fields

$$u_h \sim \mathrm{GP}(0, \alpha^{-1} C_\Delta), \qquad u_\Delta \sim \mathrm{GP}(0, C_h). \tag{63}$$

The self-consistency equations are

$$C_h(t,t') = \left\langle h(t) h(t') \right\rangle, \qquad C_\Delta(t,t') = \left\langle \Delta(t) \Delta(t') \right\rangle, \tag{64}$$

$$R_h(t,t') = \left\langle \frac{\partial h(t)}{\partial u_h(t')} \right\rangle, \qquad R_\Delta(t,t') = \left\langle \frac{\partial \Delta(t)}{\partial u_\Delta(t')} \right\rangle \tag{65}$$

The losses are diagonal correlations:

$$\mathcal{L}_{\text{test}}(t) = C_h(t,t) + \sigma^2, \qquad \mathcal{L}_{\text{train}}(t) = C_\Delta(t,t). \tag{66}$$

6.2 Marchenko–Pastur Law

Problem 13: Marchenko-Pastur Law (Medium)

The response R_h is the response of the linear system governed by $\boldsymbol{M} = \frac{1}{P}\boldsymbol{\Psi}^\top\boldsymbol{\Psi}$. Therefore its transform is the Stieltjes transform of the Wishart spectrum. Solve for the response functions in the Fourier domain $\mathcal{R}_h(\omega)$ and $\mathcal{R}_\Delta(\omega)$ and show that they have the form

$$\mathcal{R}_h(\omega) = \frac{1}{i\omega + \mathcal{R}_\Delta(\omega)} \ , \ \mathcal{R}_\Delta(\omega) = \frac{1}{1 + \alpha^{-1}\mathcal{R}_h(\omega)} \tag{67}$$

Use this result to derive the Marchenko-Pastur distribution using Stieltjes inversion

$$\rho_{\rm MP}(\lambda) = \frac{\alpha}{2\pi\lambda}\sqrt{[4\lambda\alpha^{-1} - (\alpha^{-1} - 1 + \lambda)^2]_+} + [1-\alpha]_+\delta(\lambda). \tag{68}$$

The bulk support is

$$\lambda \in \left[(1-\alpha^{-1/2})^2, (1+\alpha^{-1/2})^2\right]. \tag{69}$$

7 Random feature models and non-Hermitian dynamics

Random feature models introduce an additional random matrix $\boldsymbol{A}$ for the student's features $\boldsymbol{A\psi}$ [14, 30]. Consider a simple structured random feature model

$$f = \frac{1}{\sqrt{N}}\boldsymbol{w}^\top\boldsymbol{A\psi}(t) \ , \ y = \boldsymbol{w}^\star\cdot\boldsymbol{\psi} \tag{70}$$

$$\langle\psi_k\psi_\ell\rangle = \delta_{k\ell}\lambda_k \ , \ \sum_{k=1}^{\infty}\lambda_k < \infty \tag{71}$$

where the random feature matrix $\boldsymbol{A}$ has iid entries. Show that the gradient flow dynamics on the weights $\boldsymbol{w}(t)$ induces the following dynamics on $\boldsymbol{h}(t) \equiv \boldsymbol{w}_\star - \frac{1}{\sqrt{N}}\boldsymbol{A}^\top\boldsymbol{w}(t)$

$$\frac{d}{dt}\boldsymbol{h}(t) = -\left(\frac{1}{N}\boldsymbol{A}^\top\boldsymbol{A}\right)\left(\frac{1}{P}\boldsymbol{\Psi}^\top\boldsymbol{\Psi}\right)\boldsymbol{h}(t) \tag{72}$$

Problem 14: Spectrum of the Matrix (Medium)

The matrix $\boldsymbol{M} = \left(\frac{1}{N}\boldsymbol{A}^\top\boldsymbol{A}\right)\left(\frac{1}{P}\boldsymbol{\Psi}^\top\boldsymbol{\Psi}\right)$ is non-Hermitian. Despite being non-Hermitian, argue that $\boldsymbol{M}$ has a real and non-negative spectrum. This means that the response function $R(t,t')$ should be computed from this non-negative spectrum.
Hint: Show that $\boldsymbol{M}$ has the same eigenvalues as $\tilde{\boldsymbol{M}} = \frac{1}{N}\boldsymbol{A}\left(\frac{1}{P}\boldsymbol{\Psi}^\top\boldsymbol{\Psi}\right)\boldsymbol{A}^\top$ which is Hermitian and positive semidefinite.

The above exercise demonstrates that eigenspectrum of $\boldsymbol{M}$ is non-negative. Naively (if we thought of $\boldsymbol{M}$ as Hermitian) this would imply that the test loss would be monotone. **However, this is not true in general! In fact, when $N = P$, the loss in this model will diverge as**

$$\mathcal{L}_{\rm test} = \frac{1}{N}|\boldsymbol{h}(t)|^2 \sim \mathcal{O}(t^{1/2}) \ , \ N = P \tag{73}$$

The key issue is that $\boldsymbol{M}$ is non-normal and can amplify fluctuations [17].

Problem 15: DMFT for the Random Feature Model (Medium)

Using either a cavity approach or MSR, work out the DMFT equations for the above dynamical system. Show that the correlation function $C(t,t')$ can differ from the response $C(t,t') \neq R(t+t')$. Hint: Define the original $\boldsymbol{h}(t) = \boldsymbol{h}_0(t)$ and introduce intermediate fields $\boldsymbol{h}_1(t) = \boldsymbol{\Psi}\boldsymbol{h}_0(t)$ and $\boldsymbol{h}_2 = \frac{1}{P}\boldsymbol{\Psi}^\top \boldsymbol{h}_1(t)$ and $\boldsymbol{h}_3(t) = \boldsymbol{A}\boldsymbol{h}_2(t)$ and $\boldsymbol{h}_4(t) = \frac{1}{N}\boldsymbol{A}^\top \boldsymbol{h}_3(t)$ with $\frac{d}{dt}\boldsymbol{h}_0(t) = -\boldsymbol{h}_4(t)$.

Remark 7.1 (A useful warning)**.** For learning dynamics,

$$\text{spectrum of } \boldsymbol{M} \neq \text{loss curve} \tag{74}$$

in general. The equivalence is special to sufficiently normal or Hermitian systems. DMFT is useful because it tracks both $R(t,t')$ and $C(t,t')$. The loss actually is governed by $C(t,t')$.

8 Feature Learning in Wide and Deep Neural Networks

The preceding sections focused on linear dynamics with fixed random matrices. Neural network training is more complicated because the features themselves evolve. In a hidden layer ℓ, write schematic forward features and backward signals as

$$f_\mu = \frac{1}{N\gamma}\boldsymbol{w}^L \cdot \phi(\boldsymbol{h}_\mu^L) \ , \ \boldsymbol{h}_\mu^{\ell+1} = \frac{1}{\sqrt{N}}\boldsymbol{W}^\ell \phi(\boldsymbol{h}_\mu^\ell) \ , \ \boldsymbol{h}_\mu^1 = \frac{1}{\sqrt{D}}\boldsymbol{W}^0 \boldsymbol{x}_\mu \tag{75}$$

We will train the network on a loss function $\mathcal{L}$ with the following learning rates

$$\eta = \eta_0 \gamma^2 N \tag{76}$$

where η_0 and γ are independent of N. This scaling is known as mean field or maximum update scaling [31, 3, 8]. The factor γ can be interpreted as follows

$$\begin{aligned} &\gamma \approx 0 \implies \text{Lazy Regime (Kernels do not deviate from their Initial Values)} \\ &\gamma \text{ large} \implies \text{Significant Evolution of the Features and Kernels} \end{aligned}$$

Gradient descent updates a weight matrix by outer products

$$\boldsymbol{W}^\ell(t) = \boldsymbol{W}^\ell(0) + \frac{\eta_0\gamma}{\sqrt{N}} \sum_{t<T} \sum_{\mu \in B_t} \Delta_\mu(t) \boldsymbol{g}_\mu^{\ell+1}(t) \boldsymbol{\phi}_\mu^\ell(t)^\top, \tag{77}$$

where $\boldsymbol{\phi}_\mu^\ell(t) \equiv \phi(\boldsymbol{h}_\mu^\ell(t))$ and Δ and $\boldsymbol{g}$ are defined as

$$\Delta_\mu(t) \equiv -\frac{\partial \mathcal{L}}{\partial f_\mu} \ , \ \boldsymbol{g}_\mu^\ell \equiv N\gamma \frac{\partial f_\mu}{\partial \boldsymbol{h}_\mu^\ell} \simeq \text{gradient signal from back-propagation} \tag{78}$$

The $\boldsymbol{g}$ features satisfy the following recursion relation

$$\boldsymbol{g}_\mu^\ell(t) = \dot{\phi}(\boldsymbol{h}_\mu^\ell(t)) \odot \underbrace{\left[\frac{1}{\sqrt{N}}\boldsymbol{W}^\ell(t)^\top \boldsymbol{g}_\mu^{\ell+1}(t)\right]}_{\equiv \boldsymbol{z}_\mu^\ell(t)} \ , \ \boldsymbol{g}_\mu^L(t) = \dot{\phi}(\boldsymbol{h}_\mu^L(t)) \odot \boldsymbol{w}^L(t) \tag{79}$$

Problem 16: Deriving Single Site Equations for Infinite Width Networks (Medium)

We would like to describe typical case training dynamics in the limit as $N \to \infty$. Expand out the weight dynamics and compute the forward pass and gradient variables

$$\boldsymbol{h}_\mu^{\ell+1}(t) = \underbrace{\frac{1}{\sqrt{N}}\boldsymbol{W}^\ell(0)\phi(\boldsymbol{h}_\mu^\ell(t))}_{\boldsymbol{\chi}_\mu^{\ell+1}(t)} + \eta_0\gamma \sum_{t'<t}\sum_\nu \Delta_\nu(t')\ \boldsymbol{g}_\nu^{\ell+1}(t')\ \underbrace{\left(\frac{1}{N}\phi(\boldsymbol{h}_\nu^\ell(t'))\cdot\phi(\boldsymbol{h}_\mu^\ell(t))\right)}_{C_{\mu\nu}^{\phi,\ell}(t,t')} \tag{80}$$

$$\boldsymbol{z}_\mu^\ell(t) = \underbrace{\frac{1}{\sqrt{N}}\boldsymbol{W}^\ell(0)^\top\boldsymbol{g}_\mu^{\ell+1}(t)}_{\boldsymbol{\xi}_\mu^\ell(t)} + \eta_0\gamma \sum_{t'<t}\sum_\nu \Delta_\nu(t')\ \phi(\boldsymbol{h}_\nu^\ell(t))\ \underbrace{\left(\frac{1}{N}\boldsymbol{g}_\nu^\ell(t')\cdot\boldsymbol{g}_\mu^\ell(t)\right)}_{C_{\mu\nu}^{g,\ell+1}(t,t')} \tag{81}$$

Argue that C^ϕ and C^g are correlation functions that should concentrate as $N \to \infty$ under the DMFT. Argue that, if the C^ϕ and C^g are deterministic, then the only random variables that depend on $\boldsymbol{W}^\ell(0)$ which appear in the dynamics are the random fields $\boldsymbol{\chi}_\mu^{\ell+1}(t)$ and $\boldsymbol{\xi}_\mu^\ell(t)$. Using either a two-step cavity argument or the MSR approach, show that as $N \to \infty$ these variables can be expressed as single site stochastic processes with **colored noise + response** decomposition

$$\chi_\mu^{\ell+1}(t) = u_\mu^{\ell+1}(t) + \sum_{t'<t}\sum_\nu R_{\mu\nu}^{\phi,\ell}(t,t')g_\nu^{\ell+1}(t')\ ,\ u_\mu^{\ell+1}(t) \sim \mathcal{N}(0, \boldsymbol{C}^{\phi,\ell}) \tag{82}$$

$$\xi_\mu^\ell(t) = r_\mu^\ell(t) + \sum_{t'<t}\sum_\nu R_{\mu\nu}^{g,\ell+1}(t,t')\phi(h_\nu^\ell(t'))\ ,\ r_\mu^\ell(t) \sim \mathcal{N}(0, \boldsymbol{C}^{g,\ell+1}) \tag{83}$$

where the response functions $R^{\phi,\ell}$ and $R^{g,\ell}$ are defined as

$$R_{\mu\nu}^{\phi,\ell}(t,t') = \left\langle \frac{\partial\phi(h_\mu^\ell(t))}{\partial r_\nu^\ell(t')} \right\rangle\ ,\ R_{\mu\nu}^{g,\ell}(t,t') = \left\langle \frac{\partial g_\mu^\ell(t)}{\partial u_\nu^\ell(t')} \right\rangle \tag{84}$$

Argue that χ, ξ, h, z are generally non-Gaussian. Show that these are Gaussian in the special case of linear networks $\phi(h) = h$.

The output of the network can be computed following a similar procedure

Problem 17: Deriving Output Dynamics for Infinite Width Networks (Medium)

The output dynamics of the network can be computed in a similar manner. Show that the output follows

$$f_\mu(t) = \sum_{t'<t} \sum_\nu R^{\phi,L}_{\mu\nu}(t,t') + \sum_{t'<t} C^{\phi,L}_{\mu\nu}(t,t')\Delta_\nu(t') \tag{85}$$

Optional: show that the $\gamma \to 0$ limit results in the function f evolving according to the *initial neural tangent kernel*

$$\gamma \to 0 \implies f_\mu(t) = \sum_{t'<t} \sum_\nu \left[\sum_{\ell=0}^{L} C^{g,\ell+1}_{\mu\nu}(0,0) C^{\phi,\ell}_{\mu\nu}(0,0) \right] \Delta_\nu(t') \tag{86}$$

where the base cases are defined as $C^{\phi,0} = \frac{1}{D}\boldsymbol{x}\cdot\boldsymbol{x}'$ and $G^{L+1}_{\mu\nu} = 1$.

9 Spiked Wigner and Classic BBP Transition

Random matrices illustrate a simple example of competing effects of a high dimensional random matrix and a structured low dimensional component [32, 33]. To illustrate how we can compute the behavior of the outlier eigenvalues of such spiked matrices, let us begin with a Hermitian GOE random matrix with S symmetric spikes

$$\boldsymbol{M} = \underbrace{\boldsymbol{M}_0}_{\text{GOE}} + \underbrace{\frac{\gamma}{N}\sum_{a=1}^{S} \boldsymbol{v}_a \boldsymbol{v}_a^\top}_{\text{rank}\leq S}, \quad \Sigma^v_{ab} \equiv \frac{1}{N}\boldsymbol{v}_a \cdot \boldsymbol{v}_b \tag{87}$$

where we assumed that $\boldsymbol{M}_0$ is GOE and the vectors $\boldsymbol{v}$ have overlap matrix $\boldsymbol{\Sigma}^v \in \mathbb{R}^{S\times S}$. As before we can consider the flow on a probe variable $\boldsymbol{h}(t)$

$$\frac{d}{dt}\boldsymbol{h}(t) = \boldsymbol{M}\boldsymbol{h}(t) = \boldsymbol{M}_0\boldsymbol{h}(t) + \gamma \sum_{a=1}^{S} \boldsymbol{v}_a \underbrace{\left(\frac{1}{N}\boldsymbol{h}(t)\cdot\boldsymbol{v}_a\right)}_{\equiv C^{hv}_a(t)} \tag{88}$$

Problem 18: DMFT equations for Spiked Matrices

First, argue that the large time asymptotic dynamics for the linear system

$$\boldsymbol{h}(t) \sim \sum_a e^{z_a t} h_a(0) \ , \ t \to \infty \tag{89}$$

can reveal candidate outlier eigenvalues z_a of $\boldsymbol{M}$ where $h_a(0)$ represents the initial condition $\boldsymbol{h}(0)$ in the eigenbasis of $\boldsymbol{M}$. We will try to find such candidates. Show that averaging over the GOE matrix gives the following DMFT equation

$$\frac{d}{dt}h(t) = \xi(t) + \int dt' R(t,t')h(t') + \gamma \sum_a v_a C_a^{hv}(t) \tag{90}$$

Taking a Laplace transform from time variable t to Laplace variable z, show that you obtain $\boldsymbol{C}^{hv}(z) \in \mathbb{R}^S$ which represents the alignment and let $[\boldsymbol{C}_0^{hv}]_a = \langle h(0) v_a \rangle$ represent the alignment of our initial condition with each spike, then

$$\boldsymbol{C}^{hv}(z) = \left(z - R(z) - \gamma \boldsymbol{\Sigma}^v\right)^{-1} \boldsymbol{C}_0^{hv} \ , \ zR(z) = 1 + R(z)^2 \tag{91}$$

Argue that the poles of the above formula correspond to potential outliers. These can be interpreted as a secular equation on a $S \times S$ matrix

$$\det\left(z - R(z) - \gamma \boldsymbol{\Sigma}^v\right) = 0 \tag{92}$$

We see that the effect of the GOE matrix is to add the nonlinear $R(z)$ term to the usual eigenvalue problem for $\boldsymbol{\Sigma}^v$. Diagonalize $\boldsymbol{\Sigma}^v$ and argue that since it has real eigenvalues λ_a that we have S potential candidates $z_a = R(z_a) + \gamma \lambda_a$. Argue that a root of the above is only a genuine outlier of the matrix $\boldsymbol{M}$ if it surpasses the edge of the bulk GOE eigenvalue density $z_a > \lambda_{\max}(\text{GOE}) = 2$.

Let's work out an explicit formula for each candidate outlier

$$z_a - R(z_a) = \frac{1}{R(z_a)} = \gamma \lambda_a \implies z_a = \gamma \lambda_a + \frac{1}{\gamma \lambda_a} \tag{93}$$

Since $\gamma \lambda_a \geq 0$ this quantity is always postive. This candidate outlier is a **true outlier** only if it exceeds the edge of the bulk eigenvalue density $z_a > 2$. Thus the eigenvalues of $\boldsymbol{M}$ have $\mathcal{O}(N)$ eigenvalues that obey the GOE density and up to S spikes

$$\text{spec } \boldsymbol{M} = \text{GOE Spectrum} + \left\{ \gamma \lambda_a + \frac{1}{\gamma \lambda_a} : \gamma \lambda_a > 1 \right\} \tag{94}$$

The fact that the spikes are invisible if their signal strength is smaller than a threshold is known as the Baik-Ben Arous-Péché (BBP) phase transition [32, 33].

10 Feature Learning as Spiked Matrices

At initialization, the initial weights of a neural network have singular values that are follow a Marchenko–Pastur law. Training adds structured updates. Equation (77) suggests a spiked matrix model

$$\boldsymbol{W}(S) = \boldsymbol{W}(0) + \frac{\eta\gamma}{\sqrt{N}} \sum_{t=1}^{S} \Delta(t) \boldsymbol{g}(t) \boldsymbol{\phi}(t)^\top. \tag{95}$$

If the number of spikes $S = \mathcal{O}(1)$ as $N \to \infty$, this is a finite-rank perturbation of a random matrix. However, as we saw in previous sections this setting is quite different than the classic BBP setting.

Assumptions of the Classic BBP Analysis The BBP phase transition in the previous section used the fact that

1. Spike directions are independent of the random bulk.
2. Spike directions are generated exogenously, they do not arise from dynamical variables that depend on $\boldsymbol{W}(0)$.

Training-induced dependent spikes In trained neural networks, the spike directions are not independent of the random initialization. They are **generated endogenously by the trajectory**, and the trajectory depends on $W(0)$. To expose the dependence, define fields

$$\boldsymbol{\chi}(t) = \frac{1}{\sqrt{N}}\boldsymbol{W}(0)\boldsymbol{\phi}(t), \qquad \boldsymbol{\xi}(t) = \frac{1}{\sqrt{N}}\boldsymbol{W}(0)^\top \boldsymbol{g}(t). \tag{96}$$

In many large-width limits, the spike coordinates become independent across sites but remain statistically coupled to these fields. A useful structural form from our DMFT equations is

$$\phi_i(t) = \phi_t\left(\{\xi_i(s)\}_{s\le t}\right), \qquad g_i(t) = g_t\left(\{\chi_i(s)\}_{s\le t}\right). \tag{97}$$

Thus the **spike is a function of the random bulk through a history of projections** onto the features themselves. This invalidates the independence and exogeneity of the classical BBP.

11 Two-level DMFT for dependent spiked matrices

The above problem motivates the study of **endogenously spiked random matrices** [34]. Before describing how to apply this to a neural network, we will study a simpler case that is more revealing and where we can compare to the exogenous BBP transition.

11.1 Evolving Wigner Matrix with Endogenously Generated Spikes

Let's look at a simple modification of the GOE example where the matrix $\boldsymbol{M}$ starts as a GOE matrix but accumulates spikes that are generated through endogenous dynamics.

$$\begin{aligned} &\text{Matrix dynamics: } \boldsymbol{M}_{n+1} = \boldsymbol{M}_n + \frac{\gamma}{N}\boldsymbol{v}_n\boldsymbol{v}_n^\top \qquad \text{Spike dynamics: } \boldsymbol{v}_{n+1} = \boldsymbol{v}_n - \eta \boldsymbol{M}_n \boldsymbol{v}_n \\ &\text{Initial Condition: } \boldsymbol{M}_0 = \text{GOE} \end{aligned} \tag{98}$$

The idea here is that the matrix $\boldsymbol{M}$ generates a dynamical system over spike states $\boldsymbol{v}_n$ and that these spike states are used to update $\boldsymbol{M}$. The fact that $\boldsymbol{v}$ is evolving with a difference equation over steps will generate potentially strong correlations across steps. This is similar in spirit to the deep network case where the random matrices $\boldsymbol{W}$ generate features and gradients that are used to update the weights, which then change the features and gradients. We consider the matrix $\boldsymbol{M}_S$ after S spike steps. Expanding out the weight dynamics, we find the following

$$\text{Level 1 Dynamics: } \boldsymbol{v}_{n+1} = \boldsymbol{v}_n - \eta \boldsymbol{M}_0 \boldsymbol{v}_n - \eta \sum_{m<n} \Sigma^v_{nm}\boldsymbol{v}_m \ , \ \Sigma^v_{nm} \equiv \frac{1}{N}\boldsymbol{v}_n \cdot \boldsymbol{v}_m \tag{99}$$

We additionally introduce the usual dynamics on the probe vector $\boldsymbol{h}(t)$

$$\text{Level 2 Dynamics: } \frac{d}{dt}\boldsymbol{h}(t) = \boldsymbol{M}_S\ \boldsymbol{h}(t) = \boldsymbol{M}_0\boldsymbol{h}(t) + \gamma \sum_{n=1}^{S} C_n^{hv}(t)\boldsymbol{v}_n \tag{100}$$

Our goal is to now characterize the **typical case behavior** of this joint two-level system as $N \to \infty$. We will first do so in a naive way and then in a way that is correct and compare the two results.

Problem 19: Naive Approach to the BBP Transtion (Easy)

First, we will make the approximation that the matrix $\boldsymbol{M}_0$ that appears in the level 2 dynamics is a **different, uniquely independent GOE matrix** $\tilde{\boldsymbol{M}}$ than the random GOE matrix $\boldsymbol{M}_0$ appearing in the level one dynamics. Argue that we can now treat these levels as separate dynamical systems. Perform the DMFT average over $\boldsymbol{M}_0$ and derive the DMFT process for the spikes

$$v_{n+1} = v_n - \eta\chi_n - \eta \sum_m R^v_{nm} v_m - \eta\gamma \sum_{m<n} \Sigma^v_{nm} v_m\ ,\ R^v_{nm} = \left\langle \frac{\partial v_n}{\partial \chi_m} \right\rangle\ ,\ \Sigma^v_{nm} = \langle v_n v_m \rangle \tag{101}$$

Suppose that you have solved the above equations for $\boldsymbol{\Sigma}^v \in \mathbb{R}^{S\times S}$ argue that the level 2 dynamics is a specific instance of our previous exogenously spiked BBP calculation and that the candidate outliers z of $\boldsymbol{M}_S$ satisfy the secular equation

$$\det\left(z - R(z) - \gamma\boldsymbol{\Sigma}^v\right) = 0 \tag{102}$$

This would be nice if true! Unfortunately, the fact that both level one and level 2 share the same matrix $\boldsymbol{M}_0$ generates additional terms that alter the BBP criterion. Let's now try to identify the correct rule.

Problem 20: Accurate Two Level DMFT (Hard)

Now do it correctly, using the fact that the same matrix $\boldsymbol{M}_0$ in the probe dynamics for $\boldsymbol{h}(t)$ is also driving the dynamics for $\boldsymbol{v}_n$. Show that the v dynamics satisfy the same DMFT as above

$$v_{n+1} = v_n - \eta\chi_n - \eta\sum_m R^v_{nm} v_m - \eta\gamma \sum_{m<n} \Sigma^v_{nm} v_m \ , \ R^v_{nm} = \left\langle \frac{\partial v_n}{\partial \chi_m} \right\rangle \ , \ \Sigma^v_{nm} = \langle v_n v_m \rangle \, . \tag{103}$$

However, the dynamics for the probe variable $\boldsymbol{h}(t)$ are modified with additional response and cross correlation

$$\frac{d}{dt}h(t) = \xi(t) + \int dt' R(t,t')h(t') + \sum_n R^{h\chi}_n(t) v_n + \gamma \sum_n C^{hv}_n(t) v_n \ , \ \langle \xi(t)\chi_n \rangle = C^{hv}_n(t) \tag{104}$$

Show that the Laplace transformed equations give

$$(z - R(z))\boldsymbol{C}^{hv}(z) = \boldsymbol{R}^v \boldsymbol{C}^{hv}(z) + \boldsymbol{\Sigma}^v \boldsymbol{R}^{h\chi}(z) + \gamma \boldsymbol{\Sigma}^v \boldsymbol{C}^{hv}(z) + \boldsymbol{C}^{vh}_0 \tag{105}$$

$$(z - R(z))\boldsymbol{R}^{h\chi}(z) = (\boldsymbol{R}^v)^\top \boldsymbol{R}^{h\chi}(z) + \gamma (\boldsymbol{R}^v)^\top \boldsymbol{C}^{hv}(z) \tag{106}$$

Argue that the new secular equation for candidate outliers has the form of a zero-determinant condition on $2S \times 2S$ matrix that can be populated by the order parameters $\boldsymbol{\Sigma}^v$ and $\boldsymbol{R}^v$ from the $\boldsymbol{v}$ dynamics

$$\det \begin{bmatrix} z - R(z) - \boldsymbol{R}^v - \gamma \boldsymbol{\Sigma}^v & -\boldsymbol{\Sigma}^v \\ -\gamma (\boldsymbol{R}^v)^\top & z - R(z) - (\boldsymbol{R}^v)^\top \end{bmatrix} = 0 \tag{107}$$

Hint: Use Novikov/Stein for $\langle \xi(t)\chi_n \rangle$.

11.2 Dynamics of Spiked Matrices in μP Infinite Width Networks

Key reference

- Lauditi, Pehlevan*, Bordelon* 2026 [34]

We can use the same technology to compute the outlier dynamics and BBP transitions for infinite width feature learning networks.

Problem 20: Weight Singular Values in Infinite Width Networks (Very Hard)

Start from the defining equations on a super-index $a = (\mu, t)$ and $b = (\nu, s)$.

$$\boldsymbol{\chi}_a = \frac{1}{\sqrt{N}} \boldsymbol{W}(0) \boldsymbol{\phi}_a \ , \ \boldsymbol{\xi}_a = \frac{1}{\sqrt{N}} \boldsymbol{W}(0)^\top \boldsymbol{g}_a \tag{108}$$

coupled with the rule that $\boldsymbol{\phi}_a = \phi_a(\{\boldsymbol{\xi}_b\})$ and $\boldsymbol{g}_a = g_a(\{\boldsymbol{\chi}_b\})$ where $\phi_a(\cdot)$ and $g_a(\cdot)$ are functions that act **elementwise**. Without loss of generality, we can assume that $\boldsymbol{g}_a$ absorbs a factor of Δ_a. Define the weights after S spikes as

$$\boldsymbol{W}_S = \boldsymbol{W}(0) + \frac{\eta\gamma}{\sqrt{N}} \sum_{a=1}^{S} \boldsymbol{g}_a \boldsymbol{\phi}_a^\top \tag{109}$$

Introduce the following flow on a probe variable $\boldsymbol{\psi}$ to track the eigenvalues of $\boldsymbol{M} = \frac{1}{N} \boldsymbol{W}_S^\top \boldsymbol{W}_S$

$$\partial_t \boldsymbol{\psi}(t) = \frac{1}{N} \boldsymbol{W}_S^\top \boldsymbol{W}_S \ \boldsymbol{\psi}(t) = \frac{1}{\sqrt{N}} \boldsymbol{W}_S^\top \boldsymbol{\psi}_1(t) \ , \ \boldsymbol{\psi}_1(t) \equiv \frac{1}{\sqrt{N}} \boldsymbol{W}_S \boldsymbol{\psi}(t) \tag{110}$$

Perform the average over $\boldsymbol{W}(0)$ and argue that the outlier candidate condition is a secular equation over $4S \times 4S$ block matrix $\boldsymbol{A}(z)$. Provide the formula for $\det \boldsymbol{A}(z) = 0$ in terms of the correlation functions $\boldsymbol{C}^\phi, \boldsymbol{C}^g$ and the response functions $\boldsymbol{R}^\phi$ and $\boldsymbol{R}^g$ for the original DMFT.

$$\boldsymbol{A}(z) = \begin{bmatrix} \mathcal{G}(z)^{-1} \boldsymbol{I} & \boldsymbol{0} & -(\boldsymbol{R}^\phi + \boldsymbol{C}^\phi) & -\boldsymbol{C}^\phi \\ \boldsymbol{0} & \mathcal{G}(z)^{-1} \boldsymbol{I} & -(\boldsymbol{R}^\phi)^\top & -(\boldsymbol{R}^\phi)^\top \\ -(\boldsymbol{R}^g + \boldsymbol{C}^g) & -\boldsymbol{C}^g & (1 - \mathcal{G}(z)) \boldsymbol{I} & \boldsymbol{0} \\ -(\boldsymbol{R}^g)^\top & -(\boldsymbol{R}^g)^\top & \boldsymbol{0} & (1 - \mathcal{G}(z)) \boldsymbol{I} \end{bmatrix} \in \mathbb{R}^{4S \times 4S} \tag{111}$$

where $\mathcal{G}(z) = \frac{1}{2z} \left[z - \sqrt{z^2 - 4z} \right]$ is the resolvent of the Wishart matrix with aspect ratio $\alpha = 1$ evaluated for $z > 4$ (outside the bulk).

Optional: extend this result to the case where $\boldsymbol{W} \in \mathbb{R}^{N_1 \times N_0}$ is rectangular with aspect ratio $\alpha = N_1/N_0$. See the result of [34].

12 Summary and Main takeaways

1. DMFT reduces high-dimensional disordered dynamics to self-consistent stochastic processes.
2. The central order parameters are two-time correlations and responses.
3. For linear systems, the response is a random matrix resolvent.
4. In Hermitian systems, spectra often determine loss curves; in non-Hermitian systems, correlations and responses must both be tracked.
5. Random feature regression gives a clean learning-theoretic example where the Marchenko–Pastur law and train/test loss dynamics arise from DMFT.
6. Feature learning creates structured low-rank updates to hidden weights.
7. Trained-network spikes are generally dependent on the random initialization, unlike classical BBP spikes.

8. A two-level DMFT gives a generalized BBP criterion involving both correlations and responses.

9. Large-output regimes may require theories of bulk deformation rather than finite-rank outliers.

References

[1] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[2] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[3] 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.

[4] Nolan Dey, Bin Claire Zhang, Lorenzo Noci, Mufan Li, Blake Bordelon, Shane Bergsma, Cengiz Pehlevan, Boris Hanin, and Joel Hestness. Don't be lazy: Completep enables compute-efficient deep transformers. *arXiv preprint arXiv:2505.01618*, 2025.

[5] Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural networks. *Physical review letters*, 61(3):259, 1988.

[6] 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. *Advances in neural information processing systems*, 32, 2019.

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

[8] Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. *arXiv preprint arXiv:2205.09653*, 2022.

[9] 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.

[10] David G Clark, Blake Bordelon, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Structure, disorder, and dynamics in task-trained recurrent neural circuits. *bioRxiv*, pages 2026–03, 2026.

[11] Francesca Mignacco, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Dynamical mean-field theory for stochastic gradient descent in gaussian mixture classification. *Advances in Neural Information Processing Systems*, 33:9540–9550, 2020.

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

[13] Elisabeth Agoritsas, Giulio Biroli, Pierfrancesco Urbani, and Francesco Zamponi. Out-of-equilibrium dynamical mean-field equations for the perceptron model. *Journal of Physics A: Mathematical and Theoretical*, 51(8):085002, 2018.

[14] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In *International Conference on Machine Learning*, pages 4345–4382. PMLR, 2024.

[15] Francesca Mignacco and Pierfrancesco Urbani. The effective noise of stochastic gradient descent. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(8):083405, 2022.

[16] Cedric Gerbelot, Emanuele Troiani, Francesca Mignacco, Florent Krzakala, and Lenka Zdeborova. Rigorous dynamical mean field theory for stochastic gradient descent methods. *arXiv preprint arXiv:2210.06591*, 2022.

[17] Blake Bordelon and Cengiz Pehlevan. Disordered dynamics in high dimensions: Connections to random matrices and machine learning. *arXiv preprint arXiv:2601.01010*, 2026.

[18] Andrea Crisanti, Heinz Horner, and H-J Sommers. The spherical p-spin interaction spin-glass model: the dynamics. *Zeitschrift für Physik B Condensed Matter*, 92(2):257–271, 1993.

[19] Leticia F Cugliandolo and Jorge Kurchan. Analytical solution of the off-equilibrium dynamics of a long-range spin-glass model. *Physical Review Letters*, 71(1):173, 1993.

[20] Marc Mézard, Giorgio Parisi, Nicolas Sourlas, Gérard Toulouse, and Miguel Virasoro. Replica symmetry breaking and the nature of the spin glass phase. *Journal de Physique*, 45(5):843–854, 1984.

[21] C De Dominicis. Dynamics as a substitute for replicas in systems with quenched random impurities. *Physical Review B*, 18(9):4913, 1978.

[22] Samantha J Fournier, Alessandro Pacco, Valentina Ros, and Pierfrancesco Urbani. Non-reciprocal interactions and high-dimensional chaos: comparing dynamics and statistics of equilibria in a solvable model. *arXiv preprint arXiv:2503.20908*, 2025.

[23] Blake Bordelon and Cengiz Pehlevan. The influence of learning rule on representation dynamics in wide neural networks. *arXiv preprint arXiv:2210.02157*, 2022.

[24] A Crisanti and H Sompolinsky. Path integral approach to random neural networks. *Physical Review E*, 98(6):062120, 2018.

[25] Moritz Helias and David Dahmen. *Statistical field theory for neural networks*, volume 970. Springer, 2020.

[26] Radford M Neal. Priors for infinite networks. In *Bayesian learning for neural networks*, pages 29–53. Springer, 1996.

[27] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. *Advances in neural information processing systems*, 29, 2016.

[28] Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. *arXiv preprint arXiv:1611.01232*, 2016.

[29] David Clark and Haim Sompolinsky. Simplified derivations for high-dimensional convex learning problems. *SciPost Physics Lecture Notes*, page 105, 2025.

[30] Alexander Atanasov, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *arXiv preprint arXiv:2405.00592*, 2024.

[31] Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(11):113301, 2020.

[32] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. 2005.

[33] Jinho Baik and Jack W Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of multivariate analysis*, 97(6):1382–1408, 2006.

[34] Clarissa Lauditi, Cengiz Pehlevan, and Blake Bordelon. Spectral dynamics in deep networks: Feature learning, outlier escape, and learning rate transfer. *arXiv preprint arXiv:2605.07870*, 2026.