UNIVERSITY OF CALIFORNIA
RIVERSIDE

Topics in Quantum-Hall Physics, Game-Optimization and Generalization in Neural Networks

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Physics

by

Amartya Mitra

December 2021

Dissertation Committee:

Dr. Michael C. Mulligan, Chairperson
Dr. Bahram Mobasher
Dr. Samet Oymak

The Dissertation of Amartya Mitra is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I would like to thank my advisor Michael Mulligan for his guidance, mentorship, and help throughout my graduate career. My time in graduate school was complicated by a transition in my research focus from condensed matter physics to machine learning, which would not have been feasible without the unwavering support of Mike. I am hence sincerely thankful to him for allowing me to follow my inclination.

Secondly, I would like to thank Ioannis Mitliagkas and Aaron Courville at Mila, Montréal for generously hosting me as an intern for 2 years. A great amount of credit also goes to a lot of my peers at Mila especially Reyhane Askari Hemmat, Mohammad Pezeshki and Aristide Baratin at Mila, for patiently guiding me through every step of this transition.

I would also like to acknowledge my collaborators Guillaume Lajoie, Yoshua Bengio and my dissertation committee members, Dr. Bahram Mobasher and Dr. Samet Oymak, for their help and support. A great amount of appreciation also goes to my peer and great friend, Paromita Mukherjee for being all ears to my grumblings during grad school and providing helpful advice in navigating this meandering journey at various stages.

A sincere shout out also goes to the economy class seatings of Delta, United, American and WestJet airlines, for their hospitality while most of the work contained in this thesis was being performed.

Lastly, I am truly indebted to my parents, Gautom and Pritha, my brother Arjun, my girlfriend Debadrita and our cat Zelda, for their relentless support and encouragement during the course of my study, sans which this journey would have been immeasurably more difficult.

To my late advisor, Surajit Sengupta.

ABSTRACT OF THE DISSERTATION

Topics in Quantum-Hall Physics, Game-Optimization and Generalization in Neural Networks

by

Amartya Mitra

Doctor of Philosophy, Graduate Program in Physics
University of California, Riverside, December 2021
Dr. Michael C. Mulligan, Chairperson

This thesis contains research conducted on various topics in quantum Hall physics and deep learning theory. The first chapter studies a particular aspect of quantum Hall systems, namely their behavior around the $\nu = 1/2$ Landau level (LL) state. This work is motivated by the need to understand better this particular state in light of the two proposed distinct theoretical descriptions existing for the same. Specifically, we analyze quantum oscillations around the $\nu = 1/2$ LL state using one of the propositions to support the latter. The second and third chapters study two distinct domains in deep learning, multi and single-objective models. In particular, the second considers a specific type of multi-objective model, zero-sum games, to demonstrate existing issues in training such setups and develop an efficient optimization scheme. The final chapter involves studying a particular aspect of the generalization behavior of deep neural networks (DNNs). Specifically, it attempts to provide a theoretical framework to explain the recently observed phenomenon of "epoch-wise double descent" in such DNNs.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis presents research across a variety of different topics. Here we discuss each, in turn.

Fluctuations and magnetoresistance oscillations near the half-filled Landau level: We study theoretically the magnetoresistance oscillations near a half-filled lowest Landau level ($\nu = 1/2$) that result from the presence of a periodic one-dimensional electrostatic potential. We use the Dirac composite fermion theory of Son [Phys. Rev. X 5 031027 (2015)], where the $\nu = 1/2$ state is described by a $(2 + 1)$-dimensional theory of quantum electrodynamics. We extend previous work that studied these oscillations in the mean-field limit by considering the effects of gauge field fluctuations within a large flavor approximation. A self-consistent analysis of the resulting Schwinger–Dyson equations suggests that fluctuations dynamically generate a Chern-Simons term for the gauge field and a magnetic field-dependent mass for the Dirac composite fermions away from $\nu = 1/2$. We show how this mass results in a shift of the locations of the oscillation minima that improves the comparison with experiment [Kamburov

et. al., Phys. Rev. Lett. 113, 196801 (2014)]. The temperature-dependent amplitude of these oscillations may enable an alternative way to measure this mass. This amplitude may also help distinguish the Dirac and Halperin, Lee, and Read composite fermion theories of the half-filled Landau level. This research was conducted in collaboration with Michael Mulligan, and it was published as [242].

LEAD: Min-Max Optimization from a Physical Perspective: Adversarial formulations such as generative adversarial networks (GANs) have rekindled interest in two-player min-max games. A central obstacle in the optimization of such games is the rotational dynamics that hinder their convergence. Existing methods typically employ intuitive, carefully hand-designed mechanisms for controlling such rotations. This paper takes a novel approach to address this issue by casting min-max optimization as a physical system to motivate LEAD, an optimizer for min-max games. Next, using Lyapunov stability theory and spectral analysis, we study LEAD's convergence properties in continuous and discrete-time settings for a class of quadratic min-max games to demonstrate linear convergence to the Nash equilibrium. Finally, we empirically evaluate our method on synthetic setups and CIFAR-10 image generation to demonstrate improvements in GAN training. This research was conducted in collaboration with Reyhane Askari Hemmat, Guillaume Lajoie and Ioannis Mitliagkas, and it was preprinted as [146].

Double Descent Phenomena: A Tale of Multi-scale Feature Learning Dynamics: A key challenge in building theoretical foundations for deep learning is the complex optimization dynamics of large neural networks. Such dynamics result from high-dimensional interactions between the large number of parameters of such networks, thus leading to

non-trivial behaviors. In this regard, a particularly puzzling phenomenon is the "double descent" of the generalization error, where it undergoes two non-monotonous transitions, or descents, with increasing model complexity (model-wise) or training time (epoch-wise). While model-wise double descent has been a subject of extensive study of recent, the origins of the latter are much less clear. To bridge this gap, in this work, we leverage tools from statistical physics to study a simple teacher-student setup exhibiting epoch-wise double descent similar to deep neural networks. In this setting, we derive closed-form analytical expressions for the evolution of generalization error as a function of the training time. Crucially, this provides a new mechanistic explanation of epoch-wise double descent, suggesting that it can be attributed to different features being learned at different time scales. Summarily, **while a fast-learning feature is over-fitted, a slower-learning feature starts to fit, resulting in a non-monotonous generalization curve.** Finally, we validate our findings through simple numerical experiments where our theory accurately predicts empirical findings and remains consistent with observations in deep neural networks. This research was conducted in collaboration with Mohammad Pezeshki (Lead), Guillaume Lajoie and Yoshua Bengio.

Chapter 2

Fluctuations and magnetoresistance oscillations near the half-filled Landau level

In recent years, there has been a renewed debate about how effective descriptions of the non-Fermi liquid state at a half-filled lowest Landau level ($\nu = 1/2$) of the two-dimensional electron gas might realize an emergent Landau level particle-hole (PH) symmetry [374, 117], found in electrical Hall transport [327, 367, 277] and numerical [299, 111] experiments. The seminal theory of the half-filled Landau level of Halperin, Lee, and Read [140], which has received substantial experimental support [362], describes the $\nu = 1/2$ state in terms of non-relativistic composite fermions in an effective magnetic field that vanishes at half-filling (see [170, 97] for pedagogical introductions). However, the HLR theory appears to treat electrons and holes asymmetrically [185, 36]. For instance, it is naively unclear how composite

fermions in zero effective magnetic field might produce the Hall effect $\sigma_{xy}^{\rm cf} = -\frac{1}{4\pi}$ that PH symmetry requires [185]. (We use the convention $k_B = c = \hbar = e = 1$.)

Two lines of thought point towards a possible resolution. The first comes by way of an a priori different composite fermion theory, introduced by Son [334]. In this Dirac composite fermion theory, the half-filled Landau level is described by a $(2+1)$-dimensional theory of quantum electrodynamics in which PH symmetry is a manifest invariance. This theory is part of a larger web of $(2+1)$-dimensional quantum field theory dualities [324]. On the other hand, it has recently been shown that HLR mean-field theory *can* produce PH symmetric electrical response, if quenched disorder is properly included in the form of a precisely correlated random chemical potential and magnetic flux [351, 197, 195]. (Mean-field theory means that fluctuations of an emergent gauge field coupling to the composite fermion are ignored.) Furthermore, both composite fermion theories yield identical predictions for a number of observables in mean-field theory [334, 121, 351, 73, 196], e.g., thermopower at half-filling and magnetoroton spectra away from half-filling. These results suggest that the HLR and Dirac composite fermion theories may belong to the same universality class.

To what extent do these results extend beyond the mean-field approximation? How do alternative experimental probes constrain the description of the $\nu = 1/2$ state? The aim of this paper is to address both of these questions within the Dirac composite fermion theory. Prior work has identified observables that may possibly differ in the two composite fermion theories: Son and Levin [212] have derived a linear relation between the Hall conductivity and susceptibility that any PH symmetric theory must satisfy; Wang and Senthil [353] have determined how PH symmetry constrains the thermal Hall response of the HLR theory;

using the microscopic composite fermion wave function approach, Balram, Toke, and Jain [33] found that Friedel oscillations in the pair-correlation function are symmetric about $\nu = 1/2$.

2.0.1 Weiss oscillations and the $\nu = 1/2$

Here, we study theoretically commensurability oscillations in the magnetoresistance near $\nu = 1/2$, focusing on those oscillations that result from the presence of a periodic one-dimensional static potential [362]. These commensurability oscillations are commonly known as Weiss oscillations [360, 113, 365, 359]. For a free two-dimensional Fermi gas, the locations of the Weiss oscillation minima, say, as a function of the transverse magnetic field b, satisfy

$$\ell_b^2 = \frac{d}{2k_F}\Big(p + \phi\Big), \quad p = 1, 2, 3, \ldots, \tag{2.1}$$

where $\ell_b = 1/\sqrt{|b|}$ is the magnetic length; d is the period of the potential; k_F is the Fermi wave vector; $\phi = +1/4$ for a periodic vector potential, while $\phi = -1/4$ for a periodic scalar potential [282, 378]. (Expressions for the oscillation minima when both potentials are present can be found in Refs. [283, 112].)

Early experiments [362] saw $p = 1$ Weiss oscillation minima about $\nu = 1/2$ due to an electrostatic *scalar* potential, upon identifying, in Eq. (2.1), $b = B - 4\pi n_e$ with the effective magnetic field experienced by composite fermions (B is the external magnetic field and n_e is the electron density) and $k_F = \sqrt{4\pi n_e}$ with the composite fermion Fermi wave vector, and choosing $\phi = +1/4$. These results, along with other commensurability oscillation experiments [362], provided strong support for the general picture of the $\nu = 1/2$ state

suggested by the HLR theory. In particular, the phenomenology near the $\nu = 1/2$ state could be well described by an HLR mean-field theory in which composite fermions respond to an electronic scalar potential as a *vector* potential.

Recent improvements in sample quality and experimental design have allowed for an unprecedented refinement of these measurements. Through a careful study of the oscillation minima corresponding to the first three harmonics ($p = 1, 2, 3$), Kamburov et al. [177] came to a remarkable conclusion that is in apparent disagreement with the above hypothesis (see [329] for a review of these and related experiments): Weiss oscillation minima are well described by Eq. (2.1) upon taking $k_F = \sqrt{4\pi n_e}$ for $\nu < 1/2$, as before; but for $\nu > 1/2$, the inferred Fermi wave vector, $k_F = \sqrt{4\pi(\frac{B}{2\pi} - n_e)}$, is determined by the density of holes. In both cases, $\phi = +1/4$. Might a theory of the $\nu = 1/2$ state require two *different* composite fermion theories [177, 36], a theory of composite electrons for $\nu < 1/2$ and a theory of composite holes for $\nu > 1/2$? If $k_F = \sqrt{4\pi n_e}$ is instead taken for $1/2 < \nu < 1$, there is a roughly 2% mismatch between the locations of the $p = 1$ minimum obtained from Eq. (2.1) and the nearest observed minimum; this discrepancy between theory and experiment decreases in magnitude as p increases [177]. While the mismatch is small, it is systematic: it persists in a variety of different samples of varying mobilities and densities, as well as two-dimensional hole gases, which typically have larger effective masses (as well as near half-filling of other Landau levels [329]). (This mismatch is the same magnitude as the difference between the electrical Hall conductivities produced by an HLR theory with $\sigma_{xy}^{\mathrm{cf}} = 0$ and an HLR theory with $\sigma_{xy}^{\mathrm{cf}} = -1/4\pi$, the composite fermion Hall conductivity required by PH symmetry; an equal value of the dissipative resistance [362] is assumed in

both cases for this comparison. See Eq. (48) of [185].)

The hypothesis that composite fermions respond to an electric scalar potential as a purely magnetic one approximates HLR mean-field field theory. In fact, an electric scalar potential generates both a scalar and vector potential in the HLR theory. (This observation by Wang et al. [351] is crucial for obtaining PH symmetric electrical Hall transport within HLR mean-field theory.) However, the magnitude of the scalar potential is suppressed relative to the vector potential by a factor of $\ell_B/d \approx 1/50$ [36]. Cheung et al. [73] found that upon including the effects of the scalar potential in HLR mean-field theory, there is a slight correction to the expected locations of the oscillation minima *both* above and below $\nu = 1/2$. The nature of the corrections are such that HLR mean-field theories of composite electrons or composite holes that take either $k_F = \sqrt{4\pi n_e}$ or $k_F = \sqrt{4\pi(\frac{B}{2\pi} - n_e)}$ produce identical results. In addition, the shifted oscillation minima are in agreement with the mean-field predictions of the Dirac composite fermion theory (at least within the regime of electronic parameters probed by experiment). Unfortunately, the small disagreement between composite fermion mean-field theory and experiment persists, in this case for all values of $0 < \nu < 1$: for a given p, the observed oscillation minima are shifted inwards relative to the theoretical prediction by an amount that decreases as $\nu = 1/2$ is approached—see Fig. 2.1.

2.0.2 Outline

In this paper, we consider the mismatch from the point of view of the Dirac composite fermion theory. In perturbation theory about mean-field theory, we argue that the comparison with experiment can be improved if the effects of gauge field fluctuations

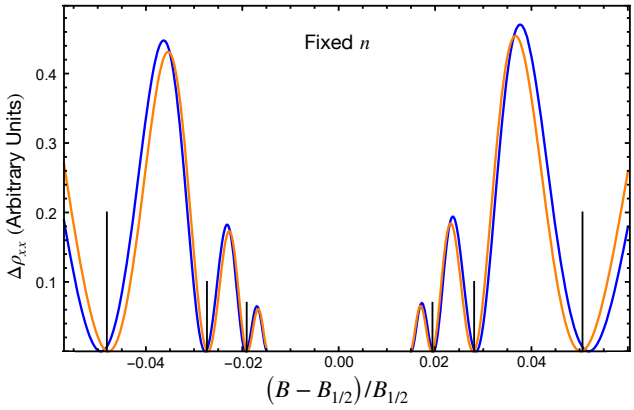

Figure 2.1: Weiss oscillations of the Dirac composite fermion theory at fixed electron density n_e and varying magnetic field B about half-filling $B_{1/2}$ ($\ell_{B_{1/2}}/d = 0.03$ and $k_B T = 0.3\sqrt{2B_{1/2}}$). The blue curve corresponds to Dirac composite fermion mean-field theory [73]. The orange curve includes the effects of a Dirac composite fermion mass $m \propto |B - 4\pi n_e|^{1/3} B^{1/6}$ induced by gauge fluctuations. Vertical lines correspond to the observed oscillation minima [177].

are considered. Our strategy is to include their effects by determining the fluctuation corrections to the mean-field Hamiltonian. We obtain this corrected Hamiltonian through an approximate large N flavor analysis of the Schwinger–Dyson equations [164] for the Dirac composite fermion theory. The resulting Dirac composite fermion propagator specifies the input parameters, namely, the chemical potential and mass, of the corrected mean-field Hamiltonian. We then follow the analysis by Cheung et al. [73] to determine the corrected Weiss oscillation curves. Our results are summarized in Fig. 2.1.

To understand our results, it is helpful to reinterpret Eq. (2.1) as a measure of a Dirac fermion density n by replacing $k_F \mapsto \sqrt{4\pi n}$ (we set the Fermi velocity to unity). Any

decrease in the density induces an inward shift of the Weiss oscillation minima determined by Eq. (2.1) towards $b = 0$. Dirac fermions of mass m, placed at chemical potential μ have a density $n = (\mu^2 - m^2)/4\pi$. Our leading order analysis of the Schwinger–Dyson equations indicates that gauge fluctuations generate a mass m away from $\nu = 1/2$, while the chemical potential is unchanged.

Such dynamical mass generation in a non-zero magnetic field is known to occur in various (2+1)-dimensional theories of Dirac fermions (see [240] for a review). For example, in the theory of a free Dirac fermion at zero density, a uniform magnetic field sources a vacuum expectation value for the mass operator. Short-ranged attractive interactions then induce a non-zero mass term in its effective Lagrangian [136]. We show how a similar phenomenon occurs in the Dirac composite fermion theory. This effect is also expected from the point of view of symmetry: PH symmetry forbids a Dirac composite fermion mass (see §2.1). (Manifest PH symmetry is the essential advantage that the Dirac composite fermion theory confers to our analysis.) Away from $\nu = 1/2$, PH symmetry is broken and so all terms, consistent with the broken PH symmetry, are expected to be present in the effective Lagrangian. Note there is no symmetry preventing corrections to the Dirac composite fermion chemical potential; rather, it is found to be unaltered to leading order within our analysis.

We also comment upon the finite-temperature behavior of quantum oscillations near $\nu = 1/2$. This behavior is interesting to consider because at finite temperatures, away from the long wavelength limit, differences in the HLR and Dirac composite fermion theories should appear. We discuss how the temperature dependence of the Weiss oscillation

amplitude might exhibit subtle differences between the two theories.

The remaining sections are organized as follows. In §2.1, we review the Dirac composite fermion theory. In §2.2, we obtain an approximate solution to the Schwinger–Dyson equations. In §2.3, we use the chemical potential and mass of the resulting Dirac composite fermion propagator as input parameters for the "fluctuation-improved" mean-field Hamiltonian and determine the resulting Weiss oscillations. We discuss a few consequences of this analysis in §2.4 and we conclude in §5. Appendix A.1 contains details of calculations summarized in the main text.

2.1 Dirac composite fermions: review

Electrons in the lowest Landau level near half-filling can be described by a Lagrangian of a 2-component Dirac electron Ψ_e [334]:

$$\mathcal{L}_e = \overline{\Psi}_e \gamma^\alpha (i\partial_\alpha + A_\alpha)\Psi_e - m_e \overline{\Psi}_e \Psi_e + \frac{1}{8\pi}\epsilon^{\alpha\beta\sigma} A_\alpha \partial_\beta A_\sigma + \ldots, \tag{2.2}$$

where A_α with $\alpha \in \{0,1,2\}$ is the background electromagnetic gauge field; $\overline{\Psi}_e = \Psi_e^\dagger \gamma^0$; the γ matrices $\gamma^0 = \sigma^3$, $\gamma^1 = i\sigma^1$, $\gamma^2 = i\sigma^2$ satisfy the Clifford algebra $\{\gamma^\alpha, \gamma^\beta\} = 2\eta^{\alpha\beta}$ with $\eta^{\alpha\beta} = \mathrm{diag}(+1,-1,-1)$; the anti-symmetric symbol $\epsilon^{012} = 1$; and we set the Fermi velocity $v_F = 1$ here and in the Dirac composite fermion dual. The benefit of the Dirac formulation is that the limit of infinite cyclotron energy $\omega_c = B/m_e$ can be smoothly achieved at fixed external magnetic field $B = \partial_1 A_2 - \partial_2 A_1 > 0$ by taking the electron mass $m_e \to 0$. The $\ldots$ include additional interactions, e.g., the Coulomb interaction and coupling to disorder.

The electron density,

$$n_e = \Psi_e^\dagger \Psi_e + \frac{B}{4\pi}. \tag{2.3}$$

Consequently, when $\nu \equiv 2\pi n_e/B = 1/2$, the Dirac electrons half-fill the zeroth Landau level. For $m_e = 0$ and $\nu = 1/2$, the Dirac Lagrangian is invariant under the anti-unitary ($i \mapsto -i$) PH transformation that takes $(t, x, y) \mapsto (-t, x, y)$,

$$\Psi_e \mapsto -\gamma^0 \Psi_e^*,$$

$$(A_0, A_1, A_2) \mapsto (-A_0, A_1, A_2), \tag{2.4}$$

and shifts the Lagrangian by a filled Landau level $\mathcal{L}_e \mapsto \mathcal{L}_e + \frac{1}{4\pi}\epsilon^{\alpha\beta\sigma} A_\alpha \partial_\beta A_\sigma$.

Son [334] conjectured that $\mathcal{L}_e$ is dual to the Dirac composite fermion Lagrangian,

$$\mathcal{L} = \overline{\psi}\gamma^\alpha (i\partial_\alpha + a_\alpha)\psi - m\overline{\psi}\psi - \frac{1}{4\pi}\epsilon^{\alpha\beta\sigma} a_\alpha \partial_\beta A_\sigma + \frac{1}{8\pi}\epsilon^{\alpha\beta\sigma} A_\alpha \partial_\beta A_\sigma - \frac{1}{4g^2} f_{\alpha\beta}^2 + \ldots, \tag{2.5}$$

where ψ is the electrically-neutral Dirac composite fermion; a_α is a dynamical $U(1)$ gauge field with field strength $f_{\alpha\beta} = \partial_\alpha a_\beta - \partial_\beta a_\alpha$ and coupling g; and $m \propto m_e$ is the Dirac composite fermion mass. A_α remains a non-dynamical gauge field, whose primary role in $\mathcal{L}$ is to determine how electromagnetism enters the Dirac composite fermion theory. As before, the $\ldots$ represent additional interactions, which can now involve the gauge field a_α. The duality between $\mathcal{L}_e$ and $\mathcal{L}$ obtains in the low-energy limit when $g \to \infty$. See [235, 352, 176, 111, 248, 251, 322, 178, 335] for additional details about this duality and [324] for a recent review.

At weak coupling, the a_0 equation of motion implies the Dirac composite fermion density,

$$\psi^\dagger \psi = \frac{B}{4\pi}. \tag{2.6}$$

At strong coupling, the right-hand side of Eq. (2.6) receives corrections from the $\ldots$ in $\mathcal{L}$ and should be replaced by $-\frac{\delta \mathcal{L}}{\delta a_0} + \psi^\dagger \psi$. In the Dirac composite fermion theory, the electron

density,

$$n_e = \frac{1}{4\pi}(-b + B), \tag{2.7}$$

where the effective magnetic field $b = \partial_1 a_2 - \partial_2 a_1$. In the Dirac composite fermion theory, the PH transformation takes $(t, x, y) \mapsto (-t, x, y)$,

$$\begin{aligned} \psi &\mapsto \gamma^2 \psi, \\ (a_0, a_1, a_2) &\mapsto (a_0, -a_1, -a_2), \\ (A_0, A_1, A_2) &\mapsto (-A_0, A_1, A_2), \end{aligned} \tag{2.8}$$

and shifts the Lagrangian by a filled Landau level. Intuitively, the PH transformation acts on the dynamical fields of $\mathcal{L}$ like a time-reversal transformation. As such, PH symmetry requires $m = 0$ and forbids a Chern-Simons term for a_α.

Away from half-filling, PH symmetry is necessarily broken since Eq. (2.7) implies the effective magnetic field $b = B - 4\pi n_e \neq 0$. Consequently, we can no longer exclude any PH breaking term allowed by symmetry. In particular, we generally expect a Dirac mass to be induced by fluctuations. Scaling implies the mass $m = \sqrt{B} f(\nu)$, where $f(\nu)$ is a scaling function of the filling fraction ν. Unbroken PH symmetry at half-filling requires $f(\nu = 1/2) = 0$; away from $\nu = 1/2$, it is possible that m can have a non-trivial dependence on B and n_e, as determined by $f(\nu)$. In the next section, we study the Schwinger–Dyson equations to determine how fluctuations generate a mass m away from $\nu = 1/2$ within an expansion where the number of Dirac composite fermion flavors $N \to \infty$.

2.2 Dynamical mass generation in an effective magnetic field

Beginning with the works of Schwinger [318] and Ritus [301], there have been a number of studies on the effects of a background magnetic field on quantum electrodynamics in various dimensions. In this paper, we rely most heavily on Refs. [128, 358, 181]; see Ref. [240] for an excellent introduction to this formalism and for additional references. We first summarize the relevant aspects of this formalism. Then, we analyze the Schwinger–Dyson equations for the Dirac composite fermion theory away from half-filling when the fluctuations of the emergent gauge field a_α about a uniform $b \neq 0$ are considered.

2.2.1 Dirac fermions in a magnetic field

At tree-level, i.e., in mean-field theory, the time-ordered real-space propagator $G_0(x,y)$ for a massive Dirac fermion in a uniform magnetic field $(\overline{a}_0, \overline{a}_1, \overline{a}_2) = (0, 0, bx_1)$ can be written in the form,

$$G_0(x,y) = e^{i\Phi(x,y)} \int \frac{d^3p}{(2\pi)^3} e^{ip_\alpha (x-y)^\alpha} G_0(p), \tag{2.9}$$

where the Schwinger phase,

$$\Phi(x,y) = -\frac{b}{2}(x_2 - y_2)(x_1 + y_1). \tag{2.10}$$

The tree-level pseudo-momentum-space propagator,

$$\begin{aligned} -iG_0(p) = i \int_0^\infty ds e^{is\Big((p_0+\mu_0+i\epsilon_{p_0})^2 - m_0^2 + i\delta - \frac{p_1^2+p_2^2}{bs}\tan(bs)\Big)} \\ \times \Big[(p_\alpha + \mu_0\delta_{\alpha,0})\gamma^\alpha - ib\Big((p_0+\mu_0)\mathbb{I} + m_0\gamma^0\Big)\tan(bs) + p_i\gamma^i\tan^2(bs)\Big], \end{aligned} \tag{2.11}$$

where the pseudo-momenta $p = (p_0, p_1, p_2)$ are analogous to the conserved momenta in a translationally-invariant system, μ_0 is a chemical potential, m_0 is a mass, $\epsilon_{p_0} = \text{sign}(p_0)\epsilon$

with the infinitesimal $\epsilon > 0$ ensures the Feynman pole prescription is satisfied, $\delta > 0$ is an infinitesimal included for convergence of the s integral, and $\mathbb{I}$ is the 2×2 identity matrix. Expanding in b:

$$-iG_0(p) \equiv \frac{(p_\alpha + \mu_0 \delta_{\alpha,0})\gamma^\alpha + m_0 \mathbb{I}}{(p_0 + \mu_0 + i\epsilon_{p_0})^2 - p_i^2 - m_0^2} + b \frac{(p_0 + \mu_0)\mathbb{I} + m_0 \gamma^0}{\Big((p_0 + \mu_0 + i\epsilon_{p_0})^2 - p_i^2 - m_0^2\Big)^2} + \mathcal{O}(b^2). \tag{2.12}$$

We imagine applying this formalism to the vicinity of $\nu = 1/2$ when the effective magnetic field b is small. As such, we drop all $\mathcal{O}(b^2)$ and higher terms in the pseudo-momentum-space propagator. For convenience, we use $G_0(p)$ to denote the linear expansion in Eq. (2.12) with higher order in b terms excluded.

The tree-level inverse propagator $G_0^{-1}(x, y)$ satisfies

$$\int d^3y \; G_0^{-1}(x, y) G_0(y, z) = \delta^{(3)}(x - z). \tag{2.13}$$

It takes a particularly simple form:

$$iG_0^{-1}(x, y) = e^{i\Phi(x,y)} \int \frac{d^3p}{(2\pi)^3} e^{ip_\alpha (x-y)^\alpha} \Big((p_\alpha + \mu_0 \delta_{\alpha,0})\gamma^\alpha - m_0 \mathbb{I}\Big). \tag{2.14}$$

In contrast to $G_0(x, y)$, the magnetic field dependence is entirely parameterized by the Schwinger phase in $G_0^{-1}(x, y)$.

Both the propagator and its inverse are obtained after performing an infinite sum over all Landau levels. Thus, $G_0(x, y)$ and $G_0^{-1}(x, y)$ in Eqs. (2.9) and (2.14) allow for a straightforward expansion about their translationally-invariant forms at $b = 0$; see [358] for further discussion. In the Dirac composite fermion theory, $G_0^{-1}(x, y)$ defines the mean-field Lagrangian, from which the Hamiltonian readily follows; the Schwinger phase $\Phi(x, y)$ reminds us to include a non-zero magnetic field by the Peierls substitution.

We use the following ansatz for the exact real-space propagator:

$$G(x,y) = e^{i\Phi(x,y)} \int \frac{d^3p}{(2\pi)^3} e^{ip_\alpha (x-y)^\alpha} G(p). \tag{2.15}$$

For the exact pseudo-momentum propagator $G(p)$, we write

$$-iG(p) = -iG^{(0)}(p) - iG^{(1)}(p), \tag{2.16}$$

where

$$-iG^{(0)}(p) = \frac{\Big(p_\alpha + \mu_0 \delta_{\alpha,0} - \Sigma_\alpha(p)\Big)\gamma^\alpha + \Sigma_m(p)\mathbb{I}}{(p_0 + \mu_0 - \Sigma_0(p) + i\epsilon_{p_0})^2 - (p_i - \Sigma_i(p))^2 - \Sigma_m^2(p)}, \tag{2.17}$$

$$-iG^{(1)}(p) = b\frac{\Big(p_0 + \mu_0 - \Sigma_0(p)\Big)\mathbb{I} + \Sigma_m(p)\gamma^0}{\Big((p_0 + \mu_0 - \Sigma_0(p) + i\epsilon_{p_0})^2 - (p_i - \Sigma_i(p))^2 - \Sigma_m^2(p)\Big)^2}. \tag{2.18}$$

In contrast to the tree-level pseudo-momentum propagator, $G_0(p)$, both $G^{(0)}(p)$ and $G^{(1)}(p)$ are expected to depend on b through the self-energies $\Sigma_m(p)$ and $\Sigma_\alpha(p)$, in addition to the explicit linear dependence that appears in $G^{(1)}(p)$. We write the exact inverse propagator as

$$iG^{-1}(x,y) = e^{i\Phi(x,y)} \int \frac{d^3p}{(2\pi)^3} e^{ip_\alpha (x-y)^\alpha} \left(\left(p_\alpha + \mu_0 \delta_{\alpha,0} - \Sigma_\alpha\left(p\right)\right)\gamma^\alpha - \Sigma_m\left(p\right)\mathbb{I}\right). \tag{2.19}$$

In $G(p)$ and $G^{-1}(p)$, we set the tree-level mass $m_0 = 0$; this is consistent with the assumption of unbroken PH symmetry at $\nu = 1/2$. The ansatze for the exact propagator and its inverse are simplifications of that which symmetry allows for a Dirac fermion in a magnetic field [358]. Nevertheless, our ansatze are consistent to leading order in a $1/N$ analysis of the Schwinger–Dyson equations described in the next section.

In general, the self-energies $\Sigma_m(p)$ and $\Sigma_\alpha(p)$ are non-trivial functions of the pseudo-momenta p. We expect the low-energy dynamics of the fermions to be dominated by

fluctuations about the Fermi surface. Thus, we replace the self-energies as follows:

$$\Sigma_m(p_{\rm FS} + \delta p) \mapsto \Sigma_m(p_{\rm FS}), \tag{2.20}$$

$$\Sigma_\alpha(p_{\rm FS} + \delta p) \mapsto \delta_{0\alpha}\Sigma_0(p_{\rm FS}) + \delta p_\alpha \Sigma'_\alpha(p_{\rm FS}), \tag{2.21}$$

where $p_{\rm FS} = (0, p_i)$ lies on the Fermi surface (in mean-field theory, this is defined by $p_i^2 = \mu_0^2$ and $p_0 = 0$), $|\delta p_\alpha| \ll \mu_0$, $\Sigma'_\alpha(p_{\rm FS}) = \partial_{p_\alpha}\Sigma_\alpha(p = p_{\rm FS})$, and there is no sum over α in Eq. (2.21).

$G^{-1}(x, y)$ determines the "fluctuation-corrected" Dirac composite fermion mean-field Hamiltonian. The tree-level chemical potential and mass are corrected by the fermion self-energies Σ_α and Σ_m. We define the physical mass,

$$m = \frac{\Sigma_m(p_{\rm FS})}{1 - \Sigma'_0(p_{\rm FS})} \equiv \frac{\Sigma_m}{1 - \Sigma'_0}, \tag{2.22}$$

and chemical potential,

$$\mu = \frac{\mu_0 - \Sigma_0}{1 - \Sigma'_0}. \tag{2.23}$$

The Schwinger phase $\Phi(x, y)$ in $G^{-1}(x, y)$ reminds us to to include the effective magnetic field b via the Peierls substitution.

2.2.2 Schwinger–Dyson equations: setup

The Schwinger–Dyson equations [164] are a set of coupled integral equations that relate the exact fermion and gauge field propagators to one another by way of the exact cubic interaction vertex Γ^α coupling the Dirac composite fermion current to a_α. We will not solve the equations exactly; rather, we seek an approximate solution that one obtains

within a large flavor generalization of the Dirac composite fermion theory. We hope this approximate solution reflects a qualitative behavior of the Dirac composite fermion theory.

Specifically, we consider the Lagrangian,

$$\mathcal{L}_N = \overline{\psi}_n \gamma^\alpha (i\partial_\alpha + a_\alpha)\psi_n - \frac{N}{4\pi}\epsilon^{\alpha\beta\sigma} a_\alpha \partial_\beta A_\sigma + \frac{N}{8\pi}\epsilon^{\alpha\beta\sigma} A_\alpha \partial_\beta A_\sigma - \frac{1}{4g^2} f_{\alpha\beta}^2, \tag{2.24}$$

where the different fermion flavors are labeled by $n = 1, \ldots, N$. When $N = 1$, we recover the Dirac composite fermion theory. In $\mathcal{L}_N$, $n_e = \delta\mathcal{L}_N/\delta A_0 = \frac{N}{4\pi}(B - b)$; thus, in our large N theory, half-filling means $\nu = N/2$. To make contact with the formalism of §2.2.1, we introduce a $SU(N)$-invariant chemical potential $\mu_0 = \sqrt{B}$ and we factor out the uniform effective magnetic field $(\overline{a}_0, \overline{a}_1, \overline{a}_2) = (0, 0, bx_1)$ that is generated away from half-filling from the dynamical fluctuations of the emergent gauge field a_α. Setting $A_\alpha = 0$, Eq. (2.24) becomes

$$\mathcal{L}_N = \overline{\psi}_n \gamma^\alpha (i\partial_\alpha + \overline{a}_\alpha + a_\alpha)\psi_n + \mu_0 \psi_n^\dagger \psi_n - \frac{1}{4g^2} f_{\alpha\beta}^2. \tag{2.25}$$

This is the large N theory that we analyze.

To leading order in N, the Ward identity implies that there are no corrections to the cubic interaction vertex at $\nu = 1/2$ [302].[1] Taking $\Gamma^\alpha = \gamma^\alpha$, the Schwinger–Dyson equations for $\mathcal{L}_N$ become:

$$iG^{-1}(x, y) - iG_0^{-1}(x, y) = \gamma^\alpha G(x, y)\gamma^\beta \Pi_{\alpha\beta}^{-1}(x - y), \tag{2.26}$$

$$i\Pi^{\alpha\beta}(x - y) - i\Pi_0^{\alpha\beta}(x - y) = N\mathrm{tr}\Big[\gamma^\alpha G(x, y)\gamma^\beta G(y, x)\Big], \tag{2.27}$$

where $\Pi^{\alpha\beta}(x-y)$ is the gauge field self-energy, $\Pi_0^{\alpha\beta}(x-y)$ is the kinetic term for a_α contributed by its Maxwell term, and we have taken the fermion propagator $G_{n,n'}(x, y) = G(x, y)\delta_{n,n'}$

[1] Furthermore, there are no corrections to this vertex if the Dirac composite fermion is given a non-zero bare mass $m_0^2 \ll \mu_0^2$ at $b = 0$. We thank N. Rombes and S. Chakravarty for correspondence on this point.

to be diagonal in flavor space. $G(x,y)$ and $G_0(x,y)$ are defined in Eqs. (2.16) and (2.12). The factor of N in Eq. (2.27) arises from the N flavors in the fermion loop.

Upon substituting the Fourier transform $\Pi^{\alpha\beta}(p)$, defined by

$$\Pi^{\alpha\beta}(x-y) = \int \frac{d^3p}{(2\pi)^3} e^{ip_\sigma (x-y)^\sigma} \Pi^{\alpha\beta}(p), \tag{2.28}$$

and Eqs (2.14), (2.16), and (2.19) into the Schwinger–Dyson equations, (2.26) and (2.27) become [358]

$$i\Sigma_\alpha(q)\gamma^\alpha + i\Sigma_m(q)\mathbb{I} = \int \frac{d^3p}{(2\pi)^3} \gamma^\alpha G(p+q)\gamma^\beta \Pi^{-1}_{\alpha\beta}(p), \tag{2.29}$$

$$i\Pi^{\alpha\beta}(\delta q) = N \int \frac{d^3p}{(2\pi)^3} \mathrm{tr}\Big[\gamma^\alpha G(p)\gamma^\beta G(p+\delta q)\Big], \tag{2.30}$$

where $q = q_{\rm FS} + \delta q$. We aim to solve these equations.

Our ansatz for the fermion self-energies is motivated by similar studies of $(2+1)$-dimensional quantum electrodynamics at zero density [287, 17, 260]. We consider the $1/N$ expansion for the fermion self-energies,

$$\begin{aligned} \Sigma_\alpha &= \Sigma_\alpha^{(1)} + \Sigma_\alpha^{(2)} + \ldots, \\ \Sigma_m &= \Sigma_m^{(1)} + \Sigma_m^{(2)} + \ldots. \end{aligned} \tag{2.31}$$

All terms and all ratios of successive terms in Eq. (2.31) vanish as $N \to \infty$. Ignoring terms with $i \geq 2$, we set $\Sigma_\alpha = \Sigma_\alpha^{(1)} = 0$ and $\Sigma_m = \Sigma_m^{(1)}$, and find a self-consistent solution to the Schwinger–Dyson equation in terms of $\Sigma_m^{(1)}$ and $\Pi^{\alpha\beta}$. This choice is consistent with the Ward identity, to leading order in $1/N$. From Eqs. (2.22) and (2.23), the resulting solution implies $m = \Sigma_m^{(1)}$ and $\mu = \mu_0$ to leading order in $1/N$. We then calculate the leading perturbative correction $\Sigma_\alpha^{(2)}$ to Σ_α and verify that $\Sigma_\alpha^{(2)}/\Sigma_m^{(1)} \to 0$ as $N \to \infty$.

2.2.3 Gauge field self-energy

The gauge field self-energy factorizes into PH symmetry even and odd parts:

$$\Pi^{\alpha\beta}(q) = \Pi^{\alpha\beta}_{\text{even}}(q) + \Pi^{\alpha\beta}_{\text{odd}}(q). \tag{2.32}$$

As the PH transformation acts like time-reversal, $\Pi^{\alpha\beta}_{\text{even}}(q)$ contains the Maxwell term for a_α, while $\Pi^{\alpha\beta}_{\text{odd}}(q)$—which can only be non-zero when PH symmetry is broken—can contain a Chern-Simons term for a_α.

To leading order in b, we substitute $G(p) = G^{(0)}(p)$ into Eq. (2.30) and first compute

$$\Pi^{\alpha\beta}_{\text{odd}}(\delta q) = i\epsilon^{\alpha\beta\sigma}\delta q_\sigma \Pi_{\text{odd}}(\delta q) = -iN\Big\{ \int \frac{d^3p}{(2\pi)^3} \text{tr}\Big[\gamma^\alpha G^{(0)}(p)\gamma^\beta G^{(0)}(p+\delta q)\Big]\Big\}_{\text{odd}}, \tag{2.33}$$

where $\{\cdot\}_{\text{odd}}$ indicates the PH odd term is isolated. We find

$$\Pi_{\text{odd}}(0) = \frac{N}{4\pi}\Big(\Theta(|\Sigma_m| - \mu_0)\frac{\Sigma_m}{|\Sigma_m|} + \Theta(\mu_0 - |\Sigma_m|)\frac{\Sigma_m}{\mu_0}\Big), \tag{2.34}$$

where $\Theta(x)$ is the step function. See Appendix A.1.1 for details. Additional momentum dependence in $\Pi_{\text{odd}}(q)$ is subdominant at low energies. For $\mu_0 > |\Sigma_m|$, Eq. (2.34) implies an effective Chern-Simons term for α_α with level,

$$k = \frac{N}{2}\frac{\Sigma_m}{\mu_0}, \tag{2.35}$$

is generated if $\Sigma_m \neq 0$. (This non-quantized Chern-Simons level is reminiscent of the anomalous Hall effect [139].)

Next, consider

$$\Pi^{\alpha\beta}_{\text{even}}(\delta q) - \Pi^{\alpha\beta}_0(\delta q) = -iN\Big\{ \int \frac{d^3p}{(2\pi)^3} \text{tr}\Big[\gamma^\alpha G^{(0)}(p)\gamma^\beta G^{(0)}(p+\delta q)\Big]\Big\}_{\text{even}}, \tag{2.36}$$

where $\{\cdot\}_{\text{even}}$ indicates the PH even term is isolated and we have again substituted $G(p) = G^{(0)}(p)$. The Maxwell kinetic term is

$$\Pi_0^{\alpha\beta}(q) = q^2\eta^{\alpha\beta} - q^\alpha q^\beta. \tag{2.37}$$

Ref. [239] finds:

$$\begin{aligned}
&\Pi^{00}_{\text{even}}(q_0, q_i) - \Pi_0^{00}(q) = \Pi_l(q_0, q_i), \\
&\Pi^{0i}_{\text{even}}(q_0, q_i) - \Pi_0^{0i}(q) = q_0 \frac{q^i}{q_i^2}\Pi_l(q_0, q_i), \\
&\Pi^{ij}_{\text{even}}(q_0, q_i) - \Pi_0^{0i}(q) = (\delta^{ij} - \frac{q^i q^j}{q_k^2})\Pi_t(q_0, q_i) + \frac{q_0^2 q^i q^j}{(q_k^2)^2}\Pi_l(q_0, q_i),
\end{aligned} \tag{2.38}$$

where

$$\begin{aligned}
&\Pi_l(q_0, q_i) = \mu_0 N\Big(\sqrt{\frac{q_0^2}{q_0^2 - q_i^2}} - 1\Big), \\
&\Pi_t(q_0, q_i) = \mu_0 N - \frac{q_0^2 - q_i^2}{q_k^2}\Pi_l(q_0, q_i).
\end{aligned} \tag{2.39}$$

We have simplified the expressions for Π_l and Π_t by taking $q_0^2 - q_i^2 > 0$ and by setting the common proportionality constant to unity. The precise behaviors of Π_l and Π_t and their effects on a_α depend upon whether $|q_0| < |q_i|$ or $|q_i| < |q_0|$. For instance, when $|q_0| < |q_i|$ (small frequency transfers, but potentially large $\sim 2k_F$ momenta transfers) and in the absence of $\Pi^{\alpha\beta}_{\text{odd}}$, Π_l gives rise to the usual Debye screening of the "electric" component of a_α and Π_t results in the Landau damping of the "magnetic" component of a_α [239], familiar from Fermi liquid theory [154]. These corrections dominate the tree-level Maxwell term for a_α at low energies.

In our analysis of the fermion self-energy in the next section, we focus on the regime $|q_i| \leq |q_0|$. In this case, Π_l and Π_t provide non-singular corrections to the Maxwell term

for a_α and will be ignored. At low energies, $g \to \infty$, the effects of the Maxwell term are suppressed compared with the Chern-Simons term [89]. Thus, to find the effective gauge field propagator $\Pi^{-1}_{\alpha\beta}(q)$ for use in Eq. (2.29), we drop $\Pi^{\alpha\beta}_{\text{even}}(q)$, add the covariant gauge fixing term $-\frac{1}{2\xi}q^\alpha q^\beta$ to $\Pi^{\alpha\beta}_{\text{odd}}(q)$, and invert. Choosing Feynman gauge $\xi = 0$, we obtain:

$$\Pi^{-1}_{\alpha\beta}(q) = \frac{2\pi}{k}\frac{\epsilon_{\alpha\beta\sigma}q^\sigma}{q^2}, \tag{2.40}$$

where k is given in Eq. (2.35). It is with this gauge field propagator that we find a self-consistent solution to the Schwinger–Dyson equation for the fermion self-energy Σ_m in §2.2.4.

Instantaneous density-density interactions between electrons give rise to additional gauge field kinetic terms in $\mathcal{L}$. Such terms, which should therefore be included in the tree-level Lagrangian $\mathcal{L}_N$, generally contribute to $\Pi^{\alpha\beta}_0 \subset \Pi^{\alpha\beta}_{\text{even}}$. To understand their possible effects in the kinematic regime $|q_i| \leq |q_0|$, we set $a_0 = 0$ and decompose the spatial components of the gauge field in terms of its longitudinal and transverse modes:

$$a_i(q) = -i\hat{q}_i a_L(q) - i\epsilon_{ji}\hat{q}_j a_T(q), \tag{2.41}$$

where the normalized spatial momenta $\hat{q}_i = q_i/|\vec{q}|$. An un-screened Coulomb interaction dualizes to a term in $\mathcal{L}$ proportional to $|\vec{q}|^{z-1}a_T(-q)a_T(q)$ with $z = 2$; a short-ranged interaction give $z = 3$ (see Sec. 3.4 of [176]). (We are working in momentum space for this analysis.) On the other hand, the effective Chern-Simons term is proportional to $iq_0 a_L(-q)a_T(q)$; there is no $a_L - a_L$ or $a_T - a_T$ Chern-Simons coupling. We consider $z > 2$ in our analysis below. In this regime, the effects of any such screened interaction are expected to be subdominant compared with those of the Chern-Simons term, as such interactions correspond to higher-order terms in the derivative expansion.

2.2.4 Fermion self-energy

We now study Eq. (2.29) for the Σ_m and Σ_0 components of the Dirac composite fermion self-energy using the effective gauge field propagator in Eq. (2.40).

Σ_m

Taking the trace of both sides of Eq. (2.29) and setting $\delta q_\alpha = 0$, we find:

$$i\Sigma_m(q_{\rm FS}) = i\mathcal{M}^{(0)}(q_{\rm FS}) + i\mathcal{M}^{(1)}(q_{\rm FS}), \tag{2.42}$$

where

$$i\mathcal{M}^{(0)}(q_{\rm FS}) = \frac{1}{2}\int \frac{d^3p}{(2\pi)^3}{\rm tr}\Big[\gamma^\alpha G^{(0)}(p+q_{\rm FS})\gamma^\beta\Big(\frac{2\pi}{k}\frac{\epsilon_{\alpha\beta\sigma}p^\sigma}{p^2}\Big)\Big], \tag{2.43}$$

$$i\mathcal{M}^{(1)}(q_{\rm FS}) = \frac{1}{2}\int \frac{d^3p}{(2\pi)^3}{\rm tr}\Big[\gamma^\alpha G^{(1)}(p+q_{\rm FS})\gamma^\beta\Big(\frac{2\pi}{k}\frac{\epsilon_{\alpha\beta\sigma}p^\sigma}{p^2}\Big)\Big], \tag{2.44}$$

and $G^{(0}(p)$ and $G^{(1)}(p)$ are given in Eqs. (2.17) and (2.18). Recall that we set $\Sigma_\alpha = 0$ and only retain Σ_m when using $G^{(0)}(p)$ and $G^{(1)}(p)$ to evaluate $\mathcal{M}^{(0)}$ and $\mathcal{M}^{(1)}$. The details of our evaluation of $\mathcal{M}^{(0)}$ and $\mathcal{M}^{(1)}$ are given in Appendix A.1.2. Here, we quote the results:

$$\mathcal{M}^{(0)} = -\frac{2\mu_0 {\rm sign}(\Sigma_m)}{N}, \tag{2.45}$$

$$\mathcal{M}^{(1)} = \frac{2}{3}\frac{b\mu_0^2}{N|\Sigma_m|^3}. \tag{2.46}$$

Thus, Σ_m solves:

$$\Sigma_m = -\frac{2\mu_0 {\rm sign}(\Sigma_m)}{N} + \frac{2}{3}\frac{b\mu_0^2}{N|\Sigma_m|^3}. \tag{2.47}$$

When $b = 0$, the only solution is $\Sigma_m = 0$, consistent with our expectation that PH symmetry is unbroken at $\nu = 1/2$. Dimensional analysis and $1/N$ scaling implies

$$\Sigma_m = \frac{\mu_0}{N} f\Big(\frac{bN^3}{\mu_0^2}\Big). \tag{2.48}$$

We find that Σ_m has the following asymptotics: for fixed $|b|/\mu_0^2 \approx 10^{-1}$,

$$\Sigma_m = \mu_0 \mathrm{sign}(b)\Big(\frac{|b|}{\mu_0^2 N}\Big)^{1/4}\Big[c_1 + c_2\Big(\frac{\mu_0^2}{|b|N^3}\Big)^{1/4} + \dots\Big], \tag{2.49}$$

where $c_1 \approx 0.9$, $c_2 \approx -0.5$, and the $\dots$ are suppressed as $N \to \infty$; while for fixed N,

$$\Sigma_m = \mu_0 \mathrm{sign}(b)\Big(\frac{|b|}{\mu_0^2}\Big)^{1/3}\Big[c_3 + c_4\Big(\frac{|b|N^3}{\mu_0^2}\Big)^{1/3} + \dots\Big], \tag{2.50}$$

where $c_3 \approx 0.69$, $c_4 \approx -0.08$, and the $\dots$ vanish as $|b|/\mu_0^2 \to 0$.

Σ_0

We now consider the leading perturbative correction to Σ_0. This allows us to calculate the corrections to Σ_0' and the chemical potential μ_0.

To evaluate the leading correction to Σ_0 that one obtains when $G(p) = G^{(0)}(p)$, we multiply both sides of Eq. (2.29) by γ^0 on the left and take the trace to find:

$$i\Sigma_0(q) = \frac{1}{2}\int \frac{d^3p}{(2\pi)^3}\mathrm{tr}\Big[\gamma^0\gamma^\alpha G^{(0)}(p+q)\gamma^\beta\Big(\frac{2\pi}{k}\frac{\epsilon_{\alpha\beta\sigma}p^\sigma}{p^2}\Big)\Big], \tag{2.51}$$

where $q^\alpha = q^\alpha_{\mathrm{FS}} + q_0\delta^{\alpha 0}$. As detailed in Appendix A.1.2, we find the leading correction $\Sigma_0^{(2)}$ to Σ_0 (see Eq. (2.31)) for $|q_0|/\mu_0 \ll \Sigma_m^2/\mu_0^2$,

$$i\Sigma_0^{(2)}(q_{\mathrm{FS}}) = -i\frac{2\mu_0}{3N|\Sigma_m|}(q_0 + \mu_0). \tag{2.52}$$

At large N, we use Eq. (2.49) for Σ_m to find $\Sigma_0 \propto \Sigma_0' \propto N^{-3/4}$. This vanishes by a factor of $N^{-1/2}$ *faster* than Σ_m and so it is relatively suppressed as $N \to \infty$. Next-order terms in Σ_α and Σ_m are obtained by self-consistently solving the Schwinger–Dyson equations with propagators corrected by the leading self-energy corrections. We have checked that the other components of Σ_α are likewise suppressed at large N; as such and because they do not

enter our subsequent calculations, we will not discuss them further. Because Σ_m vanishes at half-filling, we may only ignore Σ_0' for sufficiently large $|b|/\mu_0^2$ at large N.

Dynamically-generated mass and corrected chemical potential

We are now ready to evaluate Eq. (2.22) for the dynamically-generated mass. We extrapolate our large N solution for Σ_m to $N = 1$ using Eq. (2.50):

$$m = \frac{\Sigma_m^{(1)}}{1 - \Sigma_0'^{(1)}} \approx .69\text{sign}(b)|b|^{1/3}B^{1/6}, \tag{2.53}$$

where we set $\mu_0 = \sqrt{B}$. The specific behavior of the mass m, away from $\nu = 1/2$, depends on whether the electron density n_e or external magnetic field B is fixed. At fixed B, the magnitude of m is symmetric as function of n_e about half-filling; on the other hand, $|m|$ is asymmetric for fixed n_e and varying B. Using Eqs. (2.23) and (2.31), the chemical potential,

$$\mu = \frac{\mu_0 - \Sigma_0^{(1)}}{1 - \Sigma_0'^{(1)}} = \sqrt{B}. \tag{2.54}$$

These results imply that the Dirac composite fermion density and mass are corrected in such a way that the chemical potential is unaffected.

In our analysis of the Weiss oscillations in the next section, we ignore all higher-order in $1/N$ corrections and assume that a mass term is the dominant correction to the Dirac composite fermion mean-field Hamiltonian away from $\nu = 1/2$. The chemical potential for this fluctuation-improved mean-field Hamiltonian will be taken to be $\mu = \sqrt{B}$.

2.3 Weiss oscillations of massive Dirac composite fermions

Following earlier work [227, 342, 64, 73], we now study the effect of the field-dependent mass of Eq. (2.53) on the Weiss oscillations near $\nu = 1/2$ using the fluctuation-

improved Dirac composite fermion mean-field theory. We find that a non-zero mass results in an inward shift of the locations of the oscillation minima toward half-filling.

2.3.1 Setup

We are interested in determining the quantum oscillations in the electrical resistivity near $\nu = 1/2$ that result from a one-dimensional periodic scalar potential. In the Dirac composite fermion theory, the dc electrical conductivity,

$$\sigma_{ij} = \frac{1}{4\pi}\Big(\epsilon_{ij} - \frac{1}{2}\epsilon_{ik}(\sigma^{\psi})^{-1}_{kl}\epsilon_{lj}\Big), \tag{2.55}$$

where the (dimensionless) dc Dirac composite fermion conductivity. This equality is true at weak coupling; at strong coupling, $\langle\overline{\psi}\gamma_i\psi(-q_0)\overline{\psi}\gamma_j\psi(q_0)\rangle$ should be replaced by the exact gauge field a_α self-energy, evaluated at $q_1 = q_2 = 0$.

$$\sigma^{\psi}_{ij} = \lim_{q_0\to 0}\frac{\langle\overline{\psi}\gamma_i\psi(-q_0)\overline{\psi}\gamma_j\psi(q_0)\rangle}{iq_0}. \tag{2.56}$$

Thus, the longitudinal electrical resistivity,

$$\rho_{ii} \propto |\epsilon_{ij}|\sigma^{\psi}_{jj}, \tag{2.57}$$

where there is no sum over repeated indices. When a one-dimensional periodic scalar potential, $A_0 = V\cos(Kx_1)$ with $K = 2\pi/d$, is applied to the electronic system, the a_2 equation of motion following from the Dirac composite fermion Lagrangian (2.5) implies

$$\overline{\psi}\gamma^2\psi = -\frac{KV}{4\pi}\sin(Kx). \tag{2.58}$$

We accommodate this current modulation within Dirac composite fermion mean-field theory by turning on a modulated perturbation to the emergent vector potential,

$$\delta\vec{a} = \Big(0, W\sin(Kx_1)\Big), \tag{2.59}$$

where $W = W(V)$ vanishes when $V = 0$. (Fluctuations will also generate a modulation in the Dirac composite fermion chemical potential; we ignore such effects here.) Putting together Eqs. (2.57) and (2.59), our goal in this section is to determine the correction to σ^{ψ}_{jj} due to $\delta\vec{a}$,

$$\Delta\rho_{ii} \propto |\epsilon_{ij}|\Delta\sigma^{\psi}_{jj}. \tag{2.60}$$

In Dirac composite fermion mean-field theory, corrected by Eq. (2.53), the calculation of $\Delta\sigma^{\psi}_{ij}$ simplifies to the determination of the conductivity of a free *massive* Dirac fermion. We use the Kubo formula [69] to find the conductivity correction:

$$\Delta\sigma^{\psi}_{ij} = \frac{1}{L_1 L_2}\Sigma_M \Big(\partial_{E_M} f_D(E_M)\Big)\tau(E_M) v_i^M v_j^M, \tag{2.61}$$

where L_1 (L_2) is the length of the system in the x_1-direction (x_2-direction), $\beta^{-1} = T$ is the temperature, M denotes the quantum numbers of the single-particle states, $f_D^{-1}(E) = 1 + \exp(\beta(E-\mu))$ is the Fermi-Dirac distribution function with chemical potential $\mu = \sqrt{B}$, $\tau(E_M)$ is the scattering time for states at energy E_M, and $v_i^M = \partial_{p_i} E_M$ is the velocity correction in the x_i-direction of the state M due to the periodic vector potential. As before, the Fermi velocity is set to unity. Assuming constant $\tau(E) = \tau \neq 0$, we only need to calculate how the energies E_M are affected by $\delta\vec{a}$, which in turn will determine the velocities v_i^M. We will show that the leading correction in W to E_M only contributes to v_2^M. Calling $x_1 = x$ and $x_2 = y$, this implies the dominant correction is to $\Delta\rho_{xx} \propto \Delta\sigma^{\psi}_{yy}$. There are generally oscillatory corrections to ρ_{yy} and ρ_{xy}, however, their amplitudes are typically less prominent and so we concentrate on $\Delta\rho_{xx}$ here.

2.3.2 Dirac composite fermion Weiss oscillations

The Dirac composite fermion mean-field Hamiltonian, corrected by Eq. (2.53),

$$H = \vec{\sigma} \cdot \Big(\frac{\partial}{\partial \vec{x}} + \vec{a} \Big) + m\sigma_3, \tag{2.62}$$

where

$$\vec{a} = \Big(0, bx_1 + W \sin(Kx_1) \Big). \tag{2.63}$$

To zeroth order in W, H has the particle spectrum,

$$E_n^{(0)} = \begin{cases} \sqrt{2n|b| + m^2}, & n = 1, 2, \ldots, \\ \\ |m|, & n = 0. \end{cases}$$

with the corresponding eigenfunctions,

$$\psi_{n,p_2}(\vec{x}) = \begin{cases} \mathcal{N} e^{ip_2 x_2} \begin{pmatrix} -i\Phi_{n-1}\big(\frac{x_1 + x_b}{l_b}\big) \\ \frac{\sqrt{m^2+2n|b|}-m}{\sqrt{2n|b|}} \Phi_n\big(\frac{x_1+x_b}{l_b}\big) \end{pmatrix} & \text{for } n = 1, 2, \ldots, \\ \\ \mathcal{N} e^{ip_2 x_2} \begin{pmatrix} 0 \\ \\ \Phi_0\big(\frac{x_1+x_b}{l_b}\big) \end{pmatrix} & \text{for } n = 0, \end{cases}$$

where the normalization constant,

$$\mathcal{N} = \sqrt{\frac{n|b|}{l_b L_y (m^2 + 2n|b| - m\sqrt{m^2 + 2n|b|})}},$$

$k_2 \in \frac{2\pi}{L_2}\mathbb{Z}$ is the momentum along the x_2-direction ($L_2 \to \infty$), $x_b(p_2) \equiv x_b = p_2 l_b^2$, $l_b^{-1} = |b|$, and $\Phi_n(z) = \frac{e^{-z^2/2}}{\sqrt{2^n n! \sqrt{\pi}}} H_n(z)$ for the n-th Hermite polynomial $H_n(z)$. Thus, the states are

labeled by $M = (n, p_2)$. We are interested in how the periodic vector potential in Eq. (2.59) lifts the degeneracy of the flat Landau level spectrum and contributes to the velocity v_i^M. (Finite dissipation has already been assumed in using a finite, non-zero scattering time τ in our calculation of the oscillatory component of ρ_{xx}.)

First order perturbation theory gives the energy level corrections,

$$E^{(1)}_{n,p_2} = W\frac{\sqrt{2n}}{Kl_b}\left[\sqrt{\frac{2n|b|}{m^2+2n|b|}}\right]\cos(Kx_b)e^{-z/2}\Big[L_{n-1}(z)-L_n(z)\Big], \tag{2.64}$$

where $L_n(z)$ is the nth Laguerre polynomial, $z = K^2l_b^2/2$, and terms suppressed as $L_1, L_2 \to \infty$ have been dropped. Thus, to leading order, $v_1^{n,p_2} = 0$ and

$$v_2^{n,p_2} = \frac{\partial E^{(1)}_{n,p_2}}{\partial p_2} = -Wl_b\sqrt{2n}\left[\sqrt{\frac{2n|b|}{m^2+2n|b|}}\right]\sin(Kx_b)\mathrm{e}^{-z/2}\Big[L_{n-1}(z)-L_n(z)\Big]. \tag{2.65}$$

We substitute these v_i^{n,p_2} into the Kubo formula (2.61) to find $\Delta\sigma^{\psi}_{yy}$. To perform the integral over p_2, we approximate the Fermi-Dirac distribution function by substituting in the zeroth order energies $E_n^{(0)}$ (which are independent of p_2). Thus, we obtain the periodic potential correction to the Dirac composite fermion conductivity:

$$\Delta\sigma^{\psi}_{yy} \approx W^2\widetilde{\tau}\beta\sum_{n=0}^{\infty}\left(\frac{2n|b|}{m^2+2n|b|}\right)\frac{n\exp(\beta(E_n^{(0)}-\mu))}{\Big[1+\exp(\beta(E_n^{(0)}-\mu))\Big]^2}\mathrm{e}^{-z}\Big[L_{n-1}(z)-L_n(z)\Big]^2, \tag{2.66}$$

where $\tilde{\tau} \propto \tau$ has absorbed non-universal $\mathcal{O}(1)$ constants.

$\Delta\sigma^{\psi}_{yy}$ in Eq. (2.66) exhibits both Shubnikov–de Haas (for large $|b|$) and Weiss oscillations (for smaller $|b|$). We are interested in extracting an analytic expression that approximates Eq. (2.66) at low temperatures near $\nu = 1/2$, following the earlier analysis in [283]. In the weak field limit, $|b|/\mu^2 \ll 1$, a large number of Landau levels are filled ($n \to \infty$). Thus, we express the Laguerre polynomials L_n as

$$L_n(z) \xrightarrow[n\to\infty]{} \mathrm{e}^{z/2}\frac{\cos\left(2\sqrt{nz}-\frac{\pi}{4}\right)}{(\pi^2nz)^{1/4}} + \mathcal{O}(\frac{1}{n^{3/4}}). \tag{2.67}$$

Next, we take the continuum approximation for the summation over n by substituting

$$n \to \frac{l_b^2}{2}\Big(E^2 - m^2\Big), \quad \sum_n \to l_b^2 \int \; EdE,$$

into Eq. (2.66):

$$\Delta\sigma_{yy}^{\psi} = \mathcal{C} \int_{-\infty}^{\infty} dE \; \frac{\beta e^{\beta(E-\mu)}}{(1+e^{\beta(E-\mu)})^2} \sin^2\Big(l_b^2 K\sqrt{E^2-m^2} - \frac{\pi}{4}\Big), \tag{2.68}$$

where $\mathcal{C} = W^2\tilde{\tau} l_b^2 K$ and we have approximated $2n|b|/(m^2+2n|b|)$ by unity. (The substitution for n is motivated by the zeroth order expression for the energy of the Dirac composite fermion Landau levels.) Anticipating that at sufficiently low temperatures the integrand in Eq. (2.68) is dominated by "energies" E near the Fermi energy μ, we write:

$$E = \mu + sT \tag{2.69}$$

so that Eq. (2.68) becomes for $|s|T \ll \mu = \sqrt{B}$:

$$\Delta\sigma_{yy}^{\psi} = \mathcal{C} \int_{-\infty}^{\infty} ds \; \frac{e^s}{(1+e^s)^2} \sin^2\Big(l_b^2 K\sqrt{B-m^2} + \frac{sTl_b^2 K}{\sqrt{1-\frac{m^2}{B}}} - \frac{\pi}{4}\Big). \tag{2.70}$$

Performing the integral over s, we find the Weiss oscillations (see Eq. (2.60)):

$$\Delta\rho_{xx} \propto 1 - \frac{T/T_D}{\sinh(T/T_D)}\Big[1 - 2\sin^2\Big(\frac{2\pi l_b^2\sqrt{B-m^2}}{d} - \frac{\pi}{4}\Big)\Big], \tag{2.71}$$

where

$$T_D^{-1} = \frac{4\pi^2 l_b^2}{d}\frac{1}{\sqrt{1-\frac{m^2}{B}}}, \tag{2.72}$$

we have substituted $K = 2\pi/d$, $l_b^2 = |b|^{-1}$, and the proportionality constant is controlled by the longitudinal resistivity at $\nu = 1/2$.

Eq. (2.71) constitutes the primary result of this section. The minima of $\Delta\rho_{xx}$ occur when

$$\frac{1}{|b|} = \frac{d}{2\sqrt{B - m^2}}\Big(p + \frac{1}{4}\Big), p = 1, 2, 3, \ldots, \tag{2.73}$$

where m is given in Eq. (2.53). For either fixed electron density n_e or fixed external field B, the locations of the oscillation minima for a given p (either $B(p)$ or $n_e(p)$) are shifted inwards towards $\nu = 1/2$. This is shown in Fig. 2.1 for fixed n_e and in Fig. 2.2 for fixed B. The magnitude of this shift is symmetric for fixed B, but asymmetric for fixed n_e, given the form of the mass in Eq. (2.53). Mass dependence also appears in the temperature-dependent prefactor $\frac{T/T_D}{\sinh(T/T_D)}$. In principle, this mass dependence could be extracted from the finite-temperature scaling of $\Delta\rho_{xx}$ at the oscillation extrema.

2.4 Comparison to HLR mean-field theory at finite temperature

2.4.1 Shubnikov–de Haas oscillations

In [225], Shayegan et al. found the Shubnikov–de Haas (SdH) oscillations near half-filling to be well described over two orders of magnitude in temperature by the formula,

$$\frac{\Delta\rho_{xx}}{\rho_0} \propto \frac{\xi_{NR}}{\sinh(\xi_{NR})}\cos(2\pi\nu - \pi), \tag{2.74}$$

where $\xi_{NR} = \frac{2\pi^2 T}{\omega_c}$, $\omega_c = |b|/m^*$, m^* is an effective mass, ν is the electron filling fraction, and ρ_0 is the longitudinal resistivity at half-filling (measured at the lowest accessible temperature). (Note that these experiments were performed without any background

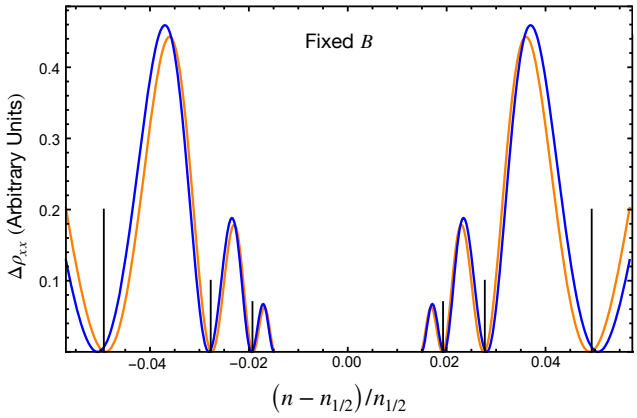

Figure 2.2: Weiss oscillations of the Dirac composite fermion theory at fixed magnetic field B and varying electron density n_e about half-filling $n_{1/2} = B_{1/2}/4\pi$ ($\ell_{B_{1/2}}/d = 0.03$ and $k_B T = 0.3\sqrt{2B_{1/2}}$). The blue curve corresponds to Dirac composite fermion mean-field theory [73]. The orange curve includes the effects of a Dirac composite fermion mass $m \propto |B - 4\pi n_e|^{1/3} B^{1/6}$ induced by gauge fluctuations. Vertical lines correspond to the observed oscillation minima [177].

periodic potential and so no Weiss oscillations were present.) Recall that we are using units where $k_B = \hbar = e = c = 1$. In particular, it was found that $m^* \propto \sqrt{B}$ for sufficiently large $|b| = |B - 4\pi n_e|$ and that m^* appeared to diverge as half-filling was approached. Interpreted within the HLR composite fermion framework, m^* corresponds to the composite fermion effective mass. The $\sqrt{B}$ behavior of the composite fermion effective mass is consistent with the theoretical expectation [140, 297] that the composite fermion mass scale at $\nu = 1/2$ is determined entirely by the characteristic energy of the Coulomb interaction. (Away from $\nu = 1/2$, scaling implies the effective mass can be a scaling function of B and n_e.)

Applying previous treatments of SdH oscillations in graphene [135, 129] to the Dirac composite fermion theory, the temperature dependence of the SdH oscillations is controlled by

$$\frac{\Delta\rho_{xx}}{\rho_0} \propto \frac{\xi_D}{\sinh(\xi_D)}, \tag{2.75}$$

where $\xi_D = \frac{2\pi^2 T\sqrt{B}}{|b|}$. Thus, $\xi_{NR} \propto \xi_D$ if $m^* \propto \sqrt{B}$. Consequently, the Dirac composite fermion theory is consistent with the observed temperature scaling with $\sqrt{B}$. We cannot account for the divergence at small $|b|$ attributed to m^* in our treatment.

2.4.2 Weiss oscillations

In [73], it was shown that the locations of the Weiss oscillation minima obtained from Dirac and HLR composite fermion mean-field theories coincide to 0.002%. This result provides evidence that the two composite fermion theories may belong to the same universality class. However, the (possible) equivalence of the two theories only occurs at long distances and so the finite-temperature behavior of the two theories will generally differ.

In HLR *mean-field theory*, the temperature dependence of the Weiss oscillations enters in the factor [283],

$$\Delta\rho_{xx} \propto \frac{T/T_{NR}}{\sinh(T/T_{NR})}, \tag{2.76}$$

where the characteristic temperature scale,

$$T_{NR}^{-1} = \frac{4\pi^2 l_b^2}{d} \frac{m^*}{\sqrt{4\pi n_e}}. \tag{2.77}$$

Assuming the effective mass $m^* \propto \sqrt{B}$, the characteristic temperatures T_D and T_{NR} generally have very different behaviors as functions of B and n_e. It would be interesting to study the effects of fluctuations in HLR theory, along the lines of the study presented here, and compare with our result in Eq. (2.71).

Chapter 3

LEAD: Min-Max Optimization from a Physical Perspective

Much of the advances in traditional machine learning can be attributed to the success of gradient-based methods. Modern machine learning formulations such as GANs [125], multi-task learning, and multi-agent settings [323] in reinforcement learning [62] require joint optimization of two or more objectives. In these *game* settings, best practices and methods developed for single-objective optimization are observed to perform noticeably poorly [233, 32, 116]. Notably, they exhibit rotational dynamics about the *Nash Equilibria* [233], slowing down convergence to the same. Recent work in game optimization [356, 228, 233, 32, 2, 216] demonstrates that intuitively introducing additional second-order terms in the optimization algorithm, helps to suppress these rotations, thereby improving convergence. Despite their relative success in many settings, several of these methods are computationally expensive to implement, preventing successful deployment in setups of relevance such as in GANs.

Taking inspiration from recent work in single-objective optimization that re-derives existing accelerated methods from a variational perspective [361, 364], in this work, we adopt a similar approach in the context of games. By likening the gradient-based optimization of two-player (zero-sum) games to the dynamics of a particular physical system, we introduce a relevant force that helps curb these rotations. We consequently utilize the dynamics of this resultant system to propose our novel second-order optimizer for games, *LEAD*.

Next, by using Lyapunov and spectral analysis, we demonstrate linear convergence of our optimizer (LEAD) in both continuous and discrete-time settings for a class of quadratic min-max games. In terms of empirical performance, LEAD achieves an FID of 10.49 on CIFAR-10 image generation, outperforming existing methods such as BigGAN [60] which is approximately 30-times larger than our baseline ResNet architecture.

What distinguishes LEAD from other second-order optimization methods for min-max games such as [233, 356, 228, 315] is its computational complexity. All these other methods, involve Jacobian (or Jacobian-inverse) vector-product computation[1], thus making a majority of them intractable in real-world problems such as GANs. On the other hand, LEAD involves computing only *one-block* of the full Jacobian of the gradient vector-field multiplied by a vector. This makes our method significantly cheaper and comparable to several first-order methods, as we show in section 3.5.1.

We summarize our contributions below:

- In Section 3.2, we model gradient descent-ascent as a physical system. Armed with the physical model, we introduce counter-rotational forces to curb the existing rotations in

[1] [356, 315] propose a conjugate-gradient approximation of the Jacobian-inverse to reduce computational cost, though still performing poorly in neural network setting. Additionally, their provided proofs of convergence rely on the expensive exact computation of the inverse.

the system. Next, we employ the principle of least action to determine the (continuous-time) dynamics. We then accordingly discretize these resultant dynamics to obtain our optimization scheme, Least Action Dynamics (LEAD).

- In Section 3.3, we use Lyapunov stability theory and spectral analysis to prove a linear convergence of LEAD in continuous and discrete-time settings for bilinear min-max games.

- Finally, in Section 3.5, we empirically demonstrate that LEAD is computationally efficient. Additionally, we demonstrate that LEAD improves the performance of GANs on different tasks such as 8-Gaussians and CIFAR-10 while comparing the performance of our method against other first and second-order methods. Furthermore, we achieve a competitive FID of 10.49 for CIFAR-10 on a ResNet architecture.

- The source code for all the experiments is available at:
`https://github.com/ReyhaneAskari/Least_action_dynamics_minmax`.

3.1 Problem Setting

Notation: Continuous time scalar variables are in uppercase letters (X), discrete-time scalar variables are in lower case (x) and vectors are in boldface ($\mathbf{A}$). Matrices are in blackboard bold ($\mathbb{M}$) and derivatives w.r.t. time are denoted as an over-dot ($\ddot{x}$).

Setting: In this work, we study the optimization problem of two-player zero-sum games,

$$\min_X \max_Y f(X, Y), \tag{3.1}$$

where $f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, and is assumed to be a convex-concave function which is continuous and twice differentiable w.r.t. $X, Y \in \mathbb{R}$. It is to be noted that though in developing our framework below, X, Y are assumed to be scalars, it is nevertheless found to hold for the more general case of vectorial X and Y, as we demonstrate both analytically (Appendix B.3) and empirically, our theoretical analysis are found to hold.

3.2 Optimization Mechanics

In our attempt to find an efficient update scheme or trajectory to optimize the min-max objective $f(X, Y)$, we note from classical physics the following: under the influence of a net force F, the trajectory of motion of a physical object of mass m, is determined by Newton's 2$^{\text{nd}}$ Law,

$$m\ddot{X} = F, \tag{3.2}$$

with the object's coordinate expressed as $X_t \equiv X$. According to the *principle of least action*[2] [200], nature "selects" this particular trajectory over other possibilities, as a quantity called the *action* is extremized along it.

Hence, the ability to model our game optimization task in terms of an object moving under a relevant set of forces, can be perceived as determining an efficient optimization path through the least action principle for the same. Regarding how such modeling may be performed, we take inspiration from Polyak's heavy-ball momentum [290] method[3] in single objective minimization of an objective $f(x)$,

$$x_{k+1} = x_k + \beta (x_k - x_{k-1}) - \eta \nabla_x f(x_k), \tag{3.3}$$

[2] Also referred to as the Principle of Stationary Action.

[3] Arbitrary momentum coefficient results in incorporating friction in the equivalent physical system

which in continuous-time translates to (see Appendix B.1),

$$m\ddot{X} = -\nabla_X f(X). \tag{3.4}$$

Comparing Eqns.(3.4) and (3.2), we notice that in this case $F = -\nabla_X f(X)$, i.e. $f(X)$ acts as a *potential* function [200]. Thus, Polyak's heavy-ball method Eq.(3.3) can be interpreted as an object (ball) of mass m rolling down under a potential $f(X)$ to reach the minimum.

Armed with this observation, we perform a straightforward extension of Eq.(3.4) to our min-max setup,

$$\begin{aligned} m\ddot{X} &= -\nabla_X f(X,Y) \\ m\ddot{Y} &= \nabla_Y f(X,Y). \end{aligned} \tag{3.5}$$

which represents the dynamics of an object moving under a *curl force* [48]:

$$\boldsymbol{F}_{\text{curl}} = (-\nabla_X f, \nabla_Y f) \tag{3.6}$$

in the 2-dimensional $X-Y$ plane. It is to be noted that discretization of Eq.(3.5) corresponds to Gradient Descent-Ascent (GDA) with momentum 1. [116] found that this optimizer is divergent in the prototypical min-max objective, $f(X,Y) = XY$ itself, thus indicating the need for further improvement.

To this end, we note that that the failure modes of the optimizer obtained from the discretization of Eq.(3.5), can be attributed to: *(a)* an outward rotatory motion by our particle of mass m, accompanied by *(b)* an increase in its velocity with time. Following these observations, we aim to introduce suitable *counter-rotational* and *dissipative* forces to our system above, in order to tackle *(a)* and *(b)* in an attempt to achieve converging dynamics. Specifically, as an initial consideration, we choose to add to our system, two ubiquitous forces:

- magnetic force,

$$\boldsymbol{F}_{\text{mag}} = \Big(-q\nabla_{XY} f\ \dot{Y}, q\nabla_{XY} f\ \dot{X} \Big) \tag{3.7}$$

known to produce rotational motion (in charged particles), to counteract the rotations introduced by $\boldsymbol{F}_{\text{curl}}$. Here, q is the charge imparted to our particle

- friction,

$$\boldsymbol{F}_{\text{fric}} = \big(\mu\dot{X}, \mu\dot{Y} \big) \tag{3.8}$$

to prevent the increase in velocity of our particle (μ: coefficient of friction)

Assimilating all the above forces $\boldsymbol{F}_{\text{curl}}$, $\boldsymbol{F}_{\text{mag}}$ and $\boldsymbol{F}_{\text{fric}}$, the equations of motion (EOMs) of our crafted system then becomes,

$$\begin{aligned} m\ddot{X} &= -\mu\dot{X} - \nabla_X f - q\nabla_{XY} f\dot{Y}, \\ m\ddot{Y} &= -\mu\dot{Y} + \nabla_Y f + q\nabla_{XY} f\dot{X}. \end{aligned} \tag{3.9}$$

Without loss of generality, from hereon we set the mass of our object to be unity.

3.2.1 Discretization

With the continuous-time trajectory of Eq.(3.9) in hand, we now proceed to discretize it using a combination of Euler's implicit and explicit discretization schemes,

$$\begin{aligned} &\text{Implicit}: \ x_{k+1} - x_k = \delta v^x_{k+1} \\ &\text{Explicit}: \ x_{k+1} - x_k = \delta v^x_k. \end{aligned} \tag{3.10}$$

to discretize $\dot{X} = V_X$ (δ: discretization step-size, k: iteration step).

Proposition: The continuous-time EOMs (3.9) can be discretized in an implicit-explicit

way, to yield,

$$\begin{aligned} x_{k+1} &= x_k + \beta(x_k - x_{k-1}) - \eta \nabla_x f\left(x_k, y_k\right) - \delta \nabla_{xy} f\left(x_k, y_k\right)\left(y_k - y_{k-1}\right), \\ y_{k+1} &= y_k + \beta(y_k - y_{k-1}) + \eta \nabla_y f\left(x_k, y_k\right) + \delta \nabla_{yx} f\left(x_k, y_k\right)\left(x_k - x_{k-1}\right), \end{aligned} \tag{3.11}$$

where we have defined $\alpha = 2q\delta, \beta = 1 - \mu\delta$ and $\eta = \delta^2$ (Proof in Appendix B.2).

Taking inspiration from the fact that Eq. (3.9) corresponds to the trajectory of a charged particle under a curl, magnetic and frictional force, as governed by the principle of least action, we refer to the discrete update rules of Eq. (3.11) as *Least Action Dynamics (LEAD)*. (Algorithm 1 details the pseudo-code of LEAD)

Terms in LEAD: Analyzing our novel optimizer, we note that it consist of three types of terms, namely,

1. Gradient Descent or Ascent: $-\nabla_x f$ or $\nabla_y f$: Each player's immediate direction of improving their own objective.

2. Momentum: Standard Polyak momentum term; known to accelerate convergence in optimization and recently in smooth games. [116, 26]

3. Coupling term: $-\nabla_{xy} f\left(x_k, y_k\right)\left(y_k - y_{k-1}\right)$ and $\nabla_{yx} f\left(x_k, y_k\right)\left(x_k - x_{k-1}\right)$: Main new term in our method. It captures the first-order interaction between players. This cross-derivative corresponds to the counter-rotational force in our physical model; it allows our method to exert control on rotations.

Algorithm 1 Least Action Dynamics (LEAD)

Input: learning rate η, momentum β, coupling coefficient α.

Initialize: $x_0 \leftarrow x_{init}$, $y_0 \leftarrow y_{init}$, $t \leftarrow 0$

while not converged **do**

$t \leftarrow t + 1$

$g_x \leftarrow \nabla_x f(x_t, y_t)$

$g_{xy}\Delta y_t \leftarrow \nabla_y(g_x)(y_t - y_{t-1})$

$x_{t+1} \leftarrow x_t + \beta(x_t - x_{t-1}) - \eta g_x - \alpha g_{xy}\Delta y_t$

$g_y \leftarrow \nabla_y f(x_t, y_t)$

$g_{xy}\Delta x_t \leftarrow \nabla_x(g_y)(x_t - x_{t-1})$

$y_{t+1} \leftarrow y_t + \beta(y_t - y_{t-1}) + \eta g_y + \alpha g_{xy}\Delta x_t$

end while

return (x_{k+1}, y_{k+1})

3.3 Convergence Analysis

We now study the behavior of LEAD on the quadratic min-max game,

$$f(\mathbf{X}, \mathbf{Y}) = \frac{h}{2}||\mathbf{X}||^2 - \frac{h}{2}||\mathbf{Y}||^2 + \mathbf{X}^T \mathbb{A} \mathbf{Y} \tag{3.12}$$

where $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^n$, $\mathbb{A} \in \mathbb{R}^n \times \mathbb{R}^n$ is a (constant) coupling matrix and h is a scalar constant. Additionally, the Nash equilibrium of the above game lies at $\mathbf{X}^* = 0, \mathbf{Y}^* = 0$. Let us further define the *vector field* $\mathbf{v}$ of the above game, f, as,

$$\boldsymbol{v} = \begin{bmatrix} \nabla_{\mathbf{X}} f(\mathbf{X}, \mathbf{Y}) \\ -\nabla_{\mathbf{Y}} f(\mathbf{X}, \mathbf{Y}) \end{bmatrix} = \begin{bmatrix} h\boldsymbol{X} + \mathbb{A}\mathbf{Y} \\ h\boldsymbol{Y} - \mathbb{A}^\top \mathbf{X} \end{bmatrix}. \tag{3.13}$$

3.3.1 Continuous Time Analysis

A general way to prove the stability of a dynamical system is to use a Lyapunov function [138, 221]. The scalar function $\mathcal{E}_t : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, is a Lyapunov function of a continuous-time dynamics if $\forall\ t$,

$$(i)\ \ \mathcal{E}_t(\mathbf{X}, \mathbf{Y}) \geq 0,$$

$$(ii)\ \ \dot{\mathcal{E}}_t(\mathbf{X}, \mathbf{Y}) \leq 0$$

The Lyapunov function $\mathcal{E}_t$ can be perceived as a generalization of the total energy of the system and the requirement (ii) ensures that this generalized energy decreases along the trajectory of evolution, leading the system to convergence as we will show next.

For the quadratic min-max game defined in Eq.(3.12), Eq.(3.9) generalizes to,

$$\begin{aligned} \ddot{\boldsymbol{X}} &= -\mu\dot{\boldsymbol{X}} - (h + \mathbb{A})\boldsymbol{Y} - q\mathbb{A}\dot{\boldsymbol{Y}} \\ \ddot{\boldsymbol{Y}} &= -\mu\dot{\boldsymbol{Y}} - (h - \mathbb{A}^T)\boldsymbol{X} + q\mathbb{A}^T\dot{\boldsymbol{X}}, \end{aligned} \tag{3.14}$$

Theorem 1 *For the dynamics of Eq.(3.14),*

$$\begin{aligned}\mathcal{E}_t = &\frac{1}{2}\left(\dot{\boldsymbol{X}} + \mu\boldsymbol{X} + \mu\mathbb{A}\boldsymbol{Y}\right)^T(\dot{\boldsymbol{x}} + \mu\boldsymbol{X} + \mu\mathbb{A}\boldsymbol{Y}) \\ &+ \frac{1}{2}\left(\dot{\boldsymbol{Y}} + \mu\boldsymbol{Y} - \mu\mathbb{A}^T\boldsymbol{X}\right)^T\left(\dot{\boldsymbol{X}} + \mu\boldsymbol{Y} - \mu\mathbb{A}^T\boldsymbol{X}\right) \\ &+ \frac{1}{2}\left(\dot{\boldsymbol{X}}^T\dot{\boldsymbol{X}} + \dot{\boldsymbol{Y}}^T\dot{\boldsymbol{Y}}\right) + \boldsymbol{X}^T(h + \mathbb{A}\mathbb{A}^T)\boldsymbol{X} + \boldsymbol{y}^T(h + \mathbb{A}^T\mathbb{A})\boldsymbol{Y}\end{aligned} \tag{3.15}$$

is a Lyapunov function of the system. Furthermore, by setting $q = (2/\mu) + \mu$*, we find* $\dot{\mathcal{E}}_t \leq -\rho\mathcal{E}_t$ *for* $\rho \leq \min\left\{\frac{\mu}{1+\mu},\ \frac{2\mu(\sigma_{\min}^2+h)}{(1+\sigma_{\min}^2+2h)(\mu^2+\mu)+2\sigma_{\min}^2}\right\}$ *with* $\sigma_{\min}$ *being the smallest singular value of* $\mathbb{A}$*. This consequently ensures linear convergence of the dynamics determined Eq.* (3.14)*,*

$$\boxed{||\boldsymbol{X}||^2 + ||\boldsymbol{Y}||^2 \leq \frac{\mathcal{E}_0}{h + \sigma_{min}^2}\exp\left(-\rho t\right)}. \tag{3.16}$$

(Proof in Appendix B.3).

3.3.2 Discrete-Time Analysis

In this Section, we next analyze the convergence behavior of LEAD, Eq.(3.11) in the case of the quadratic min-max game of Eq.(3.12), using spectral analysis,

$$\begin{aligned}\boldsymbol{x}_{k+1} &= \boldsymbol{x}_k + \beta\Delta\boldsymbol{x}_k - h\boldsymbol{x}_k - \eta\mathbb{A}\boldsymbol{y}_k - \alpha\mathbb{A}\Delta\boldsymbol{y}_k \\ \boldsymbol{y}_{k+1} &= \boldsymbol{y}_k + \beta\Delta\boldsymbol{y}_k - h\boldsymbol{y}_k + \eta\mathbb{A}^T\boldsymbol{x}_k + \alpha\mathbb{A}^T\Delta\boldsymbol{x}_k,\end{aligned} \tag{3.17}$$

where $\Delta\boldsymbol{x}_k = \boldsymbol{x}_k - \boldsymbol{x}_{k-1}$.

For brevity, consider the joint parameters $\boldsymbol{\omega}_t := (\boldsymbol{x}_t, \boldsymbol{y}_t)$. We start by studying the update operator of simultaneous gradient descent-ascent.

$$F_\eta(\boldsymbol{\omega}_t) = \boldsymbol{\omega}_t - \eta\boldsymbol{v}(\boldsymbol{\omega}_{t-1}).$$

where, the vector-field is given by Eq. (3.13). Thus, the fixed point $\boldsymbol{\omega}^*$ of $F_\eta(\boldsymbol{\omega}_t)$ satisfies

$F_\eta(\boldsymbol{\omega}^*) = \boldsymbol{\omega}^*$. Furthermore, at $\boldsymbol{\omega}^*$, we have,

$$\nabla F_\eta(\boldsymbol{\omega}^*) = \mathbb{I}_n - \eta \nabla \boldsymbol{v}(\boldsymbol{\omega}^*), \tag{3.18}$$

with $\mathbb{I}_n$ being the $n \times n$ identity matrix. Consequently the spectrum of $\nabla F_\eta(\boldsymbol{\omega}^*)$ in the quadratic game considered, is,

$$\mathrm{Sp}(\nabla F_\eta(\boldsymbol{\omega}^*)) = \{1 - \eta h - \eta\lambda \mid \lambda \in \mathrm{Sp}(\text{off-diag}[\nabla \boldsymbol{v}(\boldsymbol{\omega}^*)])\}\,. \tag{3.19}$$

Proposition:[Prop. 4.4.1 [49]] For the spectral radius,

$$\rho_{\max} := \rho\{\nabla F_\eta(\boldsymbol{\omega}^*)\} < 1 \tag{3.20}$$

and for some $\boldsymbol{\omega}_0$ in a neighborhood of $\boldsymbol{\omega}^*$, the update operator F, ensures linear convergence to $\boldsymbol{\omega}^*$ at a rate,

$$\Delta_{t+1} \leq \mathcal{O}(\rho + \epsilon)\Delta_t \;\forall\; \epsilon > 0,$$

where $\Delta_{t+1} := ||\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}^*||_2^2 + ||\boldsymbol{\omega}_t - \boldsymbol{\omega}^*||_2^2$.

Next, we proceed to define the update operator of Eq.(3.11) as $F_{\mathrm{LEAD}}\left(\boldsymbol{\omega}_t, \boldsymbol{\omega}_{t-1}\right) = \left(\boldsymbol{\omega}_{t+1}, \boldsymbol{\omega}_t\right)$. Now, for the quadratic min-max game of Eq.(3.12), the Jacobian of F_{LEAD} takes the form,

$$\nabla F_{\mathrm{LEAD}} = \begin{bmatrix} \mathbb{I}_{2n} + \beta\mathbb{I}_{2n} - (\eta + \alpha)\,\nabla \boldsymbol{v} & -\beta\mathbb{I}_{2n} + \alpha\nabla \boldsymbol{v} \\ \mathbb{I}_{2n} & 0 \end{bmatrix}. \tag{3.21}$$

Theorem 2 *The eigenvalues of* $\nabla F_{LEAD}(\boldsymbol{\omega}^*)$ *are,*

$$\mu_\pm = \frac{1 - (\eta + \alpha)\lambda + \beta - \eta h \pm \sqrt{\Delta}}{2} \tag{3.22}$$

where, $\Delta = (1 - (\eta + \alpha)\lambda + \beta - \eta h)^2 - 4(\beta - \alpha\lambda)$ *and* $\lambda \in Sp(\textit{off-diag}[\nabla \boldsymbol{v}(\boldsymbol{\omega}^*)])$. *Furthermore, for* $h, \eta, |\alpha|, |\beta| << 1$, *we have,*

$$\mu_+ \approx 1 - \eta h - \frac{\eta h \beta}{2} + \lambda \left(\frac{\eta + \alpha}{2} (\eta h - \beta) - \eta \right) \tag{3.23}$$

and

$$\mu_- \approx \beta + \frac{\eta h \beta}{2} - \lambda \left(\frac{\eta + \alpha}{2} (\eta h - \beta) + \alpha \right) \tag{3.24}$$

See Proof in Appendix B.4.

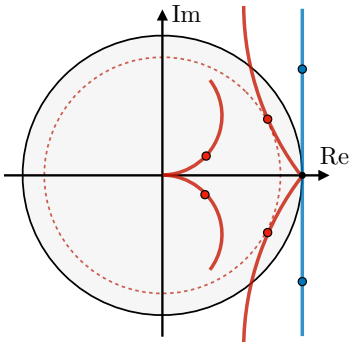

Figure 3.1: Diagram depicting positioning of the eigenvalues of GDA in blue (Eq. (3.18)) and those of LEAD (Eqns.(3.23),(3.24)) in red. Eigenvalues inside the black unit circle imply convergence such that the closer to the origin, the faster the convergence rate (Prop. 3.3.2). Every point on solid blue and red lines corresponds to a specific choice of learning rate. No choice of learning rate results in convergence for gradient ascent descent method as the blue line is tangent to the unit circle. At the same time, for a fixed value of α, LEAD shifts the eigenvalues (μ_+) into the unit circle which leads to a convergence rate proportional to the radius of the red dashed circle. Note that LEAD also introduces an extra set of eigenvalues (μ_-) which are close to zero and do not affect convergence.

In the following Proposition, we next show that locally, a choice of positive α decreases the spectral radius of $\nabla F_\eta(\boldsymbol{\omega}^*)$, $\rho := \max\{|\mu_+|^2, |\mu_-|^2\}\ \forall\ \lambda$.

Proposition: For any $\lambda \in \text{Sp}(\text{off-diag}[\nabla \boldsymbol{v}(\omega^*)])$,

$$\nabla_\alpha \rho(\lambda)\big|_{\alpha=0} < 0 \Leftrightarrow \eta \in \left(0, \frac{2}{\text{Im}(\lambda_{\max})}\right), \tag{3.25}$$

where $\text{Im}(\lambda_{\max})$ is the imaginary component of the largest eigenvalue $\lambda_{\max}$. See Proof in Appendix B.5.

Having established that a small positive value of α improves the rate of convergence, in the next theorem, we prove that for a specific choice of positive α and η in the quadratic game Eq.(3.12), a linear rate of convergence to its Nash equilibrium is attained. Before proceeding, we would like to note that we can write $\lambda_i = \pm i\sigma_i$ where $\sigma_i \equiv$ sing. values($\mathbb{A}$).

Theorem 3 *If we set $\eta = \alpha = \frac{2}{\sigma_{max}}$, then we have $\forall\ \epsilon > 0$,*

$$\Delta_{t+1} \in \mathcal{O}\left(\left(1 - 6\frac{2\sigma_{\min}^2 + h^2}{\sigma_{\max}^2} - 2h\frac{2+\beta}{\sigma_{\max}} + \frac{\beta^2}{2}\right)^t \Delta_0\right) \tag{3.26}$$

where $\sigma_{max}(\sigma_{min})$ is the largest (smallest) singular value of $\mathbb{A}$, $\Delta_{t+1} := ||\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}^*||_2^2 + ||\boldsymbol{\omega}_t - \boldsymbol{\omega}^*||_2^2$.

Theorem 3 ensures a linear convergence of LEAD in the quadratic min-max game. (Proof in Appendix B.6).

3.4 Comparison of Convergence Rate

In this Section, we perform a Big-O comparison of rates of convergence of LEAD (Eq. (3.26)), with Extragradient [188] in the quadratic min-max game of Eq. (3.12), with

$\beta = 0$. Specifically, from [25] we find,

$$\begin{aligned} r_{\text{EG}} &= \mathcal{O}\left(\frac{h^2}{L^2}\right) + \mathcal{O}\left(\frac{\sigma_{\min}^2(\mathbb{A})}{L^2}\right) + \mathcal{O}\left(\frac{h}{L}\right) \\ r_{\text{LEAD}} &= \mathcal{O}\left(\frac{h^2}{\sigma_{\max}^2(\mathbb{A})}\right) + \mathcal{O}\left(\frac{\sigma_{\min}^2(\mathbb{A})}{\sigma_{\max}^2(\mathbb{A})}\right) - \mathcal{O}\left(\frac{h}{\sigma_{\max}(\mathbb{A})}\right) \end{aligned} \tag{3.27}$$

where, $L := \max\{h, \sigma_{\max}(\mathbb{A})\}$. Therefore, for $h < \sigma_{\max}(\mathbb{A})$, we observe that $r_{\text{LEAD}} \lesssim r_{\text{EG}}$. While for $h > \sigma_{\max}(\mathbb{A})$, we note that,

$$\begin{aligned} r_{\text{EG}} &= \mathcal{O}(1) + \mathcal{O}\left(\frac{\sigma_{\min}^2(\mathbb{A})}{h^2}\right) \\ r_{\text{LEAD}} &= \mathcal{O}\left(\frac{h^2}{\sigma_{\max}^2(\mathbb{A})}\right) + \mathcal{O}\left(\frac{\sigma_{\min}^2(\mathbb{A})}{\sigma_{\max}^2(\mathbb{A})}\right) - \mathcal{O}\left(\frac{h}{\sigma_{\max}(\mathbb{A})}\right) \end{aligned} \tag{3.28}$$

Hence, for $h \gtrsim 1.62\sigma_{\max}$, we find $r_{\text{LEAD}} \gtrsim r_{\text{EG}}$.

3.5 Experiments

In this Section, we empirically validate the performance of our proposed method LEAD. Furthermore, we implement LEAD-Adam (pseudo-code in Appendix B.7.1) to be used in our experiments.

3.5.1 Comparison of Computational Cost

The Jacobian of the gradient vector field $\boldsymbol{v} = (\nabla_x f(\boldsymbol{x}, \boldsymbol{y}), -\nabla_y f(\boldsymbol{x}, \boldsymbol{y}))$ is given by,

$$\mathbb{J} = \begin{bmatrix} \nabla_x^2 f(\boldsymbol{x}, \boldsymbol{y}) & \nabla_{xy} f(\boldsymbol{x}, \boldsymbol{y}) \\ -\nabla_{yx} f(\boldsymbol{x}, \boldsymbol{y}) & -\nabla_y^2 f(\boldsymbol{x}, \boldsymbol{y}) \end{bmatrix}. \tag{3.29}$$

Considering player $\boldsymbol{x}$, a LEAD update for the same requires the computation of the term $\nabla_{xy} f(\boldsymbol{x}_k, \boldsymbol{y}_k)(\boldsymbol{y}_k - \boldsymbol{y}_{k-1})$, thereby involving only one block of the full Jacobian $\mathbb{J}$. On the other hand, the released implementation of Symplectic Gradient Adjustment (SGA) [31],

requires the full computation of two Jacobian-vector products $\mathbb{J}\boldsymbol{v}, \mathbb{J}^\top \boldsymbol{v}$. Similarly, Competitive Gradient Descent (CGD) [315] involves the computation of $\left(1 + \eta \nabla^2_{xy} f(\boldsymbol{x}_k, \boldsymbol{y}_k) \nabla^2_{yx} f(\boldsymbol{x}_k, \boldsymbol{y}_k)\right)^{-1}$ along with the Jacobian-vector product $\nabla^2_{xy} f(\boldsymbol{x}_k, \boldsymbol{y}_k) \nabla_y f(\boldsymbol{x}_k, \boldsymbol{y}_k)$. While the inverse term is approximated using conjugate gradient method in their implementation, it still involves the computation of approximately ten Jacobian-vector products for each update.

To explore these comparisons in greater detail and on models with many parameters, we experimentally compare the computational cost of our method with several other second as well as first-order methods on the 8-Gaussians problem in Figure 3.2 (architecture reported in Appendix B.7). We calculate the average wall-clock time (in milliseconds) per iteration. Results are reported on an average of 1000 iterations, computed on the same architecture and the same machine with forced synchronous execution. All the methods are implemented in PyTorch [279] and SGA is replicated based on the official implementation [4].

Furthermore, we observe that the computational cost per iteration of LEAD while being much lower than SGA and CGD, is similar to WGAN-GP and Extra-Gradient. The similarity to Extra-Gradient is due to the fact that for each player, Extra-Gradient requires the computation of a half-step and a full-step, so in total each step requires the computation of two gradients. LEAD also requires the computation of a gradient (∇f_x) which is then used to compute (∇f_{xy}) multiplied by ($\boldsymbol{y}_k - \boldsymbol{y}_{k-1}$). Using PyTorch, we do not require to compute ∇f_{xy} and then perform the multiplication. Given ∇f_x the whole term $\nabla f_{xy}(\boldsymbol{y}_k - \boldsymbol{y}_{k-1})$, is computed using PyTorch's Autograd with the computational cost of a single gradient. Thus, LEAD also requires the computation of two gradients for each step.

[4] SGA official DeepMind implementation (non-zero sum setting): `https://github.com/deepmind/symplectic-gradient-adjustment/blob/master/Symplectic_Gradient_Adjustment.ipynb`

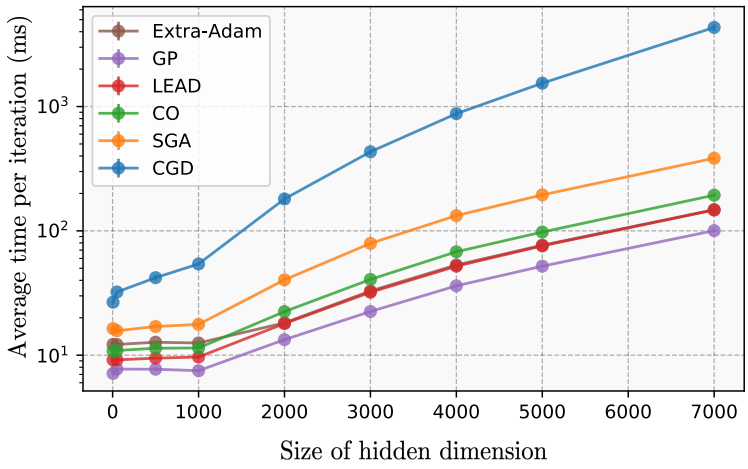

Figure 3.2: Average computational cost per iteration of several well-known methods for (non-saturating) GAN optimization. The numbers are reported on the 8-Gaussians generation task and averaged over 1000 iterations. Note that the y-axis is log-scale. We compare Competitive Gradient Descent (CGD) [315] (using official CGD optimizer code), Symplectic Gradient Adjustment (SGA) [32], Consensus Optimization (CO) [233], Extra-gradient with Adam (Extra-Adam) [115], WGAN with Gradient Penalty (WGAN GP) [130]. We observe that per-iteration time complexity of our method is very similar to Extra-Adam and WGAN GP and is much cheaper than other second order methods such as CGD. Furthermore, by increasing the size of the hidden dimension of the generator and discriminator's networks we observe that the gap between different methods increases.

3.5.2 Generative Adversarial Networks

8-Gaussians: We first compare LEAD-Adam with vanilla-Adam [183] on the generation task of a mixture of 8-Gaussians. Standard optimization algorithms such as vanilla-Adam suffer from mode collapse in this simple task, implying the generator cannot produce samples from one or several of the distributions present in the real data. Through

Fig. 3.3, we demonstrate that LEAD-Adam fully captures all the modes in the real data in both saturating and non-saturating losses.

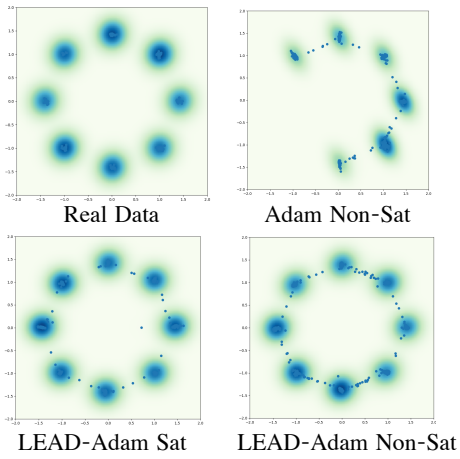

Figure 3.3: Performance of LEAD-Adam on the generation task of 8-Gaussians. All samples are shown after 10k iterations. Samples generated using Adam exhibit mode collapse, while LEAD-Adam does not suffer from this issue.

CIFAR-10: We additionally evaluate LEAD-Adam on the task of CIFAR-10 [189] image generation with a non-zero-sum formulation (non-saturating) on a DCGAN architecture similar to [130]. As shown in Table. 3.1, we compare with several first-order and second order methods and observe that LEAD-Adam outperforms the rest in terms of Fréchet Inception Distance (FID) [151][5], reaching a score of 19.27±0.10 which outperforms OMD [232] and CGD [315]. See also Figure 3.4.

Furthermore, we evaluate LEAD-Adam on more complex and deep architectures such as the ResNet architecture in [243]. We compare with several state of the art results on

[5]The FID is a metric for evaluating the quality of generated samples of a generative model. Lower FID corresponds to better sample quality.

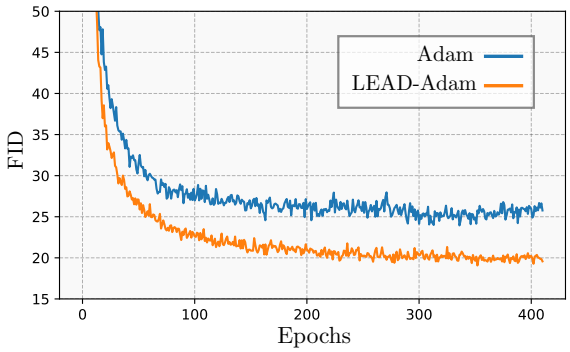

Figure 3.4: Plot showing the evolution of the FID over 400 epochs for our method (LEAD-Adam) vs vanilla Adam on a DCGAN architecture. Note that the FID is computed over 50k iterations.

the task of image generation on CIFAR-10 using ResNets. See Table 3.1 for a full comparison.

We report our results against a properly tuned version of SNGAN that achieves an FID of 12.36 using the code base of SNGAN PyTorch[6]. Our method obtains a competitive FID of 10.49.

We give a detailed description of these experiments and full detail on the architecture and hyper-parameters in Appendix B.7. See also Figure 3.5 for a sample of generated samples on a ResNet using LEAD.

3.6 Related Work

Game Optimization: With increasing interest in games, significant effort is being spent in understanding common issues affecting optimization in this domain. These issues range from convergence to non-Nash equilibrium points, to exhibiting rotational dynamics

[6] `https://github.com/GongXinyuu/sngan.pytorch`

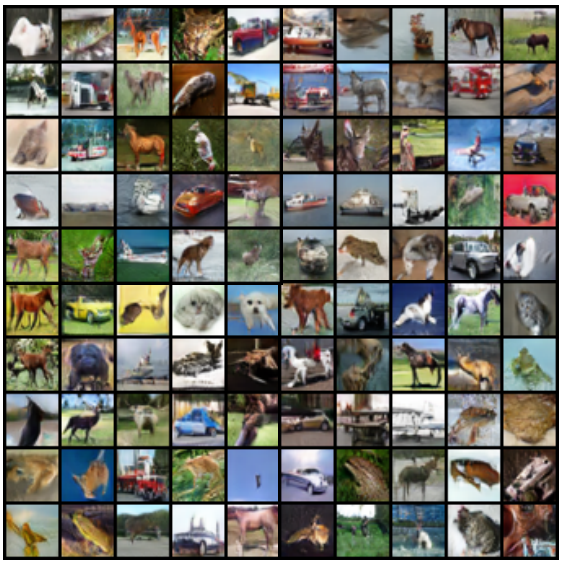

Figure 3.5: Generated sample of LEAD-Adam on CIFAR-10. LEAD-Adam achieves an FID of 10.49

around the equilibrium which hampers convergence. [233] provides a discussion on how the eigenvalues of the Jacobian govern the local convergence properties of GANs. They argue that the presence of eigenvalues with zero real-part and large imaginary part results in oscillatory behavior. To mitigate this issue, they propose Consensus Optimization (CO). Along similar lines, [32, 108, 210, 216] use the *Hamiltonian* of the gradient vector-field, to improve the convergence in games through disentangling the convergent parts of the dynamics from the rotational. Another line of attack taken in [315] is to use second-order information as a regularizer of the dynamics and motivate the use of Competitive Gradient Descent (CGD). In [356], Follow the Ridge (FtR) is proposed. They motivate the use of a second order term for one of the players (follower) as to avoid the rotational dynamics in a sequential formulation of the zero-sum game. See appendix B.8 for full discussion on the

comparison of LEAD versus other second-order methods.

Another approach taken by [116], demonstrate how applying negative momentum over GDA can improve convergence in min-max games, while also proving a linear rate of convergence in the case of bilinear games. [83] show that extrapolating the next value of the gradient using previous history, aids convergence. In the same spirit, [70], proposes LookAhead GAN (LA-GAN) and show that the LookAhead algorithm is a compelling candidate in improving convergence in GANs. [115] also explores this line of thought by introducing averaging to develop a variant of the extra-gradient algorithm and proposes Extra-Adam-Averaging. Similar to Extra-Adam-Averaging is SN-EMA [373] which uses the SN-GAN and achieves great performance by applying an exponential moving average on the parameters. Recently [344] proposes to train mixtures of BigGANs [60] to achieve state of the art performance on the task of image generation with GANs on CIFAR-10.

Lastly, in regard to convergence analysis in games, [122] provide last iterate convergence rate for convex-concave saddle point problems. [270] propose a multi-step variant of gradient descent-ascent, to show it can find a game's ϵ–first-order stationary point. Additionally, [25] and [159] provide spectral lower bounds for the rate of convergence in the bilinear setting for an accelerated algorithm developed in [26] for a specific families of bilinear games. Furthermore, [95] use Lyapunov analysis to provide convergence guarantees for gradient descent ascent using timescale separation and in [155], authors show that commonly used algorithms for min-max optimization converge to attractors that are not optimal.

Single-objective Optimization and Dynamical Systems: The authors of [339]

[6]For FtR, we provide the update for the second player given the first player performs gradient descent. Also note that in this table SGA is simplified for the two player zero-sum game. Non-zero sum formulation of SGA such as the one used for GANs require the computation of $\mathbb{J}\mathbf{v}, \mathbb{J}^\top\mathbf{v}$.

started a new trend in single-objective optimization by studying the continuous-time dynamics of Nesterov's accelerated method [263]. Their analysis allowed for a better understanding of the much-celebrated Nesterov's method. In a similar spirit, [361, 364] study continuous-time accelerated methods within a Lagrangian framework, while analyzing their stability using Lyapunov analysis. These works show that a family of discrete-time methods can be derived from their corresponding continuous-time formalism using various discretization schemes. Additionally, several recent work [250, 28, 223, 307] cast game optimization algorithms as dynamical systems so to leverage its rich theory, to study the stability and convergence of various continuous-time methods. [253] also analyzes the local stability of GANs as an approximated continuous dynamical system.

DCGAN	**FID**	**IS**
Adam	24.38 ± 0.13	
LEAD-Adam	19.27 ± 0.10	
CGD-WGAN [315]	21.3	7.2
OMD [83]	29.6 ± 0.19	5.74 ± 0.1
ResNet		
SNGAN	12.10 ± 0.31	8.58 ± 0.03
LEAD-Adam (ours)	**10.49** ± 0.11	**8.82** ± 0.05
ExtraAdam [115]	16.78 ± 0.21	8.47 ± 0.1
LA-GAN [70]	12.67 ± 0.57	8.55 ± 0.04
ODE-GAN [291]	11.85 ± 0.21	8.61 ± 0.06
Evaluated with 5k samples		
SN-GAN (DCGAN) [243]	29.3	7.42 ± 0.08
SN-GAN (ResNet) [243]	21.7 ± 0.21	8.22 ± 0.05

Table 3.1: Performance of several methods on CIFAR-10 image generation task. Methods that are not cited are reported using our own implementation where we compute and report the mean and standard-deviation over 5 random runs. The FID and IS is reported over 50k samples unless mentioned otherwise.

Chapter 4

Double Descent Phenomena: A Tale of Multi-scale Feature Learning Dynamics

Classical wisdom in statistical learning theory predicts a trade-off between the generalization ability of a machine learning model and its complexity, with highly complex models less likely to generalize well [98]. If the number of parameters measures complexity, deep learning models sometimes go against this prediction [See for example [379]]: deep neural networks trained by stochastic gradient descent exhibit a so-called *double descent* behavior [42] as with increasing model parameters. Specifically, with increasing complexity, the generalization error first obeys the traditional "U" shaped curve consistent with statistical learning theory. However, a second regime emerges as the number of parameters is further increased past a transition threshold where generalization error drops again, hence the

"double descent" or more specifically, *model-wise double descent* [254].

[254] show that double descent is not limited to varying model size but is also observed as the training time proceeds, specifically in the presence of label noise. Once again, the so-called *epoch-wise double descent* is in apparent contradiction with the classical understanding of overfitting [347], where one expects that longer training of a sufficiently large model beyond a certain threshold should result in overfitting. This phenomenon has important consequences for practitioners. It suggests the practice of early stopping, perhaps the most widely used regularization method in deep learning [124], prevents models from being trained at their fullest potential.

Although the term 'double descent' has been introduced recently to refer to such non-trivial behaviors of deep neural networks (DNNs), a similar phenomenon had already been studied in several decades-old works in the regression setting [190, 273, 274, 55] under a statistical physics framework. More recently, these behaviors have been investigated in the context of modern machine learning, both from an empirical [15, 372] and theoretical [4, 230, 91, 110, 90] point of view, to determine various limiting behavior of the training and generalization error.

In this work, we build upon early work on double descent, with roots in statistical physics, to provide a theory explaining epoch-wise double descent. Particularly,

- In Section 4.1, we introduce a novel linear data model in a teacher-student paradigm that, despite its simplicity, exhibits some of the puzzling properties of generalization dynamics in deep neural networks, namely epoch-wise double descent.
- In the limit of high dimensionality, we leverage the replica method of statistical physics

to derive closed-form expressions for the generalization dynamics of our teacher-student setup as a function of training time (Eq. (4.11)).

- Our theory provides an explanations for the existence of the epoch-wise double descent through the lens of multi-scale feature learning. Simply put, features that are learned on a faster time-scale are responsible for the conventional U-shaped generalization curve, while the second descent can be attributed to the features that are learned at a slower time-scale.
- Our analytical results closely match with simulations demonstrating epoch-wise double descent. To ensure reproducibility, we provide the code at: `GitHub repository`.

4.1 Theoretical Results

4.1.1 Prelude

This Section provides a theoretical framework to study the generalization dynamics of a high-dimensional regression model from a statistical physics perspective. Before diving into the theory, we invite the reader to recall a simple equation from thermodynamics. Consider an ideal gas in a container with its large number of molecules moving around, colliding with each other, all while obeying Newton's laws. While the exact dynamics of each of such molecules are intractable, the system's macroscopic behavior can be characterized in terms of a handful of scalar quantities, namely, the pressure P, the volume V, and the temperature T. By averaging over suitable probability measures and applying the principle of free-energy minimization, one consequently arrives at a remarkably simple relationship between these three macroscopic variables, i.e., the well-known $PV = nRT$ (n: number of

moles of gas, R: gas constant) [298].

The same principle can be ported to neural networks. Stochastic Gradient Descent (SGD) — the de facto optimization algorithm for neural networks — exhibits complex dynamics arising from the interactions between a large number of parameters [198]. However, **the idea** is to describe the high-dimensional *microscopic* dynamics of neural networks in terms of low-dimensional *macroscopic* entities. In a series of seminal papers by Gardner [100, 101, 102], the *replica method* of statistical physics was adopted to derive expressions describing the generalization behavior of large linear models trained using SGD. In this paper, we employ the so-called Gardner analysis to build upon an established line of work studying linear and generalized linear models [325, 175, 193]. While most of these previous works study the asymptotic generalization behavior in the limit of large training time, our contribution is to adapt these methods to study transient learning dynamics of generalization. We apply these tools to a simplified teacher-student model that exhibits key characteristics of modern neural network use cases, which we now describe.

4.1.2 A Teacher-Student Setup

Teacher: We study a supervised linear regression problem on a dataset $\mathcal{D}\{(\boldsymbol{x}^\mu, y^{*\mu})\}_{\mu=1}^n$ containing n training data-points generated by a linear teacher of width d (Fig. 4.1). Additionally, each input datapoint $\boldsymbol{x}^\mu \in \mathbb{R}^d$, is assumed to be sampled from an isotropic Gaussian distribution,

$$\boldsymbol{x}^\mu \ \sim \ \mathcal{N}(\mathbf{0}, \mathbb{I}_d), \tag{4.1}$$

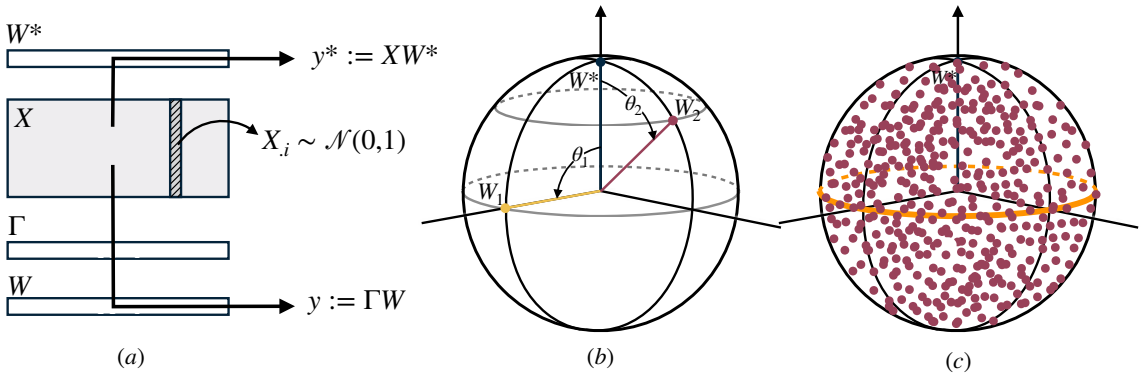

Figure 4.1: **(a):** A visual depiction of the teacher-student setup of Sec. 4.1.2. The linear teacher having access to the whole input $\boldsymbol{x}$, generates labels y^* via a weight vector $\boldsymbol{W}^*$. However, the student's access to the input is regulated through a pre-factor diagonal matrix Γ. (Only the diagonal of Γ is depicted for illustration) **(b, c):** An intuitive illustration of generalization behaviour using the macroscopic variables R and Q at initialization. We assume $Q = ||\boldsymbol{W}||_2^2 = 1$. R represents the cosine similarity ($cos(\theta_i)$) between the teacher's weight (taken to be along the z-axis) and the students weights W_i (red dots).

with the corresponding true label $y^{*\mu} \in \mathbb{R}$ being determined as,

$$y^{*\mu} := \frac{1}{\sqrt{d}} \ (\boldsymbol{x}^{\mu})^T \, \boldsymbol{W}^*, \tag{4.2}$$

where $\boldsymbol{W}^* \in \mathbb{R}^d$ represents the (fixed) weights of the teacher.

Student: Our student network is correspondingly chosen to be a similar shallow network, governed by trainable weights $\boldsymbol{W} \in \mathbb{R}^d$,

$$y^{\mu} := \frac{1}{\sqrt{d}} \ (\boldsymbol{x}^{\mu})^T \, (\Gamma \boldsymbol{W}), \tag{4.3}$$

with the important difference of an additional diagonal matrix[1] $\Gamma \in \mathbb{R}^{d \times d}$ ($\Gamma_{ii} \in [0, 1]$) acting on $\boldsymbol{W}$. This matrix can be perceived to be a mask, regulating the student's access to data features x_i^{μ}. With more features available, the student can learn a richer model.

[1]General matrix, SVD

Learning algorithm and Loss function: To train our student network, we use stochastic gradient descent (SGD) on the mean squared loss, defined on the n training examples as,

$$\mathcal{L}_{\mathcal{T}} := \frac{1}{2n}\sum_{\mu=1}^{n}(y^{*\mu} - y^{\mu})^2 \tag{4.4}$$

yielding the student weight update,

$$W_{k+1} = W_k - \eta\nabla_W \mathcal{L}_T\left(W_k\right) + \epsilon, \tag{4.5}$$

with k denoting the training step, η the learning rate while $\epsilon \sim \mathcal{N}(0, \sigma^2)$ models the stochasticity of our optimization scheme as an uncorrelated Gaussian noise.

The quantity of interest in this work, is the generalization error of the student model *on the entire task*, determined by averaging the network loss over all inputs $\boldsymbol{x}$:

$$\mathcal{L}_{\mathcal{G}} := \frac{1}{2}\mathbb{E}_{\boldsymbol{x}}\left[(y^*(\boldsymbol{x}) - y(\boldsymbol{x}))^2\right]. \tag{4.6}$$

Macroscopic variables: In the high-dimensional limit of $n, d \to \infty$ with a finite ratio $\alpha := \frac{n}{d}$, the generalization error of Eq. (4.6), is a function of scalar macroscopic variables H, R and Q defined as,

$$H = \frac{1}{d}\left(\boldsymbol{W}^*\right)^T\boldsymbol{W}^*, \tag{4.7}$$

$$R = \frac{1}{d}\left(\boldsymbol{W}^*\right)^T\left(\Gamma\boldsymbol{W}\right), \qquad Q = \frac{1}{d}\left(\Gamma\boldsymbol{W}\right)^T\left(\Gamma\boldsymbol{W}\right), \tag{4.8}$$

which as shown in [55, 190], lead to

$$\mathcal{L}_{\mathcal{G}}\left(R, Q\right) = \frac{1}{2}\left(H - 2RH + Q\right), \tag{4.9}$$

4.1.3 Main Result

In this Section, we present our main analytical results, with Section 4.1.4 containing a sketch of their derivations. (Detailed proofs in Appendix C.2).

Closed-from expressions. By employing the replica method of statistical physics [100, 101], we derive expressions for the dynamics of R and Q for the particular case,

$$\Gamma_{ii} = \begin{cases} \gamma_1, & \text{for } i = 1, p \\ \gamma_2, & \text{for } i = p+1, d \end{cases} \tag{4.10}$$

such that $\gamma_1 \gg \gamma_2$, to yield,

$$\begin{aligned} R(a_1, a_2, \alpha_1, \alpha_2) &:= \frac{\alpha_1}{a_1} H_1 + \frac{\alpha_2}{a_2} H_2, \\ Q(a_1, a_2, \alpha_1, \alpha_2) &:= \frac{\alpha_1}{a_1^2 - \alpha_1} \left(H_1 - \frac{2 - a_1}{a_1} \alpha_1 H_1 \right) + \frac{\alpha_2}{a_2^2 - \alpha_2} \left(H_2 - \frac{2 - a_2}{a_2} \alpha_2 H_2 \right). \end{aligned} \tag{4.11}$$

Here, H_1 and H_2 correspond to the parts of the teacher norm for the respective bi-partitions. While a_1 and a_2 are related to the learning time-scales t_1, t_2 of the two student bi-partitions, and are defined as (see Section 4.1.4),

$$\begin{aligned} a_1 &= \frac{\beta(Q_1^{(0)} - Q_1)}{1 + \beta(Q_1^{(0)} - Q_1)} \\ a_2 &= \frac{\beta(Q_2^{(0)} - Q_2)}{1 + \beta(Q_2^{(0)} - Q_2)} \end{aligned} \tag{4.12}$$

Here,

$$\begin{aligned} Q_1^{(0)} = \frac{1}{p} \sum_{i=1}^{p} \Gamma_{ii}^2 \langle W_i^2 \rangle, \; Q_1 = \frac{1}{p} \sum_{i=1}^{p} \Gamma_{ii}^2 \langle W_i \rangle^2 \\ Q_2^{(0)} = \frac{1}{d-p} \sum_{i=p+1}^{d} \Gamma_{ii}^2 \langle W_i^2 \rangle, \; Q_1 = \frac{1}{d-p} \sum_{i=p+1}^{d} \Gamma_{ii}^2 \langle W_i \rangle^2 \end{aligned} \tag{4.13}$$

Substituting Eqs. (4.11) into Eq. (4.9) then provides a closed-from analytic expression for the generalization error. These equations provide some qualitative insights: in high dimension,

generalization error is fully characterized by two scalar macroscopic observables, determined from the specific structure of the teacher and the student networks.

4.1.4 Sketch of derivations

In this Section, we sketch the key steps in the derivation of our main results.

Derivation of Eq. (4.9) The generalization error of Eq. (4.9) comprises an average over the input distribution $\boldsymbol{x}$. Since x's are i.i.d. and drawn from Gaussian distributions, the variables (y, y^*) is a bi-variate Gaussian with zero mean and a variance of, $\Sigma = \begin{bmatrix} 1 & R \\ R & Q \end{bmatrix}$, which implies a correlation between y and y^* obstructing the calculation of the average. Following [55, 190], we define decoupled variables $\tilde{y}^*$ and $\tilde{y}$ as follows,

$$y^* = \tilde{y}^*, \quad \text{and} \quad y = R\tilde{y}^* + \sqrt{Q - R^2}. \tag{4.14}$$

Simply replacing y^* and y into Eq. 4.6 and averaging over independent Gaussian variables of $\tilde{y}^*$ and $\tilde{y}$, result in Eq. 4.9.

Warm-up: Generalization at the initialization

It is instructive to start with studying the generalization at initialization. For the sake of clarity, here we assume the pre-factor matrix in Eq. 4.3 is $\Gamma = I$. For this warm-up exercise we also assume that the teacher and the student are initialized with unit norm, i.e., $||W^*||_2^2 = 1$ and $Q = ||W||_2^2 = 1$. That means that both the teacher and the student live on a d-dimensional sphere with unit radius. Since y^* and y in Eq. 4.6 are Gaussian variables with unit variances, $\mathcal{L}_{\mathcal{G}}$ is expected to be close to 2.

Since Q is assumed to be 1, the generalization error in Eq. 4.9 is only dependent on R. It is also useful to think about it from a geometrical perspective as advocated in [91].

Fig. 4.1 (b, c) depict that the variable R simply represents the cosine similarity between the teacher's weight and the students' weight. Since the teacher is only sampled once, we let W^* be aligned with the z-axis. Note that as R grows, the angle between the teacher and the student becomes smaller. This leads to lower generalization error.

Therefore, the question becomes: What is the typical value for R if we randomly initialize students? To answer it, we group students based on the angle they make with the teacher and simply count how many students fall into each group. It is evident that the majority of the students live near the equator and hence one may conclude that majority of the students have an $R = 0$. Substituting $R = 0$ into Eq. 4.9 shows that the typical generalization error is $L_G = 2.0$, the value we expected.

More formally, let $\Omega(R)$ denote the volume of the students with a cosine similarity of R with the teacher. With students randomly initialized over the surface of an d-dimensional sphere, it is straightforward to show the following,

$$\begin{aligned}\Omega(R) &:= \int d\boldsymbol{W}\ \delta(||W||_2^2 - 1)\ \delta\Big(\frac{1}{d}(W^*)^T W - R\Big) \\ &\propto \exp(d\big[\frac{1}{2}(1 + \ln 2\pi) + \frac{1}{2}(1 - R^2)\big]),\end{aligned} \tag{4.15}$$

in which d, the number of dimensions, appears in the exponent. This suggests that as the number of dimensions grows, $\Omega(R)$ becomes exponentially larger for the students that maximize the term inside brackets. It is consistent with the general intuition that in high-dimensions, every random student is perpendicular to the teacher with overwhelming probability.

Generalization during training

To track the generalization error during training, we first turn to the free energy of the system, which is defined to be the logarithm of the *partition function*, i.e., $f \propto -\ln(Z)$. The free energy is a self-averaging quantity where its mode coincides with its mean. Consequently, it allows us to compute the values of our macroscopic quantities at equilibrium. Next steps are: 1) Analytically derive the expression for the free-energy which is typically done by using the replica method of disordered systems [237], and, 2) Solve for values of R and Q which minimize the free energy and consequently provide us with the typical generalization behavior.

The free-energy. The first step is to note that SGD as defined in Eq. 4.5, in the long run, follow a Boltzmann distribution over the student weight, $P(W) \propto \exp(-\mathcal{L}_T/T)$, where T denotes the temperature and $\mathcal{L}_T$ is the training loss. It signifies that for a large T, the distribution of $P(W)$ is almost uniform while as $T \to 0$, $P(W)$ becomes more concentrated around the minimum (minima) of the training loss. While most of work in this literature study the case of zero temperature, here we take the approach of [55] to provide more general expressions dependent on the temperature.

The partition function is then defined as,

$$Z := \int e^{-n\beta E_{\mathcal{T}}} \delta\left(\Gamma \boldsymbol{W} - dQ\right) d\boldsymbol{W}, \tag{4.16}$$

in which δ is the Dirac delta function. The free-energy is the self-average of the logarithm of the partition function w.r.t. n training examples,

$$f := -\left\langle \left\langle \frac{\ln Z}{\beta n} \right\rangle \right\rangle_{\boldsymbol{W}^*, \boldsymbol{x}}. \tag{4.17}$$

Since logarithm is inside the expectation, analytical computation of Eq. 4.17 is intractable. However, the replica method allows us to take the logarithm outside according to the following identity also referred to as the replica trick,

$$\mathbb{E}\left[\ln Z\right] = \lim_{r \to 0} \frac{\mathbb{E}\left[Z\right]^r - 1}{r}. \tag{4.18}$$

Accordingly, the analytical expression for the free-energy reads,

$$\begin{aligned} -\beta f = &\frac{1}{2}\frac{Q_1 - R_1^2}{Q_1^{(0)} - Q_1} + \frac{1}{2}\ln(Q_1^{(0)} - Q_1) - \frac{\alpha}{2}\ln\left[1 + \beta(Q_1^{(0)} - Q_1)\right] - \frac{\alpha\beta}{2}\frac{H_1 - 2R_1H_1 + Q_1}{1 + \beta(Q_1^{(0)} - Q_1)}. \\ &+ \frac{1}{2}\frac{Q_2 - R^2}{Q_2^{(0)} - Q_2} + \frac{1}{2}\ln(Q_2^{(0)} - Q_2) - \frac{\alpha}{2}\ln\left[1 + \beta(Q_2^{(0)} - Q_2)\right] - \frac{\alpha\beta}{2}\frac{H_2 - 2R_2H_2 + Q_2}{1 + \beta(Q_2^{(0)} - Q_2)} \end{aligned} \tag{4.19}$$

Solutions for R and Q. Given the analytical form of the free-energy, we can simply solve for values of R and Q that minimize the free-energy to obtain the expressions of Eq. (4.11).

Intuitively, one may see that as the temperature drops, variables R and Q reach their asymptotic values. Hence, the last piece of puzzle is to formalize the relationship between the temperature, T, and the training time, t.

Training time is inversely proportional to the temperature. For a linear model, such as the one in Eq. 4.3, t iterations of gradient descent is shown to be equivalent to the same model trained with L_2 regularization with a coefficient inversely proportional to training time. Formally, Thm. 3 of [10] provides an *upper bound* on the excess risk of stochastic gradient flow at time t, over regularized regression $\lambda = 1/t$, for all $t \geq 0$.

Accordingly, to incorporate training time in the free-energy, one may add an L_2 regularization to the training loss with a coefficient of $\lambda = 1/t$. Hence the expressions for

the loss and the free-energy are updated as,

$$\mathcal{L}_T \to \mathcal{L}_T + \frac{1}{2t}||\boldsymbol{W}||_2^2 \quad \text{, and} \quad f \to f + \frac{Q_1^{(0)}}{2\gamma_1^2 t} + \frac{Q_2^{(0)}}{2\gamma_2^2 t}. \tag{4.20}$$

Now, we can minimize the free-energy w.r.t. Q_0 and derive an expression for the temperature T as a function of time, t.,

$$\nabla_{Q_1^{(0)}} f = 0 \;\Rightarrow\; a_1 = \frac{1}{b_1} + 1 \quad \Rightarrow\; b_1 = \frac{1 - \alpha_1 - \frac{1}{\gamma_1^2 t} \pm \sqrt{\left(1 - \alpha_1 - \frac{1}{\gamma_1^2 t}\right)^2 + \frac{4}{\gamma_1^2 t}}}{\frac{2}{\gamma_1^2 t}} \tag{4.21}$$

$$\nabla_{Q_2^{(0)}} f = 0 \;\Rightarrow\; a_2 = \frac{1}{b_2} + 1 \;\Rightarrow\; b_2 = \frac{1 - \alpha_2 - \frac{1}{\gamma_2^2 t} \pm \sqrt{\left(1 - \alpha_2 - \frac{1}{\gamma_2^2 t}\right)^2 + \frac{4}{\gamma_2^2 t}}}{\frac{2}{\gamma_2^2 t}} \tag{4.22}$$

which coincides with the intuition that training longer implies lower temperature.

4.2 Experimental Results

In this Section, we provide numerical simulations to validate our analytical theory. Furthermore, we demonstrate that our teacher-student setup exhibits generalization behavior which is qualitatively similar to that of deep neural networks. The experiments are designed to provide a better understanding of the epoch-wise double descent phenomenon.

In simulations and evaluation of the theory, we set the number of training examples $n = 150$, the dimensionality of data $d = 200$. Also the cutting point of the diagonal elements of Γ matrix is set to $p = 100$. Numerical simulations are averaged over 100 random runs. App. C.3 provides additional examples with different data and model settings. To ensure reproducibility of the results, we provide further details in App. C.3 and include the complete source code in a `GitHub repository`.

4.2.1 Match between theory and simulations

We conduct an experiment on the classification task of CIFAR-10 [189] and monitor the generalization error (0-1 test error) during the course of training. We follow the setup of [254] and add 15% random label noise to the training set, which leads to epoch-wise double descent. Fig. 4.2 **(Left)** depicts the generalization curve for two models of ResNet-18 [143] with different widths. It can be observed that the network with a smaller width displays a typical overfitting behavior in which the generalization error decreases first and then overfitting occurs, resulting in worse generalization. However, the network with a larger width exhibits a double descent generalization curve, *i.e.*, with more training, the generalization error will eventually improve.

To compare with the teacher-student setup in Sec. 4.1.2, we consider the following two cases for the pre-factor diagonal matrix Γ,

$$\Gamma_{ii}^{(A)} = \begin{cases} 1 \quad \text{for} \quad i < p \\ \\ 0 \quad \text{for} \quad i \geq p \end{cases} \text{, and} \quad \Gamma_{ii}^{(B)} = \begin{cases} 1 \quad \text{for} \quad i < p \\ \\ 0.1 \quad \text{for} \quad i \geq p \end{cases} . \tag{4.23}$$

Fig. 4.2 **(Right)** presents the analytical generalization dynamics of Eq. 4.11 for the two cases above and provides comparison between the theory and simulation results of the same model. We observe that the theory and simulations accurately match. We also note that Fig. (Right) and (Left) qualitatively match with each other, suggesting that the proposed teacher-student model is a valid approximation of deep neural networks' generalization dynamics.

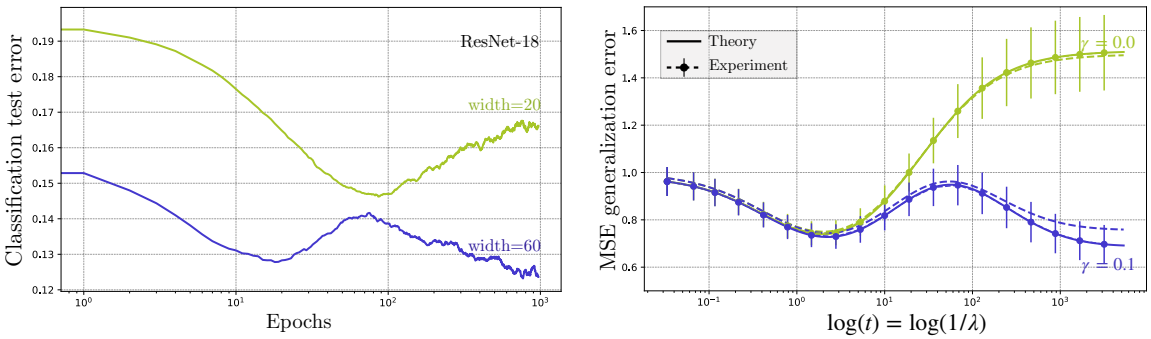

Figure 4.2: Comparison between generalization performance predicted by theory and ResNet-18 on CIFAR-10, as function of training time. We observe that the qualitative dynamics match on the left and right plots. **Left:** Generalization curves for a ResNet-18 model, following the setup in [254], we add 15% label noise. The plots depict two networks with different width. The green curve corresponds to a network with width 20. It undergoes a typical over-fitting behavior. The blue curve corresponds to a network with width 60 which undergoes two descents and with more training the generalization error eventually improves. **Right:** The teacher-student set-up in Sec. (4.1.2). We compare the analytical solutions in Eq. 4.11 to simulations performed on our teacher-student setup with $d = 200$, $p = 100$, $n = 150$ and we plot the error bars over 100 random seeds. The solutions and the simulations match closely and we observe double descent over the generalization error in both cases of the blue curve.

4.2.2 The Phase diagram

To further understand the transition between the two phases of *decent-ascent* and *decent-ascent-descent*, we explore the phase diagram.

Before discussing the structure of the phase diagram, let us highlight the fact that given Eq. 4.9, one can fully characterize the evolution of the generalization dynamics in terms of two scalar variables instead of the d-dimensional parameter space. R and Q presented in Eq. 4.8 are macroscopic variables that together represent the angle between the student and the teacher. Hence, a better generalization performance is achieved with

larger R and smaller Q.

Fig. 4.3 illustrates the generalization loss for all pairs of $(R, Q) \in [0.0, 0.8] \times [0.0, 1.6]$. However, R and Q are not free parameters and both depend on the training dynamics through Eq. 4.11. Specifically, at the time of initialization, $(R, Q) = (0, 0)$ as the students are initialized at zero. As training time proceeds, values of R and Q follow the depicted trajectories. In Fig. 4.3, different trajectories correspond to different choices of the pre-factor matrix Γ where,

$$\Gamma_{ii} = \begin{cases} 1 \quad \text{for} \quad i < p \\ \\ \gamma \quad \text{for} \quad i \geq p \end{cases}, \text{ for } \gamma \in [0.0, 0.1]. \tag{4.24}$$

The yellow curve which corresponds to the case with $\gamma = 0$ exhibits traditional over-fitting due to over-training, i.e., the yellow trajectory starts at $(0, 0)$ and moves towards Point A which has the lowest generalization error of this curve. Then as the training continues, Q increases and as $t \to \infty$ the trajectory lands at Point B which has the worse generalization error.

The curves in orange, green and blue correspond to trajectories of $\gamma > 0$. They follow the case of $\gamma = 0$ up to the vicinity of Point B, but then the trajectories slowly incline towards another fixed point, Point C signalling a better generalization performance.

The phase diagram along with the corresponding generalization curves in Fig. 4.2 suggest that features that are learned on a faster time-scale are responsible for the conventional U-shaped generalization curve, while the second descent can be attributed to the features that are learned at a slower time-scale.

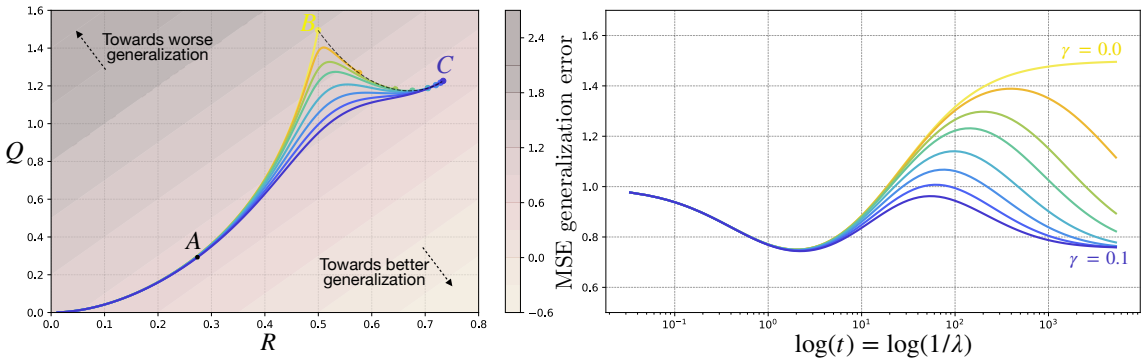

Figure 4.3: **Left:** Phase diagram of the generalization error as a function of R and Q (Eq. (4.8)). The generalization loss for all pairs of $(R, Q) \in [0.0, 0.8] \times [0.0, 1.6]$ is contour-plotted in the background in shades of beige, with the best generalization performance being attained on the lower right part of the plot. The trajectories, starting from $(0, 0)$, on the other hand, correspond to the values of R and Q as training proceeds. Each trajectory correspond to a different choice of γ in Eq. (4.24), with $\gamma = 0$ (bright yellow) exhibiting traditional over-fitting, while for $\gamma > 0$ the test error demonstrates epoch-wise double descent **Right:** The corresponding generalization curves for different values of $\gamma \in [0, 0.1]$.

4.3 Related Work

If we consider plots where the generalization error on the y-axis is plotted against other quantities on the x-axis, we find earlier works that have identified double descent behavior for quantities such as the number of parameters, the dimensionality of the data, the number of training samples, or the training time on the x-axis. In this paper, we study epoch-wise double descent phenomena, *i.e.* we plot the training time t, or the number of training epochs, on the x-axis. Literature displaying double descent phenomena in generalization behavior *wrt* other quantities do so in the limit of $t \to \infty$. Nevertheless, since our work builds upon past studies, it is relevant to review different perspectives taken in the existing literature towards studying other forms of non-monotonicity of the generalization error.

Random matrix theory perspective. [205, 141, 5], and [44] are among works which have analytically studied the spectral density of the Hessian matrix. According to their analyses, at intermediate levels of complexity, the presence of small but non-zero eigenvalues in the Hessian matrix results in high generalization error as the inverse of the Hessian is calculated for the pseudo-inverse solution. In an influential work, [230] extend the same analysis to a random feature model and theoretically derive the model-wise double descent curve for a model with Tikhonov regularization. [168] also study double descent in ridge estimators and show an equivalence to kernel ridge regression.

Bias/variance trade-off. [107], and more recently, [261] empirically observe that while bias is monotonically decreasing, variance could be decreasing too or unimodal as the number of parameters increases, thus manifesting a double descent generalization curve. [141] analytically study the variance. More recently, [372] provides a new bias/variance decomposition of bias exhibiting double desc-nt in which the variance follows a bell-shaped curve. However, the decrease in variance as the model size increases remains unexplained. For high dimensional regression with random features, [90] provides an asymptotic expression for the bias/variance decomposition and identifies three sources of variance with non-monotonous behavior as the model size or dataset size varies. [82] also employs the analysis of random feature models and identifies two forms of overfitting which leads to the so-called sample-wise triple descent. More recently, [71] show that as a result of the interaction between the data and the model, one may design generalization curves with multiple descents.

Statistical physics perspective. [273, 54, 55, 274] are among the first studies which theoretically observe sample-wise double-descent in a ridge regression setup where the

solution is obtained by the pseudo-inverse method. Most of these studies employ the "Gardner analysis" [100, 101, 102] for models where the number of parameters and the dimensionality of data are coupled and hence the observed form of double descent is different from that observed in deep neural networks. A beautiful extended review of this line of work is provided in [91]. Among recent works, [110] also apply the Gardner analysis but to a novel generalized data generating process called the hidden manifold model and derive the model-wise double-descent equations analytically. In this work, we also leverage the tools of statistical physics, namely the replica method. We adopt the approach of [55] for introducing training time into the equations through finite-temperature learning in our teacher-student setup.

Other related work. On the empirical side, [254], study the different forms of double descent in deep networks; this is the first work that identifies epoch-wise double descent and shows that it manifests in the presence of noisy training labels. In our experiments on deep networks, we follow [254] and explicitly add label noise to the training data to simulate the noise that can naturally occur in wild datasets.

Towards providing an explanation for the epoch-wise double descent, we argue that *the epoch-wise double descent can be attributed to different features being learned at different time-scales*, resulting in a non-monotonous generalization curve.

A related result was obtained by [144] in the framework of the bias/variance decomposition. The authors argue that epoch-wise double descent is observed due to the overlapping of two or more U-shape generalization curves with different minimas. Several other related works have identified that different model components are learned at different

rates depending on the data and model structure. [313, 314, 4, 5, 114, 199, 66] study the learning dynamics of linear or linearized networks and show that learning along each principal component of the NTK [167] or input covariance matrix progresses at a different rate. [285] show that different rates of learning at the presence of cross-entropy loss could result in failure to capture slower-learning features.

Chapter 5

Conclusions

In the following we provide concluding remarks on each of the three distinct topics discussed in the thesis.

Fluctuations and magnetoresistance oscillations near the half-filled Landau level: In this work, we studied theoretically commensurability oscillations about $\nu = 1/2$ that are produced by a one-dimensional scalar potential using the Dirac composite fermion theory. Through an approximate large N analysis of the Schwinger–Dyson equations, we considered how corrections to Dirac composite fermion mean-field theory affect the behavior of the predicted oscillations. We focused on corrections arising from the exchange of an emergent gauge field whose low-energy kinematics satisfy $|\vec{q}| \leq |q_0|$. In addition, we only considered screened electron-electron interactions. Remarkably within this restricted parameter regime, we found a self-consistent solution to the Schwinger–Dyson equations in which a Chern-Simons term for the gauge field and mass for the Dirac composite fermion are dynamically generated. The Dirac mass resulted in a correction to the locations of the commensurability

oscillation minima which improved comparison with experiment.

There are a variety of directions for future exploration. It would be interesting to consider the effects of the exchange of emergent gauge fields with $|q_0| < |\vec{q}|$. In this regime, Landau damping of the "magnetic" component of the gauge field propagator is expected to result in IR dominant Dirac composite fermion self-energy corrections [208, 234, 249]. In particular, it would be interesting to understand this regime when a dynamically-generated Chern-Simons term for the gauge field is present. These studies are expected to be highly sensitive to the nature of the electron-electron interactions. At $\nu = 1/2$ when the effective magnetic field vanishes, single-particle properties depend upon whether this interaction is short or long ranged [182]. It is important to understand the interplay of this physics with a non-zero effective magnetic field that is generated away from $\nu = 1/2$ and its potential observable effects.

The corrections to the predicted commensurability oscillations relied on a solution to the Schwinger–Dyson equations, obtained in a large N flavor approximation, that was extrapolated to $N = 1$. The study of higher-order in $1/N$ effects may provide additional insight into the validity of this extrapolation. Alternatively, study of the 't Hooft large N limit of the Dirac composite fermion theory dual conjectured in [157] may complement our analysis.

Recent works [351, 197, 195] have shown that PH symmetry at $\nu = 1/2$ and reflection symmetry about $\nu = 1/2$ rely on precisely correlated electric and magnetic perturbations. (This correlation is implemented by the Chern-Simons gauge field in the HLR theory.) Specifically, a periodic scalar potential $V(\mathbf{x})$ generates a periodic magnetic

flux $b(\mathbf{x})$ via,

$$b(\mathbf{x}) = -2m^* V(\mathbf{x}). \tag{5.1}$$

How might fluctuations about HLR mean-field theory affect Eq. (5.1) and potentially modify its predicted commensurability oscillations and other observables?

LEAD: Min-Max Optimization from a Physical Perspective: In this work, we leverage tools from physics to propose a novel second-order optimization scheme LEAD, to address the issue of rotational dynamics in min-max games. By casting min-max game optimization as a physical system, we use the principle of least action to discover an effective optimization algorithm for this setting. Subsequently, with the use of Lyapunov stability theory and spectral analysis, we prove LEAD to be convergent at a linear rate in bilinear min-max games. We supplement our theoretical analysis with experiments on GANs, demonstrating improvements over baseline methods. Specifically for GAN training, we observe that our method outperforms other second-order methods, both in terms of sample quality and computational efficiency.

Our analysis underlines the advantages of physical approaches in designing novel optimization algorithms for games as well as for traditional optimization tasks. It is important to note in this regard that our crafted physical system is *a* way to model min-max optimization physically. Alternate schemes to perform such modeling can involve other choices of counter-rotational and dissipative forces which can be explored in future work.

Double Descent Phenomena: A Tale of Multi-scale Feature Learning Dynamics: Leveraging tools from statistical physics such as Gardner analysis and Replica method, we derive explicit equations for the generalization error as a function of model size and training

time in a teacher-student setup. We believe our analysis introduces a convenient approach to study the generalization dynamics of neural networks. We provide important insights while characterizing some of the aspects of deep neural networks with simple analytical equations. Particularly, we provide an explanation for the epoch-wise double descent by characterizing the dynamics of a teacher-student setup with two microscopic variables. In short, epoch-wise double descent can be explained by the interaction of different learning speeds for subsets of features in the data.

Limitations. It should be noted that studying finer details of the dynamics would require a more precise model of the neural networks. Clearly, our proposed model is not a universal and unique way to model the dynamics of the complex, over-parameterized deep neural networks.

Appendix A

Chapter 2: Appendix

A.1 Integrals

In this appendix, we give details for the calculations of the gauge and fermion self-energy integrals quoted in the main text.

A.1.1 Gauge field self-energy

We begin with the gauge field self-energy given in §2.2.3. We are interested in computing the PH odd component of the gauge field self-energy Π_{odd}:

$$\Pi^{\alpha\beta}(q) = \Pi^{\alpha\beta}_{\text{even}}(q) + i\epsilon^{\alpha\beta\tau}q_\tau\Pi_{\text{odd}}(q). \tag{A.1}$$

To leading order in b, we substitute $G(p) = G^{(0)}(p)$ from Eq. (2.17) with $\Sigma_\alpha = 0$ for $\alpha \in \{0, 1, 2\}$ into Eq. (2.30):

$$\begin{aligned} i\epsilon^{\alpha\beta\tau} q_\tau \Pi_{\text{odd}}(q) &= N\Big\{ \int \frac{d^3p}{(2\pi)^3} \text{tr}\Big[\gamma^\alpha G^{(0)}(p)\gamma^\beta G^{(0)}(p+q)\Big]\Big\}_{\text{odd}} \\ &= -N\Big\{ \int \frac{d^3p}{(2\pi)^3} \text{tr}\Big[\gamma^\alpha \frac{i(\gamma^\sigma(p_\sigma + \mu_0\delta_{\sigma,0}) + \Sigma_m)}{(p_0+\mu_0)^2 - p_i^2 - \Sigma_m^2}\gamma^\beta \\ &\times \frac{i(\gamma^\tau(p_\tau + q_\tau + \mu_0\delta_{\tau,0}) + \Sigma_m)}{(p_0+q_0+\mu_0)^2 - (p_i+q_i)^2 - \Sigma_m^2}\Big]\Big\}_{\text{odd}}. \end{aligned} \tag{A.2}$$

We have suppressed the $i\epsilon_{p_0}$ factor in Eq. (2.17) that defines the Feynman contour for the Minkowski-signature p_0 integration because we will will evaluate the above integral in Euclidean signature. In subsequent sections of this appendix, we will likewise suppress the $i\epsilon_{p_0}$ factor for the same reason without further comment. Recall that the factor of N arises from the fermion loop over N flavors of Dirac composite fermions and that $\mu_0 > 0$.

To leading order in the derivative expansion, i.e., $\Pi_{\text{odd}}(q = 0)$, the expression for $\Pi_{\text{odd}}(0)$ simplifies to

$$\Pi_{\text{odd}}(0) = -2iN\Sigma_m \int \frac{d^3p}{(2\pi)^3} \frac{1}{\Big((p_0+\mu_0)^2 - p_i^2 - \Sigma_m^2\Big)^2}. \tag{A.3}$$

Here, we have used the trace identities,

$$\begin{aligned} \text{tr}\Big[\gamma^\alpha\gamma^\beta\Big] &= 2\eta^{\alpha\beta}, \\ \text{tr}\Big[\gamma^\alpha\gamma^\beta\gamma^\tau\Big] &= -2i\epsilon^{\alpha\beta\tau}. \end{aligned} \tag{A.4}$$

To compute this integral, we first Wick rotate, $p_0 \mapsto i(p_E)_3$ and $d^3p \mapsto id^3p_E$, and then

sequentially integrate over $(p_E)_3$ and the spatial momenta $(p_E)_i$ $(i = 1, 2)$ to find:

$$\begin{aligned}
\Pi_{\text{odd}}(0) &= 2N\Sigma_m \int \frac{d^3p_E}{(2\pi)^3} \frac{1}{(i(p_E)_3 + \mu_0)^2 - (p_E)_i^2 - \Sigma_m^2)^2} \\
&= 2N\Sigma_m \int \frac{d^3p_E}{(2\pi)^3} \frac{1}{((p_E)_3 - \omega_+)^2((p_E)_3 - \omega_-)^2} \\
&= \frac{N\Sigma_m}{2} \int \frac{d^2p_E}{(2\pi)^2} \frac{\Theta(|\Sigma_m| - \mu_0) + \Theta(\mu_0 - |\Sigma_m|)\Theta(|(p_E)_i| - \sqrt{\mu_0^2 - \Sigma_m^2})}{(|(p_E)_i|^2 + \Sigma_m^2)^{3/2}} \\
&= \frac{N}{4\pi}\Big(\Theta(|\Sigma_m| - \mu_0)\frac{\Sigma_m}{|\Sigma_m|} + \Theta(\mu_0 - |\Sigma_m|)\frac{\Sigma_m}{\mu_0}\Big),
\end{aligned} \tag{A.5}$$

where the step function $\Theta(|(p_E)_i| - \sqrt{\mu_0^2 - \Sigma_m^2})$ in the third line ensures the double poles $\omega_\pm = i(\mu_0 \pm \sqrt{(p_E)_i^2 + \Sigma_m^2})$ occur on opposite sides of the real $(p_E)_3$ axis. Eq. (A.5) implies that, for $\mu_0 > |\Sigma_m| > 0$, the gauge field obtains a correction to its propagator that corresponds to an effective Chern-Simons term with level,

$$k = \frac{N}{2}\frac{\Sigma_m}{\mu_0}. \tag{A.6}$$

A.1.2 Fermion self-energy

Next, we calculate the fermion self-energies Σ_m and Σ_0 quoted in §2.2.4.

We begin with Σ_m. Taking the trace of both sides of Eq. (2.29) and setting $\delta q_\alpha = 0$, we find:

$$i\Sigma_m(q_{\text{FS}}) = i\mathcal{M}^{(0)}(q_{\text{FS}}) + i\mathcal{M}^{(1)}(q_{\text{FS}}), \tag{A.7}$$

where

$$i\mathcal{M}^{(0)}(q_{\text{FS}}) = \frac{1}{2}\int \frac{d^3p}{(2\pi)^3}\text{tr}\Big[\gamma^\alpha G^{(0)}(p + q_{\text{FS}})\gamma^\beta\Big(\frac{2\pi}{k}\frac{\epsilon_{\alpha\beta\sigma}p^\sigma}{p^2}\Big)\Big], \tag{A.8}$$

$$i\mathcal{M}^{(1)}(q_{\text{FS}}) = \frac{1}{2}\int \frac{d^3p}{(2\pi)^3}\text{tr}\Big[\gamma^\alpha G^{(1)}(p + q_{\text{FS}})\gamma^\beta\Big(\frac{2\pi}{k}\frac{\epsilon_{\alpha\beta\sigma}p^\sigma}{p^2}\Big)\Big], \tag{A.9}$$

$G^{(0}(p)$ and $G^{(1)}(p)$ are given in Eqs. (2.17) and (2.18), k is given in Eq. (A.6), and $q_{\rm FS} = (0, \mu_0\hat{n})$ for the unit vector $\hat{n}$ (e.g., $\hat{n} = (\cos(\varphi), \sin(\varphi))$ where φ parameterizes a point on the Fermi surface) normal to the (assumed) spherical Fermi surface. As before, we set $\Sigma_\alpha = 0$ for $\alpha \in \{0, 1, 2\}$ and only retain Σ_m when using $G^{(0)}(p)$ and $G^{(1)}(p)$ to evaluate $\mathcal{M}^{(0)}$ and $\mathcal{M}^{(1)}$, as well as Σ_0 below. It is convenient to define $Q = (\mu_0, \mu_0\hat{n})$ so that

$$i\mathcal{M}^{(0)}(q_{\rm FS}) = \frac{1}{2}\int \frac{d^3p}{(2\pi)^3}\mathrm{tr}\Big[\gamma^\alpha\Big(\frac{i(\gamma^\sigma(p+Q)_\sigma + \Sigma_m)}{(p+Q)^2 - \Sigma_m^2}\Big)\gamma^\beta\Big(\frac{2\pi}{k}\frac{\epsilon_{\alpha\beta\tau}p^\tau}{p^2}\Big)\Big], \tag{A.10}$$

$$i\mathcal{M}^{(1)}(q_{\rm FS}) = \frac{1}{2}\int \frac{d^3p}{(2\pi)^3}\mathrm{tr}\Big[\gamma^\alpha\Big(\frac{ib(\mathbb{I}(p+Q)_0 + \gamma^0\Sigma_m)}{((p+Q)^2 - \Sigma_m^2)^2}\Big)\gamma^\beta\Big(\frac{2\pi}{k}\frac{\epsilon_{\alpha\beta\tau}p^\tau}{p^2}\Big)\Big]. \tag{A.11}$$

We first consider $\mathcal{M}^{(0)} = \mathcal{M}^{(0)}(q_{\rm FS})$. Using the trace identities in Eq. (A.4), we find

$$\begin{aligned} i\mathcal{M}^{(0)} &= \frac{\pi i}{k}\int \frac{d^3p}{(2\pi)^3}\mathrm{tr}\Big[\gamma^\alpha\Big(\frac{(\gamma^\sigma(p+Q)_\sigma + \Sigma_m)}{(p+Q)^2 - \Sigma_m^2}\Big)\gamma^\beta\Big(\frac{\epsilon_{\alpha\beta\tau}p^\tau}{p^2}\Big)\Big] \\ &= -\frac{4\pi}{k}\int \frac{d^3p}{(2\pi)^3}\frac{(p+Q)_\sigma p^\sigma}{((p+Q)^2 - \Sigma_m^2)p^2}. \end{aligned} \tag{A.12}$$

Next, we combine denominators using the Feynman parameter x and then shift the integration by defining $\ell_\alpha = p_\alpha + Q_\alpha x$:

$$\begin{aligned} i\mathcal{M}^{(0)} &= -\frac{4\pi}{k}\int \frac{d^3p}{(2\pi)^3}\int_0^1 dx \frac{(p+Q)_\sigma p^\sigma}{(p^2 + 2p\cdot Qx + Q^2x - \Sigma_m^2 x)^2} \\ &= -\frac{4\pi}{k}\int \frac{d^3\ell}{(2\pi)^3}\int_0^1 dx \frac{\ell^2 + \ell\cdot Q(1-2x) - x(1-x)Q^2}{(\ell^2 + Q^2x(1-x) - \Sigma_m^2 x)^2} \\ &= -\frac{4\pi}{k}\int \frac{d^3\ell}{(2\pi)^3}\int_0^1 dx \frac{\ell^2}{(\ell^2 - \Sigma_m^2 x)^2}, \end{aligned} \tag{A.13}$$

where we evaluated $Q^2 = 0$ and dropped the linear in ℓ term in the third line since it vanishes upon integration over ℓ. Next, we Wick rotate by taking $\ell_0 \mapsto i(\ell_E)_3$, $\ell^2 \mapsto -\ell_E^2$, and $d^3\ell \mapsto id^3\ell_E$, integrate over ℓ_E via dimensional regularization, and finally integrate over

x:

$$\begin{aligned} i\mathcal{M}^{(0)} &= \frac{4\pi i}{k}\int\frac{d^3\ell_E}{(2\pi)^3}\int_0^1 dx\frac{\ell_E^2}{(\ell_E^2+\Sigma_m^2 x)^2} \\ &= -\frac{12\pi^{3/2}i|\Sigma_m|}{k(4\pi)^{3/2}}\int_0^1 dx\ x^{1/2} \\ &= -i\frac{|\Sigma_m|}{k} \\ &= -i\frac{2\mu_0\text{sign}(\Sigma_m)}{N}, \end{aligned} \tag{A.14}$$

where we substituted in the Chern-Simons level given in Eq. (A.6) in the final line.

Next, consider $\mathcal{M}^{(1)} = \mathcal{M}^{(1)}(q_{\text{FS}})$. Using the trace identities in Eq. (A.4), we find

$$\begin{aligned} i\mathcal{M}^{(1)} &= -\frac{4\pi b\Sigma_m}{k}\int\frac{d^3p}{(2\pi)^3}\frac{p_0}{((p+Q)^2-\Sigma_m^2)^2p^2} \\ &= -\frac{4\pi b\Sigma_m}{k}I(\Sigma_m^2,Q). \end{aligned} \tag{A.15}$$

With the help of the formal identity,

$$I(\Sigma_m^2,Q) = -\partial_{\Sigma_m^2}J(\Sigma_m^2,Q) = -\partial_{\Sigma_m^2}\int\frac{d^3p}{(2\pi)^3}\frac{p_0}{((p+Q)^2-\Sigma_m^2)p^2}, \tag{A.16}$$

we rewrite

$$i\mathcal{M}^{(1)} = \frac{4\pi b\Sigma_m}{k}\partial_{\Sigma_m^2}\int\frac{d^3p}{(2\pi)^3}\frac{p_0}{((p+Q)^2-\Sigma_m^2)p^2}. \tag{A.17}$$

This integral has the same basic form as the one we encountered in calculating $\mathcal{M}^{(0)}$ and so we will follow the same steps as before: combine denominators with the Feynman parameter x, shift the integration $\ell_\alpha = p_\alpha + Q_\alpha x$, and substitute in $Q_0 = \mu_0$ and $Q^2 = 0$:

$$i\mathcal{M}^{(1)} = -\frac{4\pi b\Sigma_m\mu_0}{k}\partial_{\Sigma_m^2}\int\frac{d^3\ell}{(2\pi)^3}\int_0^1 dx\frac{x}{(\ell^2-\Sigma_m^2 x)^2}. \tag{A.18}$$

Next, we Wick rotate by taking $\ell_0 \mapsto i(\ell_E)_3$, integrate over ℓ_E via dimensional regularization, integrate over x, take the derivative with respect to Σ_m^2, and then evaluate $k = \frac{N}{2}\frac{\Sigma_m}{\mu_0}$:

$$\begin{aligned} i\mathcal{M}^{(1)} &= -i\frac{4\pi b\Sigma_m\mu_0}{k}\partial_{\Sigma_m^2}\int\frac{d^3\ell_E}{(2\pi)^3}\int_0^1 dx\frac{x}{(\ell_E^2+\Sigma_m^2 x)^2} \\ &= -i\frac{b\Sigma_m\mu_0}{k}\partial_{\Sigma_m^2}\frac{1}{(\Sigma_m^2)^{1/2}}\int_0^1 dx\ x^{1/2} \\ &= i\frac{2}{3}\frac{b\mu_0^2}{N|\Sigma_m|^3}. \end{aligned} \tag{A.19}$$

Finally, we calculate $\Sigma_0(q_{\rm FS})$ and $\Sigma_0'(q_{\rm FS})$, which we obtain from evaluating the derivative with respect to q_0 of $\Sigma_0(P)$ at the Fermi surface:

$$i\Sigma_0(P) = \frac{1}{2}\int\frac{d^3p}{(2\pi)^3}\mathrm{tr}\Big[\gamma^0\gamma^\alpha G^{(0)}(p+P)\gamma^\beta\Big(\frac{2\pi}{k}\frac{\epsilon_{\alpha\beta\sigma}p^\sigma}{p^2}\Big)\Big], \tag{A.20}$$

where $P = (q_0+\mu_0, \mu_0\hat{n})$. First, we note that

$$\begin{aligned} \mathrm{tr}[\gamma^0\gamma^\alpha\gamma^\sigma\gamma^\beta](p+P)_\sigma p^\tau\epsilon_{\alpha\beta\tau} &= 2(\eta^{0\alpha}\eta^{\sigma\beta}-\eta^{0\sigma}\eta^{\alpha\beta}+\eta^{0\beta}\eta^{\alpha\sigma})(p+P)_\sigma p^\tau\epsilon_{\alpha\beta\tau} \\ &= (p+P)^\beta p^\tau\epsilon_{0\beta\tau}+(p+P)^\alpha p^\tau\epsilon_{\alpha 0\tau} \\ &= 0. \end{aligned} \tag{A.21}$$

Therefore, only the term proportional to Σ_m in the numerator of $G^{(0)}$ contributes. Using the trace identities in Eq. (A.4), we find

$$i\Sigma_0(P) = \frac{4\pi\Sigma_m}{k}\int\frac{d^3p}{(2\pi)^3}\frac{p^0}{((p+P)^2-\Sigma_m^2)p^2}. \tag{A.22}$$

As above, we combine denominators, shift the integration variable $\ell_\alpha = p_\alpha + P_\alpha x$, and drop any linear in ℓ terms in the numerator:

$$i\Sigma_0(P) = -\frac{4\pi\Sigma_m(q_0+\mu_0)}{k}\int\frac{d^3\ell}{(2\pi)^3}\int_0^1 dx\frac{x}{(\ell^2+x(1-x)P^2-\Sigma_m^2 x)^2}. \tag{A.23}$$

We assume $\Sigma_m^2 > |P^2| \approx |2\mu q_0|$. Wick rotating $\ell_0 \mapsto i(\ell_E)_3$ and sequentially performing the ℓ_E and x integrals, we find:

$$\begin{aligned} i\Sigma_0(P) &= -i\frac{4\pi\Sigma_m(q_0+\mu_0)}{k}\int\frac{d^3\ell_E}{(2\pi)^3}\int_0^1 dx\frac{x}{(\ell_E^2 - x(1-x)(2\mu_0 q_0) + \Sigma_m^2 x)^2} \\ &= -i\frac{\Sigma_m(q_0+\mu_0)}{2k}\int_0^1 dx\frac{x}{(\Sigma_m^2 x - 2\mu_0 q_0 x(1-x))^{1/2}} \\ &= -i\frac{\Sigma_m(q_0+\mu_0)}{3k|\Sigma_m|}\Big(1+\frac{2\mu_0 q_0}{5|\Sigma_m|^2}+\mathcal{O}(q_0^2)\Big) \\ &= -i\frac{2\mu_0}{3N|\Sigma_m|}(q_0+\mu_0)\Big(1+\frac{2\mu_0 q_0}{5|\Sigma_m|^2}+\mathcal{O}(q_0^2)\Big). \end{aligned} \tag{A.24}$$

Taking the derivative of $\Sigma_0(P)$ with respect to q_0, evaluating at $q = (0, \mu_0\hat{n})$, and retaining only the first term ($\mu_0 q_0 \ll |\Sigma_m|^2$), we obtain

$$i\Sigma_0'(q_{\rm FS}) = -i\frac{2\mu_0}{3N|\Sigma_m|}. \tag{A.25}$$

Appendix B

Chapter 3: Appendix

B.1 Derivation of Eq. 3.3

Proof. Polyak's heavy ball method with unit momentum for the minimization of a single objective $f(x)$ is given by,

$$x_{k+1} = x_k + (x_k - x_{k-1}) - \eta \nabla_x f(x_k), \tag{B.1}$$

where η is the learning rate. One can rewrite this equation as,

$$\frac{(x_{k+\delta} - x_k) - (x_k - x_{k-\delta})}{\delta^2} = -\frac{\eta}{\delta^2} \nabla_x f(x_k), \tag{B.2}$$

where δ is the discretization step-size. In the limit $\delta, \eta \to 0$, Eq.(B.2) then becomes $(x_k \to X(t) \equiv X)$,

$$m\ddot{X} = -\nabla_X f(X) \tag{B.3}$$

This is equivalent to Newton's 2$^{\text{nd}}$ Law of motion (Eq.(3.2)) of a particle of mass $m = \delta^2/\eta$, if we identify $F = -\nabla_X f(X)$. ■

B.2 Proof of Proposition 3.2.1

Proof. The EOMs of the quadratic game in continuous-time (Eq.(3.9)), can be discretized in using a combination of implicit and explicit update steps as [331],

$$x_{k+1} - x_k = \delta v_{k+1}^x, \tag{B.4a}$$

$$y_{k+1} - y_k = \delta v_{k+1}^y, \tag{B.4b}$$

$$v_{k+1}^x - v_k^x = -q\delta \nabla_{xy} f\left(x_k, y_k\right) v_k^y - \mu\delta v_k^x - \delta \nabla_x f\left(x_k, y_k\right) \tag{B.4c}$$

$$v_{k+1}^y - v_k^y = q\delta \nabla_{xy} f\left(x_k, y_k\right) v_k^x - \mu\delta v_k^y + \delta \nabla_y f\left(x_k, y_k\right) \tag{B.4d}$$

where δ is the discretization step-size. Using Eqns.(B.4a) and (B.4b), we can further re-express Eqns. (B.4c), (B.4d) as,

$$\begin{aligned} x_{k+1} &= x_k + \beta \Delta x_k - \eta \nabla_x f\left(x_k, y_k\right) - \alpha \nabla_{x,y} f\left(x_k, y_k\right) \Delta y_k \\ y_{k+1} &= y_k + \beta \Delta y_k + \eta \nabla_y f\left(x_k, y_k\right) + \alpha \nabla_{x,y} f\left(x_k, y_k\right) \Delta x_k \end{aligned} \tag{B.5}$$

where $\Delta x_k = x_k - x_{k-1}$, and,

$$\beta = 1 - \mu\delta,\ \eta = \delta^2,\ \alpha = 2q\delta \tag{B.6}$$

■

B.3 Continuous-time Convergence Analysis: Quadratic Min-Max Game

Proof. For the class of quadratic min-max games,

$$f\left(\boldsymbol{X}, \boldsymbol{Y}\right) = \frac{h}{2}|\boldsymbol{X}|^2 - \frac{h}{2}|\boldsymbol{Y}|^2 + \boldsymbol{X}^T \mathbb{A} \boldsymbol{Y} \tag{B.7}$$

where $\boldsymbol{X} \equiv \left(X^1, \cdots, X^n\right), \boldsymbol{Y} \equiv \left(Y^1, \cdots, Y^n\right) \in \mathbb{R}^n$ and $\mathbb{A}_{n\times n}$ is a constant positive-definite matrix, the continuous-time EOMs of Eq.(3.9) become:

$$\begin{aligned} \ddot{\boldsymbol{X}} &= -\mu\dot{\boldsymbol{X}} - h\boldsymbol{X} - \mathbb{A}\boldsymbol{Y} - q\mathbb{A}\dot{\boldsymbol{Y}} \\ \ddot{\boldsymbol{Y}} &= -\mu\dot{\boldsymbol{Y}} - h\boldsymbol{Y} + \mathbb{A}^T\boldsymbol{X} + q\mathbb{A}^T\dot{\boldsymbol{X}} \end{aligned} \tag{B.8}$$

We next define our continuous-time Lyapunov function in this case to be,

$$\begin{aligned} \mathcal{E}_t = &\frac{1}{2}\left(\dot{\boldsymbol{X}} + \mu\boldsymbol{X} + \mu\mathbb{A}\boldsymbol{Y}\right)^T\left(\dot{\boldsymbol{X}} + \mu\boldsymbol{X} + \mu\mathbb{A}\boldsymbol{Y}\right) \\ &+ \frac{1}{2}\left(\dot{\boldsymbol{Y}} + \mu\boldsymbol{Y} - \mu\mathbb{A}^T\boldsymbol{X}\right)^T\left(\dot{\boldsymbol{X}} + \mu\boldsymbol{Y} - \mu\mathbb{A}^T\boldsymbol{X}\right) \\ &+ \frac{1}{2}\left(\dot{\boldsymbol{X}}^T\dot{\boldsymbol{X}} + \dot{\boldsymbol{Y}}^T\dot{\boldsymbol{Y}}\right) + \boldsymbol{X}^T(h + \mathbb{A}\mathbb{A}^T)\boldsymbol{X} + \boldsymbol{Y}^T(h + \mathbb{A}^T\mathbb{A})\boldsymbol{Y} \\ \geq &\, 0 \;\forall\; t \end{aligned} \tag{B.9}$$

The time-derivative of $\mathcal{E}_t$ is then given by,

$$\begin{aligned} \dot{\mathcal{E}}_t = &\left(\dot{\boldsymbol{X}} + \mu\boldsymbol{X} + \mu\mathbb{A}\boldsymbol{Y}\right)^T\left(\ddot{\boldsymbol{X}} + \mu\dot{\boldsymbol{X}} + \mu\mathbb{A}\dot{\boldsymbol{Y}}\right) + \left(\dot{\boldsymbol{Y}} + \mu\boldsymbol{Y} - \mu\mathbb{A}^T\boldsymbol{X}\right)^T\left(\ddot{\boldsymbol{Y}} + \mu\dot{\boldsymbol{Y}} - \mu\mathbb{A}^T\dot{\boldsymbol{X}}\right) \\ &+ \left(\dot{\boldsymbol{X}}^T\ddot{\boldsymbol{X}} + \dot{\boldsymbol{Y}}^T\ddot{\boldsymbol{Y}}\right) + 2\left(\boldsymbol{X}^T(h + \mathbb{A}\mathbb{A}^T)\dot{\boldsymbol{X}} + \boldsymbol{Y}^T(h + \mathbb{A}^T\mathbb{A})\dot{\boldsymbol{Y}}\right) \\ = &\left(\dot{\boldsymbol{X}}^T + \mu\boldsymbol{X}^T + \mu\boldsymbol{Y}^T\mathbb{A}^T\right)\left((-q + \mu)\,\mathbb{A}\dot{\boldsymbol{Y}} - \mathbb{A}\boldsymbol{Y}\right) + \dot{\boldsymbol{X}}^T\left(-q\mathbb{A}\dot{\boldsymbol{Y}} - \mu\dot{\boldsymbol{X}} - \mathbb{A}\boldsymbol{Y}\right) \\ &+ \left(\dot{\boldsymbol{Y}}^T + \mu\boldsymbol{Y}^T - \mu\boldsymbol{X}^T\mathbb{A}\right)\left((q - \mu)\,\mathbb{A}^T\dot{\boldsymbol{X}} + \mathbb{A}^T\boldsymbol{X}\right) + \dot{\boldsymbol{Y}}^T\left(q\mathbb{A}^T\dot{\boldsymbol{X}} - \mu\dot{\boldsymbol{Y}} + \mathbb{A}^T\boldsymbol{X}\right) \\ &+ 2\left(\boldsymbol{X}^T(h + \mathbb{A}\mathbb{A}^T)\dot{\boldsymbol{X}} + \boldsymbol{Y}^T(h + \mathbb{A}^T\mathbb{A})\dot{\boldsymbol{Y}}\right) \\ = &\left(\mu\,(q - \mu) - 2\right)\left(\boldsymbol{Y}^T\mathbb{A}^T\dot{\boldsymbol{X}} - \boldsymbol{X}^T\mathbb{A}\dot{\boldsymbol{Y}}\right) - \left(\mu\,(q - \mu) - 2\right)\left(\boldsymbol{X}^T\mathbb{A}\mathbb{A}^T\dot{\boldsymbol{X}} + \boldsymbol{Y}^T\mathbb{A}^T\mathbb{A}\dot{\boldsymbol{Y}}\right) \\ &- \mu\left(\boldsymbol{X}^T(h + \mathbb{A}\mathbb{A}^T)\boldsymbol{X} + \boldsymbol{Y}^T(h + \mathbb{A}^T\mathbb{A})\boldsymbol{Y}\right) - \mu\left(\dot{\boldsymbol{X}}^T\dot{\boldsymbol{X}} + \dot{\boldsymbol{Y}}^T\dot{\boldsymbol{Y}}\right) \end{aligned} \tag{B.10}$$

where we have used the fact that $\boldsymbol{X}^T\mathbb{A}\boldsymbol{Y}$ being a scalar thus implying $\boldsymbol{X}^T\mathbb{A}\boldsymbol{Y} = \boldsymbol{Y}^T\mathbb{A}^T\boldsymbol{X}$. If we now set $q = (2/\mu) + \mu$ in the above, then that further leads to,

$$\begin{aligned}\dot{\mathcal{E}}_t &= -\mu\left(\boldsymbol{X}^T(h + \mathbb{A}\mathbb{A}^T)\boldsymbol{X} + \boldsymbol{Y}^T(h + \mathbb{A}^T\mathbb{A})\boldsymbol{Y}\right) - \mu\left(\dot{\boldsymbol{X}}^T\dot{\boldsymbol{X}} + \dot{\boldsymbol{Y}}^T\dot{\boldsymbol{Y}}\right) \\ &= -\mu\left(h||\boldsymbol{X}||^2 + h||\boldsymbol{Y}||^2 + \left|\left|\mathbb{A}^T\boldsymbol{X}\right|\right|^2 + ||\mathbb{A}\boldsymbol{Y}||^2\right) - \mu\left(\left|\left|\dot{\boldsymbol{X}}\right|\right|^2 + \left|\left|\dot{\boldsymbol{Y}}\right|\right|^2\right) \leq 0 \;\forall\; t\end{aligned} \tag{B.11}$$

exhibiting that the Lyapunov function, Eq.(3.15) is *asymptotically stable* at all times t.

Next, consider the following expression,

$$\begin{aligned}&-\rho\mathcal{E}_t - \frac{\rho\mu}{2}\left|\left|\boldsymbol{X} - \dot{\boldsymbol{X}}\right|\right|^2 - \frac{\rho\mu}{2}\left|\left|\boldsymbol{Y} - \dot{\boldsymbol{Y}}\right|\right|^2 - \frac{\rho\mu}{2}\left|\left|\dot{\boldsymbol{X}} - \mathbb{A}\boldsymbol{Y}\right|\right|^2 - \frac{\rho\mu}{2}\left|\left|\mathbb{A}^T\boldsymbol{X} + \dot{\boldsymbol{Y}}\right|\right|^2 \\ &\quad= -\rho\mathcal{E}_t - \frac{\rho\mu}{2}\left(||\boldsymbol{X}||^2 + ||\boldsymbol{Y}||^2\right) + \rho\mu\left(\boldsymbol{X}^T\dot{\boldsymbol{X}} + \boldsymbol{Y}^T\dot{\boldsymbol{Y}}\right) - \rho\mu\left(\left|\left|\dot{\boldsymbol{X}}\right|\right|^2 + ||\boldsymbol{Y}||^2\right) \\ &\qquad- \rho\mu\left(\boldsymbol{X}^T\mathbb{A}\dot{\boldsymbol{Y}} - \dot{\boldsymbol{X}}^T\mathbb{A}\boldsymbol{Y}\right) - \frac{\rho\mu}{2}\left(\left|\left|\mathbb{A}^T\boldsymbol{X}\right|\right|^2 + ||\mathbb{A}\boldsymbol{Y}||^2\right) \\ &\quad= -\rho\left(1 + \mu\right)\left(\left|\left|\dot{\boldsymbol{X}}\right|\right|^2 + \left|\left|\dot{\boldsymbol{Y}}\right|\right|^2\right) - \frac{\rho}{2}\left(\mu^2 + \mu + 2h\right)\left(||\boldsymbol{X}||^2 + ||\boldsymbol{Y}||^2\right) \\ &\qquad- \frac{\rho}{2}\left(\mu^2 + \mu + 2\right)\left(\left|\left|\mathbb{A}^T\boldsymbol{X}\right|\right|^2 + ||\mathbb{A}\boldsymbol{Y}||^2\right) \\ &\quad\leq -\rho\mathcal{E}_t\end{aligned} \tag{B.12}$$

where ρ is some positive definite constant. This implies that the above expression is negative semi-definite by construction given $\mu \geq 0$. Now, for a general square matrix $\mathbb{A}$, we can perform a singular value decomposition (SVD) as $\mathbb{A} = \mathbb{V}^{\mathrm{T}}\mathbb{S}\mathbb{U}$. Here, $\mathbb{U}$ and $\mathbb{V}$ are the right and left unitaries of $\mathbb{A}$, while $\mathbb{S}$ is a diagonal matrix of singular values (σ_i) of $\mathbb{A}$. Using this

decomposition in Eq.(B.12), then allows us to write,

$$
\begin{aligned}
&-\rho\left(1+\mu\right)\left(\left|\left|\dot{\boldsymbol{X}}\right|\right|^2+\left|\left|\dot{\boldsymbol{Y}}\right|\right|^2\right)-\frac{\rho}{2}\left(\mu^2+\mu+2h\right)\left(||\boldsymbol{X}||^2+||\boldsymbol{Y}||^2\right)\\
&\qquad-\frac{\rho}{2}\left(\mu^2+\mu+2\right)\left(\left|\left|\mathbb{A}^T\boldsymbol{X}\right|\right|^2+||\mathbb{A}\boldsymbol{y}||^2\right)\\
&=-\rho\left(1+\mu\right)\left(\left|\left|\mathbb{V}\dot{\boldsymbol{X}}\right|\right|^2+\left|\left|\mathbb{U}\dot{\boldsymbol{Y}}\right|\right|^2\right)-\frac{\rho}{2}\left(\mu^2+\mu+2h\right)\left(||\mathbb{V}\boldsymbol{X}||^2+||\mathbb{U}\boldsymbol{Y}||^2\right)\\
&\qquad-\frac{\rho}{2}\left(\mu^2+\mu+2\right)\left(||\mathbb{S}\mathbb{V}\boldsymbol{X}||^2+||\mathbb{S}\mathbb{U}\boldsymbol{Y}||^2\right)\\
&=-\rho\left(1+\mu\right)\left(\left|\left|\dot{\boldsymbol{\mathcal{X}}}\right|\right|^2+\left|\left|\dot{\boldsymbol{\mathcal{Y}}}\right|\right|^2\right)-\frac{\rho}{2}\left(\mu^2+\mu+2h\right)\left(||\boldsymbol{\mathcal{X}}||^2+||\boldsymbol{\mathcal{Y}}||^2\right)\\
&\qquad-\frac{\rho}{2}\left(\mu^2+\mu+2\right)\left(||\mathbb{S}\boldsymbol{\mathcal{X}}||^2+||\mathbb{S}\boldsymbol{\mathcal{Y}}||^2\right)\\
&=-\sum_{j=1}^{n}\rho\left(1+\mu\right)\left(\left|\left|\dot{\mathcal{X}}^j\right|\right|^2+\left|\left|\dot{\mathcal{Y}}^j\right|\right|^2\right)\\
&\qquad-\sum_{j=1}^{n}\frac{\rho}{2}\left(\left(1+\sigma_j^2+2h\right)\left(\mu^2+\mu\right)+2\sigma_j^2\right)\left(\left|\left|\mathcal{X}^j\right|\right|^2+\left|\left|\mathcal{Y}^j\right|\right|^2\right)
\end{aligned}
\tag{B.13}
$$

where we have made use of the relations $\mathbb{U}^{\mathrm{T}}\mathbb{U}=\mathbb{U}\mathbb{U}^{\mathrm{T}}=\mathbb{I}_n=\mathbb{V}^{\mathrm{T}}\mathbb{V}=\mathbb{V}\mathbb{V}^{\mathrm{T}}$, and additionally performed a basis change, as $\boldsymbol{\mathcal{X}}=\mathbb{V}\boldsymbol{X}$ and $\boldsymbol{\mathcal{Y}}=\mathbb{U}\boldsymbol{Y}$. Now, we know from Eq.(B.11) that,

$$
\begin{aligned}
\dot{\mathcal{E}}_t&=-\mu\left(h||\boldsymbol{X}||^2+h||\boldsymbol{Y}||^2+\left|\left|\mathbb{A}^T\boldsymbol{X}\right|\right|^2+||\mathbb{A}\boldsymbol{Y}||^2\right)-\mu\left(\left|\left|\dot{\boldsymbol{X}}\right|\right|^2+\left|\left|\dot{\boldsymbol{Y}}\right|\right|^2\right)\\
&=-\mu\left(h||\boldsymbol{X}||^2+h||\boldsymbol{Y}||^2+\left|\left|\mathbb{U}^{\mathrm{T}}\mathbb{S}\mathbb{V}\boldsymbol{X}\right|\right|^2+\left|\left|\mathbb{V}^{\mathrm{T}}\mathbb{S}\mathbb{U}\boldsymbol{Y}\right|\right|^2\right)-\mu\left(\left|\left|\mathbb{V}\dot{\boldsymbol{X}}\right|\right|^2+\left|\left|\mathbb{U}\dot{\boldsymbol{Y}}\right|\right|^2\right)\\
&=-\mu\left(h||\boldsymbol{\mathcal{X}}||^2+h||\boldsymbol{\mathcal{Y}}||^2+||\mathbb{S}\boldsymbol{\mathcal{X}}||^2+||\mathbb{S}\boldsymbol{\mathcal{Y}}||^2\right)-\mu\left(\left|\left|\dot{\boldsymbol{\mathcal{X}}}\right|\right|^2+\left|\left|\dot{\boldsymbol{\mathcal{Y}}}\right|\right|^2\right)\\
&=-\sum_{j=1}^{n}\mu\left(\sigma_j^2+h\right)\left(\left|\left|\mathcal{X}^j\right|\right|^2+\left|\left|\mathcal{Y}^j\right|\right|^2\right)-\sum_{j=1}^{n}\mu\left(\left|\left|\dot{\mathcal{X}}^j\right|\right|^2+\left|\left|\dot{\mathcal{Y}}^j\right|\right|^2\right)
\end{aligned}
\tag{B.14}
$$

Comparing the above expression with Eq.(B.13), we note that a choice of ρ as,

$$
\rho\leq\min\left\{\frac{\mu}{1+\mu},\ \frac{2\mu(\sigma_{\min}^2+h)}{\left(1+\sigma_{\min}^2+2h\right)\left(\mu^2+\mu\right)+2\sigma_{\min}^2}\right\}\ \forall\ j\in[1,n] \tag{B.15}
$$

implies,

$$
\begin{aligned}
\dot{\mathcal{E}}_t &\leq -\rho\mathcal{E} \\
\Rightarrow \mathcal{E}_t &\leq \mathcal{E}_0 \exp\left(-\rho t\right) \\
\Rightarrow X^T\left(h + \mathbb{A}\mathbb{A}^T\right)X + Y^T\left(h + \mathbb{A}^T\mathbb{A}\right)Y &\leq \mathcal{E}_0 \exp\left(-\rho t\right) \\
\Rightarrow \mathcal{X}^T(h + \mathbb{S}^2)\mathcal{X} + \mathcal{Y}^T(h + \mathbb{S}^2)\mathcal{Y} &\leq \mathcal{E}_0 \exp\left(-\rho t\right) \\
\Rightarrow \sum_{j=1}^{n}(h + \sigma_j^2)\left(||\mathcal{X}^j||^2 + ||\mathcal{Y}^j||^2\right) &\leq \mathcal{E}_0 \exp\left(-\rho t\right) \\
\Rightarrow \sum_{j=1}^{n}(h + \sigma_j^2)\left(||X^j||^2 + ||Y^j||^2\right) &\leq \mathcal{E}_0 \exp\left(-\rho t\right) \\
\therefore\ ||\boldsymbol{X}||^2 + ||\boldsymbol{Y}||^2 &\leq \frac{\mathcal{E}_0}{h + \sigma_{\min}^2} \exp\left(-\rho t\right)\ \forall\ j
\end{aligned}
\tag{B.16}
$$

■

B.4 Proof of Theorem 2

Theorem: The eigenvalues of $\nabla F_{\text{LEAD}}(\boldsymbol{\omega}^*)$ about the Nash equilibrium $\boldsymbol{\omega}^* = (x^*, y^*)$ of the quadratic min-max game are,

$$
\mu_{\pm}(\alpha, \beta, \eta) = \frac{1 - (\eta + \alpha)\lambda + \beta - \eta h \pm \sqrt{\Delta}}{2} \tag{B.17}
$$

where, $\Delta = \left(1 - (\eta + \alpha)\,\lambda + \beta - \eta h\right)^2 - 4\left(\beta - \alpha\lambda\right)$ and $\lambda \in \text{Sp}(\text{off-diag}[\nabla \boldsymbol{v}(\boldsymbol{\omega}^*)])$. Furthermore, for $h, \eta, |\alpha|, |\beta| << 1$, we have,

$$
\begin{aligned}
\mu_{+}^{(i)}(\alpha, \beta, \eta) &\approx 1 - \eta h + \frac{(\eta + \alpha)^2\,\lambda_i^2 + \eta^2 h^2 + \beta^2 - 2\eta h\beta}{4} \\
&\quad + \lambda_i\left(\frac{\eta + \alpha}{2}\left(\eta h - \beta\right) - \eta\right) \\
\mu_{-}^{(i)}(\alpha, \beta, \eta) &\approx \beta - \frac{(\eta + \alpha)^2\,\lambda_i^2 + \eta^2 h^2 + \beta^2 - 2\eta h\beta}{4} \\
&\quad + \lambda_i\left(\frac{\eta + \alpha}{2}\left(\beta - \eta h\right) - \alpha\right)
\end{aligned}
\tag{B.18}
$$

Proof. For the quadratic game (B.7), the Jacobian of the vector field $\boldsymbol{v}$ is given by,

$$\nabla \boldsymbol{v} \equiv \nabla \begin{bmatrix} \nabla_x f(\boldsymbol{x}_t, \boldsymbol{y}_t) \\ -\nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t) \end{bmatrix} = \begin{bmatrix} h\mathbb{I}_{2n} & \mathbb{A} \\ -\mathbb{A}^\top & h\mathbb{I}_{2n} \end{bmatrix} \in \mathbb{R}^{2n} \times \mathbb{R}^{2n}. \tag{B.19}$$

Let us next define a matrix $\mathbb{D}_q$ as,

$$\mathbb{D}_q = \begin{bmatrix} \nabla^2_{xy} f(\boldsymbol{x}, \boldsymbol{y}) & 0 \\ 0 & -\nabla^2_{xy} f(\boldsymbol{x}, \boldsymbol{y}) \end{bmatrix} = \begin{bmatrix} \mathbb{A} & 0 \\ 0 & -\mathbb{A}^\top \end{bmatrix} \in \mathbb{R}^{2n} \times \mathbb{R}^{2n} \tag{B.20}$$

Consequently, the update rule for LEAD can be written as:

$$\begin{aligned} \begin{bmatrix} \boldsymbol{x}_{t+1} \\ \boldsymbol{y}_{t+1} \end{bmatrix} &= \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{y}_t \end{bmatrix} + \beta \begin{bmatrix} \boldsymbol{x}_t - \boldsymbol{x}_{t-1} \\ \boldsymbol{y}_t - \boldsymbol{y}_{t-1} \end{bmatrix} - \eta \begin{bmatrix} \nabla_x f(\boldsymbol{x}_t, \boldsymbol{y}_t) \\ -\nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t) \end{bmatrix} - \alpha \begin{bmatrix} \nabla^2_{xy} f(\boldsymbol{x}_t, \boldsymbol{y}_t) \Delta \boldsymbol{y}_t \\ -\nabla^2_{xy} f(\boldsymbol{x}_t, \boldsymbol{y}_t) \Delta \boldsymbol{x}_t \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{y}_t \end{bmatrix} + \beta \begin{bmatrix} \boldsymbol{x}_t - \boldsymbol{x}_{t-1} \\ \boldsymbol{y}_t - \boldsymbol{y}_{t-1} \end{bmatrix} - \eta \boldsymbol{v} - \alpha \mathbb{D}_{\mathrm{q}} \begin{bmatrix} \Delta \boldsymbol{y}_t \\ \Delta \boldsymbol{x}_t \end{bmatrix} \end{aligned} \tag{B.21}$$

where $\Delta \boldsymbol{y}_t = \boldsymbol{y}_t - \boldsymbol{y}_{t-1}$ and $\Delta \boldsymbol{x}_t = \boldsymbol{x}_t - \boldsymbol{x}_{t-1}$.

Next, by making use of the permutation matrix $\mathbb{P}$,

$$\mathbb{P} := \begin{bmatrix} 0 & \mathbb{I}_n \\ \mathbb{I}_n & 0 \end{bmatrix} \in \mathbb{R}^{2n} \times \mathbb{R}^{2n}$$

we can re-express Eq. (B.21) as,

$$\begin{aligned} \begin{bmatrix} \boldsymbol{\omega}_{t+1} \\ \boldsymbol{\omega}_t \end{bmatrix} &= \begin{bmatrix} \mathbb{I}_{2n} & 0 \\ \mathbb{I}_{2n} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\omega}_t \\ \boldsymbol{\omega}_{t-1} \end{bmatrix} + \beta \begin{bmatrix} \mathbb{I}_{2n} & -\mathbb{I}_{2n} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\omega}_t \\ \boldsymbol{\omega}_{t-1} \end{bmatrix} - \eta \begin{bmatrix} \boldsymbol{v} \\ 0 \end{bmatrix} - \alpha \begin{bmatrix} \mathbb{D}_{\mathrm{q}} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbb{P} & -\mathbb{P} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\omega}_t \\ \boldsymbol{\omega}_{t-1} \end{bmatrix} \\ &= \begin{bmatrix} \mathbb{I}_{2n} & 0 \\ \mathbb{I}_{2n} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\omega}_t \\ \boldsymbol{\omega}_{t-1} \end{bmatrix} + \beta \begin{bmatrix} \mathbb{I}_{2n} & -\mathbb{I}_{2n} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\omega}_t \\ \boldsymbol{\omega}_{t-1} \end{bmatrix} - \eta \begin{bmatrix} \boldsymbol{v} \\ 0 \end{bmatrix} - \alpha \begin{bmatrix} \mathbb{D}_{\mathrm{q}}\mathbb{P} & -\mathbb{D}_{\mathrm{q}}\mathbb{P} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\omega}_t \\ \boldsymbol{\omega}_{t-1} \end{bmatrix} \end{aligned} \tag{B.22}$$

where $\boldsymbol{\omega}_t \equiv (\boldsymbol{x}_t, \boldsymbol{y}_t)$. Hence, the Jacobian of F_{LEAD} is then given by,

$$\begin{aligned}\nabla F_{\text{LEAD}} &= \begin{bmatrix} \mathbb{I}_{2n} & 0 \\ \mathbb{I}_{2n} & 0 \end{bmatrix} + \beta \begin{bmatrix} \mathbb{I}_{2n} & -\mathbb{I}_{2n} \\ 0 & 0 \end{bmatrix} - \eta \begin{bmatrix} \nabla \boldsymbol{v} & 0 \\ 0 & 0 \end{bmatrix} - \alpha \begin{bmatrix} \mathbb{D}_{\text{q}}\mathbb{P} & -\mathbb{D}_{\text{q}}\mathbb{P} \\ 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} (1+\beta)\,\mathbb{I}_{2n} - \eta\nabla\boldsymbol{v} - \alpha\mathbb{D}_{\text{q}}\mathbb{P} & -\beta\mathbb{I}_{2n} + \alpha\mathbb{D}_{\text{q}}\mathbb{P} \\ \mathbb{I}_{2n} & 0 \end{bmatrix}\end{aligned} \tag{B.23}$$

It is to be noted that, for games of the form of Eq. (B.7), we specifically have,

$$\nabla \boldsymbol{v} = \mathbb{D}_{\text{q}}\mathbb{P} + h\mathbb{I}_{2n}$$

and,

$$\text{off-diag}[\nabla \boldsymbol{v}] = \mathbb{D}_{\text{q}}\mathbb{P}$$

Therefore, Eq. (B.23) becomes,

$$\nabla F_{\text{LEAD}} = \begin{bmatrix} (1+\beta-\eta h)\,\mathbb{I}_{2n} - (\eta+\alpha)\,\mathbb{D}_{\text{q}}\mathbb{P} & -\beta\mathbb{I}_{2n} + \alpha\mathbb{D}_{\text{q}}\mathbb{P} \\ \mathbb{I}_{2n} & 0 \end{bmatrix} \tag{B.24}$$

We next proceed to study the eigenvalues of this matrix which will determine the convergence properties of LEAD around the Nash equilibrium. Using Lemma 1 of [116], we can then

write the characteristic polynomial of ∇F_{LEAD} as,

$$\begin{aligned}
&\det\left(X\mathbb{I}_{4n} - \nabla F_{\text{LEAD}}\right) = 0 \\
\Rightarrow &\det\left(\begin{bmatrix} (X-1)\,\mathbb{I}_{2n} - (\beta-\eta h)\,\mathbb{I}_{2n} + (\eta+\alpha)\,\mathbb{D}_{\text{q}}\mathbb{P} & \beta\mathbb{I}_{2n} - \alpha\mathbb{D}_{\text{q}}\mathbb{P} \\ -\mathbb{I}_{2n} & X\mathbb{I}_{2n} \end{bmatrix}\right) = 0 \\
\Rightarrow &\det\left(\left[(X-1)\,(X-\beta)\,\mathbb{I}_{2n} + X\eta h\mathbb{I}_{2n} + (X\eta + X\alpha - \alpha)\,\mathbb{D}_{\text{q}}\mathbb{P}\right]\right) = 0 \\
\Rightarrow &\det\left(\left[((X-1)\,(X-\beta) + X\eta h)\,\mathbf{U}\mathbf{U}^{-1} + (X\eta + X\alpha - \alpha)\,\mathbf{U}\Lambda\mathbf{U}^{-1}\right]\right) = 0 \\
\Rightarrow &\det\left(\left[((X-1)\,(X-\beta) + X\eta h)\,\mathbb{I}_{2n} + (X\eta + X\alpha - \alpha)\,\Lambda\right]\right) = 0 \\
\Rightarrow &\prod_{i=1}^{2n}\left[(X-1)\,(X-\beta) + X\eta h + (X\eta + \alpha\,(X-1))\,\lambda_i\right] = 0
\end{aligned} \tag{B.25}$$

Where, in the above, we have performed an eigenvalue decomposition of $\mathbb{D}_{\text{q}}\mathbb{P} = \mathbf{U}\Lambda\mathbf{U}^{-1}$.

Therefore,

$$\begin{aligned}
&X^2 - X\,(1 - (\eta+\alpha)\lambda_i + \beta - \eta h) + \beta - \alpha\lambda = 0,\ \lambda_i \in \text{Sp}(\mathbb{D}_{\text{q}}\mathbb{P}) \\
&\Rightarrow X^{(i)} \equiv \mu_{\pm}^{(i)} = \frac{1 - (\eta+\alpha)\lambda_i + \beta - \eta h \pm \sqrt{\Delta}}{2}
\end{aligned} \tag{B.26}$$

with,

$$\Delta = (1 - (\eta+\alpha)\,\lambda_i + \beta - \eta h)^2 - 4\,(\beta - \alpha\lambda_i) \tag{B.27}$$

Furthermore for $h, \eta, |\beta|, |\alpha| << 1$, we can approximate the above roots to be,

$$\begin{aligned}
\mu_{+}^{(i)}(\alpha,\beta,\eta) &\approx 1 - \eta h + \frac{(\eta+\alpha)^2\,\lambda_i^2 + \eta^2 h^2 + \beta^2 - 2\eta h\beta}{4} + \lambda_i\left(\frac{\eta+\alpha}{2}\,(\eta h - \beta) - \eta\right) \\
\mu_{-}^{(i)}(\alpha,\beta,\eta) &\approx \beta - \frac{(\eta+\alpha)^2\,\lambda_i^2 + \eta^2 h^2 + \beta^2 - 2\eta h\beta}{4} + \lambda_i\left(\frac{\eta+\alpha}{2}\,(\beta - \eta h) - \alpha\right)
\end{aligned} \tag{B.28}$$

■

B.5 Proof of Proposition 3.3.2

Proposition: For any $\lambda \in \text{Sp}(\text{off-diag}[\nabla \boldsymbol{v}(\omega^*)])$,

$$\nabla_\alpha \rho(\lambda)\big|_{\alpha=0} < 0 \Leftrightarrow \eta \in \left(0, \frac{2}{\text{Im}(\lambda_{\max})}\right), \tag{B.29}$$

where $\text{Im}(\lambda_{\max})$ is the imaginary component of the largest eigenvalue $\lambda_{\max}$.

We observe from Proposition 3.3.2 above that for $h, \eta, |\alpha|, |\beta| << 1$,

$$\begin{aligned} \rho(\alpha, \eta, \beta) &:= \max\{|\mu_+^{(i)}|^2, |\mu_-^{(i)}|^2\} \ \forall \ i \\ &= \max\{\left|\mu_+^{(i)}\right|^2\} \ \forall \ i \end{aligned} \tag{B.30}$$

$$\begin{aligned} \therefore \nabla_\alpha \rho\big|_{\alpha=0} &\approx \max\Big\{ \frac{\eta^2 |\lambda_i|^2 - \eta^2 h^2 - \beta^2}{4} \eta |\lambda_i|^2 + \frac{\eta h \beta - (\eta h - \beta)^2}{2} \eta |\lambda_i|^2 \\ &\quad - (1+\beta)\eta |\lambda_i|^2 \Big\} \ \forall \ i \\ &\approx \max\left\{ \frac{\eta^3}{4} |\lambda_i|^4 - \left(1 + \beta + \frac{3\beta^2}{4}\right) \eta |\lambda_i|^2 \right\} \ \forall \ i \\ &< \max\left\{ \left(\frac{\eta^2}{4} |\lambda_i|^2 - 1\right) \eta |\lambda_i|^2 \right\} \ \forall \ i \end{aligned} \tag{B.31}$$

where we have retained only terms up to cubic-order in $\eta, |\beta|$ and h. Hence, choosing $\eta \in \left(0, \frac{2}{\text{Im}(\lambda_{\max})}\right)$, ensures:

$$\nabla_\alpha \rho\big|_{\alpha=0} < 0 \ \forall \ i, \tag{B.32}$$

We thus posit, that a choice of a positive α causes the norm of the limiting eigenvalue μ_+ of F_{LEAD} to decrease.

B.6 Proof of Theorem 3

Theorem: If we set $\eta = \alpha = \frac{2}{\sigma_{\max}(\mathbb{A})}$, then we have $\forall \ \epsilon > 0$,

$$\Delta_{t+1} \in \mathcal{O}\left(\left(1 - 6\frac{2\sigma_{\min}^2 + h^2}{\sigma_{\max}^2} - 2h\frac{2+\beta}{\sigma_{\max}} + \frac{\beta^2}{2}\right)^t \Delta_0\right) \tag{B.33}$$

where $\sigma_{max}(\sigma_{min})$ is the largest (smallest) singular value of $\mathbb{A}$, $\Delta_{t+1} := ||\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}^*||_2^2 + ||\boldsymbol{\omega}_t - \boldsymbol{\omega}^*||_2^2$.

Proof: From Eq. (B.26), we recall that the eigenvalues of $\nabla F_{\text{LEAD}}(\boldsymbol{\omega}^*)$ for the quadratic game are,

$$\mu_{\pm}^{(i)}(\alpha,\beta,\eta) = \frac{(1-(\alpha+\eta)\lambda_i+\beta-\eta h)}{2}\left(1 \pm \sqrt{1 - \frac{4\left(\beta-\eta\lambda_i\right)}{(1-(\alpha+\eta)\lambda_i+\beta-\eta h)^2}}\right) \tag{B.34}$$

with $\lambda_i \in \text{Sp}(\text{off-diag}[\nabla \boldsymbol{v}(\omega^*)])$. Now, since in the quadratic-game setting considered, we have,

$$\text{off-diag}[\nabla \boldsymbol{v}(\omega^*)] = \mathbb{D}_{\text{q}}\mathbb{P} = \begin{bmatrix} 0 & \mathbb{A} \\ -\mathbb{A}^T & 0 \end{bmatrix} \tag{B.35}$$

hence, $\lambda_i = \pm i\sigma_i$ with σ_i being the singular values of $\mathbb{A}$. This, then allows us to write,

$$\mu_{\pm}^{(i)}(\alpha,\beta,\eta) = \frac{(1-(\alpha+\eta)(\pm i\sigma_i)+\beta)}{2}\left(1 \pm \sqrt{1 - \frac{4\left(\beta-\alpha(\pm i\sigma_i)\right)}{(1-(\alpha+\eta)(\pm i\sigma_i)+\beta)^2}}\right) \tag{B.36}$$

Now, according to Proposition 3.3.2, the convergence behavior of LEAD is determined as, $\Delta_{t+1} \leq \mathcal{O}(\rho+\epsilon)\Delta_t \;\forall\; \epsilon > 0$, where (setting $\eta = \alpha$),

$$\begin{aligned} \rho &:= \max\{|\mu_+^{(i)}|^2, |\mu_-^{(i)}|^2\} \;\forall\; i \\ &= |\mu_+^{(i)}|^2 \;\forall\; i \\ &\approx 1 - 3\eta^2\sigma_{\min}^2 - \frac{\beta^2}{2} - \frac{3}{2}\eta^2 h^2 + (2+\beta)\,\eta h \\ &\equiv 1 - r_{\text{LEAD}} \end{aligned} \tag{B.37}$$

Here, $r_{\text{LEAD}} = 3\eta^2\sigma_{\min}^2 - \frac{\beta^2}{2} - \frac{3}{2}\eta^2 h^2 + (2+\beta)\,\eta h$, is defined to be the *rate of convergence* of LEAD. Furthermore, using the largest learning rate η as prescribed by Proposition 3.3.2, in the above, we find,

$$r_{\text{LEAD}} = 6\frac{2\sigma_{\min}^2 - h^2}{\sigma_{\max}^2} + 2h\frac{2+\beta}{\sigma_{\max}} - \frac{\beta^2}{2} \tag{B.38}$$

Therefore,

$$\begin{aligned}\Delta_{t+1} &\leq \mathcal{O}\left(\left(1 - r_{\text{LEAD}}\right)^t \Delta_0\right) \\ &= \mathcal{O}\left(\left(1 - 6\frac{2\sigma_{\min}^2 + h^2}{\sigma_{\max}^2} - 2h\frac{2+\beta}{\sigma_{\max}} + \frac{\beta^2}{2}\right)^t \Delta_0\right)\end{aligned} \tag{B.39}$$

where $\Delta_{t+1} := ||\boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}^*||_2^2 + ||\boldsymbol{\omega}_t - \boldsymbol{\omega}^*||_2^2$.

B.7 Experiments and Implementation Details

B.7.1 LEAD-Adam Pseudocode

See Page 109.

B.7.2 Simple Experiment On Quadratics

In this Section, we provide an experimental setting of the quadratic min-max game,

$$f(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\boldsymbol{x}^T \mathbb{H} \boldsymbol{x} + \boldsymbol{x}^T \mathbb{A} \boldsymbol{y} - \frac{1}{2}\boldsymbol{y}^T \mathbb{G} \boldsymbol{y}. \tag{B.40}$$

where we set, the matrices $\mathbb{H}, \mathbb{G}$ and $\mathbb{A}$ as,

$$\begin{aligned} \mathbb{H} = \mathbb{G} &:= \mathbb{R}_{\theta=90^\circ} \Lambda \mathbb{R}_{\theta=90^\circ}^T \\ \mathbb{A} &:= R_{\theta_\mathrm{A}} \Lambda \mathbb{R}_{\theta=90^\circ}^T \end{aligned} \tag{B.41}$$

where,

$$\mathbb{R}_\theta = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}, \ \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \tag{B.42}$$

We next vary θ_A from 90° ($\mathbb{A}$ fully aligned with $\mathbb{H}$, $\mathbb{G}$) to 0° ($\mathbb{A}$ fully unaligned with $\mathbb{H}$, $\mathbb{G}$) and compare different methods (Figure B.1). For $\theta_A = 90^\circ$, the matrices are simultaneously diagonalizable, implying decoupled dynamics between the different parameters of players $\boldsymbol{x}$ and $\boldsymbol{y}$. As we decrease θ_A, the dynamics become more coupled. We observe that at 0° (fully unaligned), LEAD outperforms Extra-Grad with momentum (the optimal 1^{st}-order method) and all the other 2^{nd}-order methods. We conjecture that superiority of LEAD is the result of the term $\nabla_{xy} f(x, y)(x_k - x_{k-1})$ for player $\boldsymbol{x}$ and equivalently for player $\boldsymbol{y}$. We

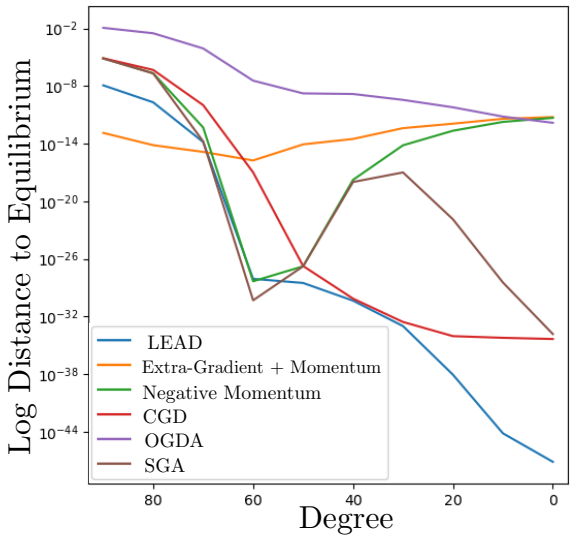

Figure B.1: Comparison of the performance of LEAD vs. several other first-order and second-order methods on a variant of the quadratic min-max game. We start with a game where the matrices are all simultaneously diagonizable and slowly move to the case where they are fully unaligned. We see that LEAD converges faster to the solution of quadratic games whose Jacobian consists of blocks that are not simultaneously diagonizable.

would like to state that for every angle choice, all the methods are fully tuned (a budget of 1000 hyper-parameters was given to each method).

B.7.3 Mixture of Eight Gaussians

Dataset The real data is generated by 8-Gaussian distributions their mean are uniformly distributed around the unit circle and their variance is 0.05. The code to generate the data is included in the source code.

Architecture The architecture for Generator and Discriminator, each consists of four layers of affine transformation, followed by ReLU non-linearity. The weight initialization

is default PyTorch's initialization scheme. See a schematic of the architecture in Table B.1.

Generator	**Discriminator**
Input: $z \in \mathbb{R}^{64} \sim \mathcal{N}(0, I)$	*Input:* $x \in \mathbb{R}^2$
Linear $(64 \to 2000)$	Linear $(2 \to 2000)$
ReLU	ReLU
Linear $(2000 \to 2000)$	Linear $(2000 \to 2000)$
ReLU	ReLU
Linear $(2000 \to 2000)$	Linear $(2000 \to 2000)$
ReLU	ReLU
Linear $(2000 \to 2)$	Linear $(2000 \to 1)$

Table B.1: Architecture used for the Mixture of Eight Gaussians.

Other Details: We use the Adam [183] optimizer on top of our algorithm in the reported results. Furthermore, we use batchsize of 128.

B.7.4 CIFAR 10 DCGAN

Dataset The CIFAR10 dataset is available for download at the following link; `https://www.cs.toronto.edu/~kriz/cifar.html`

Architecture The discriminator has four layers of convolution with LeakyReLU and batch normalization. Also, the generator has four layers of deconvolution with ReLU and batch normalization. See a schematic of the architecture in Table B.2.

Generator	Discriminator
Input: $z \in \mathbb{R}^{100} \sim \mathcal{N}(0, I)$	*Input:* $x \in \mathbb{R}^{3\times32\times32}$
conv. (ker: 4×4, 100 → 1024; stride: 1; pad: 0)	conv. (ker: 4×4, 3 → 256; stride: 2; pad: 1)
Batch Normalization	LeakyReLU
ReLU	conv. (ker: 4×4, 256 → 512; stride: 2; pad: 1)
conv. (ker: 4×4, 1024 → 512; stride: 2; pad: 1)	Batch Normalization
Batch Normalization	LeakyReLU
ReLU	conv. (ker: 4×4, 512 → 1024; stride: 2; pad: 1)
conv. (ker: 4×4, 512 → 256; stride: 2; pad: 1)	Batch Normalization
Batch Normalization	LeakyReLU
ReLU	conv. (ker: 4×4, 1024 → 1; stride: 1; pad: 0)
conv. (ker: 4×4, 256 → 3; stride: 2; pad: 1)	
Tanh	Sigmoid

Table B.2: Architecture used for CIFAR-10 DCGAN.

Other Details For the baseline we use Adam with β_1 set to 0.5 and β_2 set to 0.99. Generator's learning rate is 0.0002 and discriminator's learning rate is 0.0001. The same learning rate and momentum were used to train LEAD model. We also add the mixed derivative term with $\alpha_d = 0.3$ and $\alpha_g = 0.0$.

The baseline is a DCGAN with the standard non-saturating loss (non-zero sum formulation). In our experiments, we compute the FID based on 50,000 samples generated from our model vs 50,000 real samples.

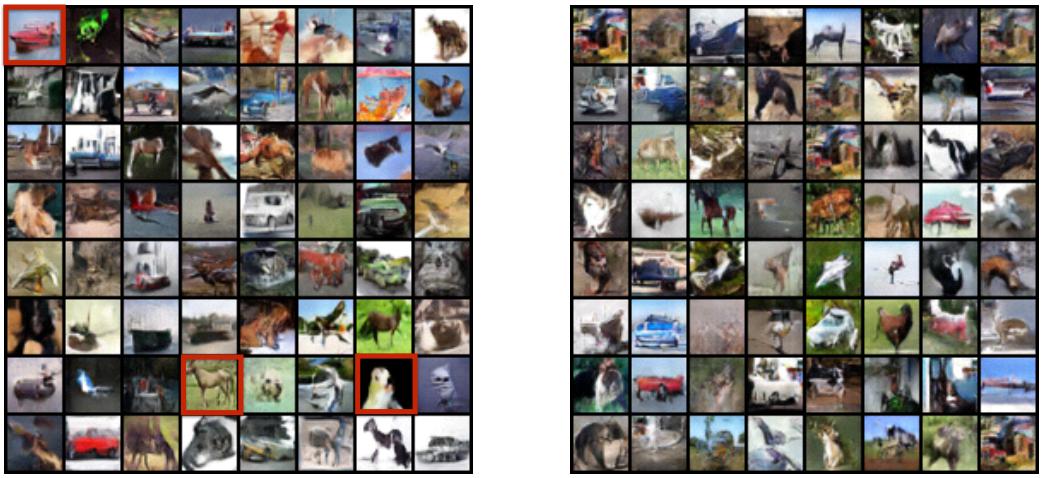

Figure B.2: Performance of LEAD on CIFAR-10 image generation task on a DCGAN architecture. **Left**: LEAD achieves FID 19.27. **Right**: Vanilla Adam achieves FID 24.38. LEAD is able to generate better sample qualities from several classes such as ships, horses and birds (red). Best performance is reported after 100 epochs.

B.7.5 CIFAR 10 ResNet

Dataset The CIFAR10 dataset is available for download at the following link; `https://www.cs.toronto.edu/~kriz/cifar.html`

Architecture See Table B.4 for a schematic of the architecture used for the CIFAR10 experiments with ResNet.

Other Details: The baseline is a ResNet with non-saturating loss (non-zero sum formulation). Similar to [243], for every time that the generator is updated, the discriminator is updated 5 times. For both the Baseline SNGAN and LEAD-Adam we use a β_1 of 0.0 and β_2 of 0.9 for Adam. Baseline SNGAN uses a learning rate of 0.0002 for both the generator and

Gen–Block
Shortcut:
Upsample(×2)
Residual:
Batch Normalization
ReLU
Upsample(×2)
conv. (ker: 3×3, 256 → 256; stride: 1; pad: 1)
Batch Normalization
ReLU
conv. (ker: 3×3, 256 → 256; stride: 1; pad: 1)

Dis–Block
Shortcut:
downsample
conv. (ker: 1×1, $3_{\ell=1}/128_{\ell\neq 1} \rightarrow 128$; stride: 1)
Spectral Normalization
[AvgPool (ker:2×2, stride:2)], if $\ell \neq 1$
Residual:
[ReLU], if $\ell \neq 1$
conv. (ker: 3×3, $3_{\ell=1}/128_{\ell\neq 1} \rightarrow 128$; stride: 1; pad: 1)
Spectral Normalization
ReLU
conv. (ker: 3×3, 128 → 128; stride: 1; pad: 1)
Spectral Normalization
AvgPool (ker:2×2)

Table B.3: ResNet blocks used for the ResNet architectures (see Table B.4).

the discriminator. LEAD-Adam also uses a learning rate of 0.0002 for the generator but 0.0001 for the discriminator. LEAD-Adam uses an α of 0.5 and 0.01 for the generator and the discriminator respectively. Furthermore, we evaluate both the baseline and our method on an exponential moving average of the generator's parameters.

In our experiments, we compute the FID based on 50,000 samples generated from our model vs 50,000 real samples and reported the mean and variance over 5 random runs. We have provided pre-trained models as well as the source code for both LEAD-Adam and Baseline SNGAN in our GitHub repository.

Generator	Discriminator
Input: $z \in \mathbb{R}^{64} \sim \mathcal{N}(0, I)$	*Input:* $x \in \mathbb{R}^{3\times32\times32}$
Linear($64 \to 4096$)	D–ResBlock
G–ResBlock	D–ResBlock
G–ResBlock	D–ResBlock
G–ResBlock	D–ResBlock
Batch Normalization	ReLU
ReLU	AvgPool (ker:8×8)
conv. (ker: 3×3, $256 \to 3$; stride: 1; pad:1)	Linear($128 \to 1$)
$Tanh(\cdot)$	Spectral Normalization

Table B.4: ResNet architectures used for experiments on CIFAR10.

B.8 Comparison to other methods

In this section we compare our method with several other second order methods in the min-max setting.

The distinction of LEAD from SGA and LookAhead, can be understood by considering the 1^{st}-order approximation of $x_{k+1} = x_k - \eta\nabla_x f\left(x_k, y_k + \eta\Delta y_k\right)$, where $\Delta y_k = \eta\nabla_y f\left(x_k + \eta\Delta x, y_k\right)$.

[1]For FtR, we have provided the update for the second player given the first player performs gradient descent on f.

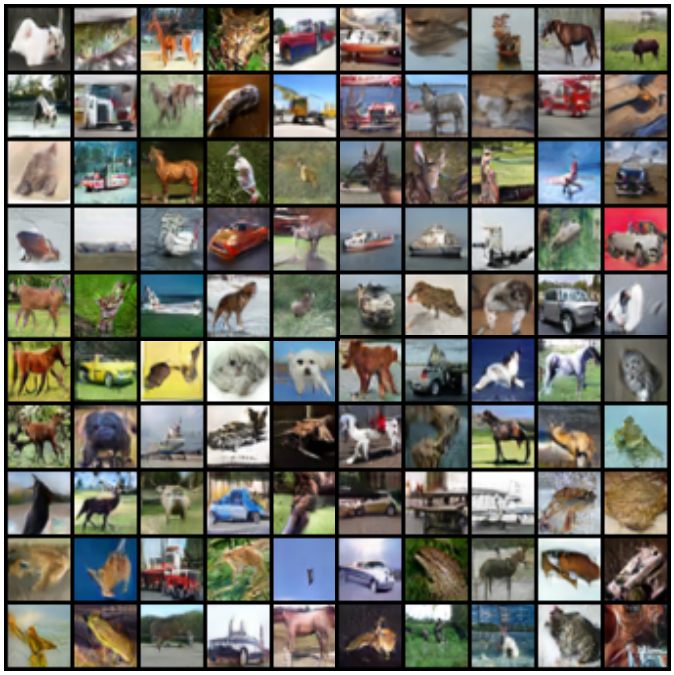

Figure B.3: Generated sample of LEAD-Adam on CIFAR-10 after 50k iterations on a ResNet architecture. We achieve an FID score of 10.49 using learning rate $2e-4$ for the generator and the discriminator, α for the generator is 0.01 and for the discriminator is 0.5.

This gives rise to:

$$x_{k+1} = x_k - \eta \nabla_x f\left(x_k, y_k\right) - \eta^2 \nabla^2_{xy} f\left(x_k, y_k\right) \Delta y, \tag{B.43}$$

$$y_{k+1} = y_k + \eta \nabla_y f\left(x_k, y_k\right) + \eta^2 \nabla^2_{xy} f\left(x_k, y_k\right) \Delta x, \tag{B.44}$$

with $\Delta x, \Delta y$ corresponding to each player accounting for its opponent's potential next step. However, SGA and LookAhead additonally *model* their opponent as *naive* learners i.e. $\Delta x = -\nabla_x f(x_k, y_k)$, $\Delta y = \nabla_y f(x_k, y_k)$. On the contrary, our method does away with such specific assumptions, instead modeling the opponent based on its most recent move.

Furthermore, there is a resemblance between LEAD and OGDA that we would like to address. The 1^{st} order Taylor expansion of the difference in gradients term of OGDA

		Coefficient	Momentum	Gradient	Interaction-xy	Interaction-xx
GDA	$\Delta x_{k+1} =$	1	0	$-\eta\nabla_x f$	$-\eta\nabla_x f$	0
LEAD	$\mathbf{\Delta x_{k+1}} =$	**1**	$\beta\mathbf{\Delta x_k}$	$-\eta\nabla_{\mathbf{x}}\mathbf{f}$	$-\alpha\nabla^{\mathbf{2}}_{\mathbf{xy}}\mathbf{f}\mathbf{\Delta y_k}$	0
SGA[32]	$\Delta x_{k+1} =$	1	0	$-\eta\nabla_x f$	$-\eta\gamma\nabla^2_{xy} f\nabla_y f$	0
CGD[315]	$\Delta x_{k+1} =$	$\mathcal{C}^{-1}$	0	$-\eta\nabla_x f$	$-\eta^2\nabla^2_{xy} f\nabla_y f$	0
CO[233]	$\Delta x_{k+1} =$	1	0	$-\eta\nabla_x f$	$-\eta\gamma\nabla^2_{xy} f\nabla_y f$	$-\eta\gamma\nabla^2_{xx} f\nabla_x f$
FtR[356]	$\Delta y_{k+1} =$	1	0	$\eta_y\nabla_y f$	$\eta_x \left(\nabla^2_{yy} f\right)^{-1}\nabla^2_{yx} f\nabla_x f$	0
LOLA[96]	$\Delta x_{k+1} =$	1	0	$-\eta\nabla_x f$	$-2\eta\alpha\nabla_{xy} f\nabla_y f$	0

Table B.5: Comparison of several second-order methods in min-max optimization. Each update rule, corresponding to a particular row, can be constructed by adding cells in that row from Columns 4 to 7 and then multiplying that by the value in Column 1. Furthermore, $\Delta x_{k+1} = x_{k+1} - x_k$, while $\mathcal{C} = \left(\mathrm{I} + \eta^2\nabla^2_{xy} f\nabla^2_{yx} f\right)$. We compare the update rules of the first player[1] for the following methods: Gradient Descent-Ascent (GDA), Least Action Dynamics (LEAD, ours), Symplectic Gradient Adjustment (SGA), Competitive Gradient Descent (CGD), Consensus Optimization (CO), Follow-the-Ridge (FtR) and Learning with Opponent Learning Awareness (LOLA), in a zero-sum game.

yields the update (for x):

$$x_{k+1} = x_k - \eta\nabla_x f - \eta^2\nabla^2_{xy} f\nabla_y f + \eta^2\nabla^2_{xx} f\nabla_x f, \tag{B.45}$$

which contains an extra 2^{nd} order term $\nabla^2_{xx} f$ compared to ours. As noted in [315], the $\nabla^2_{xx} f$ term does not systematically aid in curbing the min-max rotations, rather causing convergence to non-Nash points in some settings. For e.g., let us consider the simple game $f(x, y) = \gamma(x^2 - y^2)$, where x, y, γ are all scalars, with the Nash equilibrium of this game

located at $(x^* = 0, y^* = 0)$. For a choice of $\gamma \geq 6$, OGDA fails to converge for any learning rate while methods like LEAD, Gradient Descent Ascent (GDA) and CGD ([315]) that do not contain the $\nabla_{xx} f(\nabla_{yy} f)$ term do exhibit convergence. See Figure B.4 and [315] for more discussion.

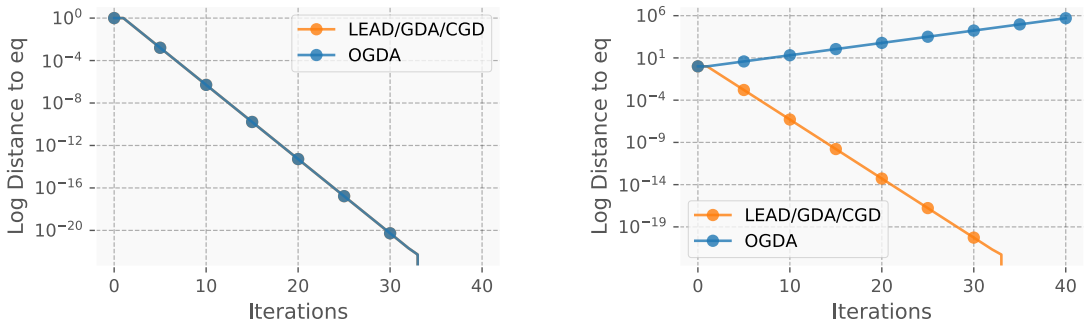

Figure B.4: Figure depicting the convergence/divergence of several algorithms on the game of $f(x, y) = \gamma(x^2 - y^2)$ (Nash equilibrium at $x^* = 0, y^* = 0$). **Left**: For $\gamma = 1$, OGDA and LEAD/GDA/CGD (overlaying) are found to converge to the Nash eq. **Right**: For $\gamma = 6$, we find that OGDA fails to converge while LEAD/GDA/CGD (overlaying) converge. We conjecture that the reason behind this observation is the existence of $\nabla^2_{xx} f$ term in the optimization algorithm of OGDA.

Algorithm 2 Least Action Dynamics Adam (LEAD-Adam)

Input: learning rate η, momentum β, coupling coefficient α.

Initialize: $x_0 \leftarrow x_{init}$, $y_0 \leftarrow y_{init}$, $t \leftarrow 0$, $m_0^x \leftarrow 0$, $v_0^x \leftarrow 0$ $m_0^y \leftarrow 0$, $v_0^y \leftarrow 0$

while not converged **do**

$t \leftarrow t + 1$

$g_x \leftarrow \nabla_x f(x_t, y_t)$

$g_{xy}\Delta y \leftarrow \nabla_y(g_x)(y_t - y_{t-1})$

$g_t^x \leftarrow g_{xy}\Delta y + g_x$

$m_t^x \leftarrow \beta_1.m_{t-1}^x + (1-\beta_1).g_t^x$

$v_t^x \leftarrow \beta_2.v_{t-1}^x + (1-\beta_2).(g_t^x)^2$

$\hat{m}_t \leftarrow m_t/(1-\beta_1^t)$

$\hat{v}_t \leftarrow v_t/(1-\beta_2^t)$

$x_{t+1} \leftarrow x_t - \eta\ \hat{m}_t/(\sqrt{\hat{v}_t} + \epsilon)$

$g_y \leftarrow \nabla_y f(x_{t+1}, y_t)$

$g_{xy}\Delta x \leftarrow \nabla_x(g_y)(x_{t+1} - x_t)$

$g_t^y \leftarrow g_{xy}\Delta x + g_y$

$m_t^y \leftarrow \beta_1.m_{t-1}^y + (1-\beta_1).g_t^y$

$v_t^y \leftarrow \beta_2.v_{t-1}^y + (1-\beta_2).(g_t^y)^2$

$\hat{m}_t^y \leftarrow m_t^y/(1-\beta_1^t)$

$\hat{v}_t^y \leftarrow v_t^y/(1-\beta_2^t)$

$y_{t+1} \leftarrow y_t + \eta\ \hat{m}_t^y/(\sqrt{\hat{v}_t^y} + \epsilon)$

end while

return (x, y)

Appendix C

Chapter 4: Appendix

C.1 Self-averaging and the replica trick

Before proceeding, we note that using the replica trick [237], one can recast Eq. (C.2) as,

$$-\beta f = \frac{1}{d} \lim_{r \to 0} \frac{\langle \langle Z^r \rangle \rangle_{\boldsymbol{x}, \boldsymbol{W}^*} - 1}{r}, \quad r \in \mathbb{Z} \tag{C.1}$$

C.2 Theoretical Details

C.2.1 Generalization Error

The self-averaged free energy (per network weight) is given by [91],

$$-\beta f = \frac{1}{d} \langle \langle \ln Z \rangle \rangle_{\boldsymbol{x}, \boldsymbol{W}^*} \tag{C.2}$$

where $\beta = 1/T$ is the inverse temperature, d the student (and teacher) width, and Z is the partition function of the system defined to be,

$$\begin{aligned} Z &= \int_{-\infty}^{\infty} \mathrm{d}\mu\left(\mathbf{W}\right) e^{-n\beta E_{\mathcal{T}}} \\ \text{with,} \quad \mathrm{d}\mu\left(\mathbf{W}\right) &= \prod_{i=1}^{d} \frac{\mathrm{d}W_i}{\left(\sqrt{2\pi}\right)^d} \delta\left(\gamma_i^2 \sum_{i=1}^{d} \left(W_i\right)^2 - dQ_0\right) \end{aligned} \tag{C.3}$$

Here, the γ_i's correspond to the respective mask values acting on each student neuron while n corresponds to the number of training examples. Additionally, $\langle\langle\cdot\rangle\rangle$ in Eq. (C.2) refers to self-averaging over the (normal Gaussian) input and (uniform) teacher weight distributions. Next, by making use of the replica trick, one can rewrite Eq. (C.2) as,

$$-\beta f = \frac{1}{d} \lim_{r\to 0} \frac{\langle\langle Z^r\rangle\rangle_{\boldsymbol{x},\boldsymbol{W}^*} - 1}{r}, \quad r \in \mathbb{Z} \tag{C.4}$$

To evaluate $\langle\langle Z^r\rangle\rangle_{\boldsymbol{x},\boldsymbol{W}^*}$, we first introduce the teacher and the student outputs $y^{*\mu}$ and $y^{a,\mu}$ (Eq. (4.2),(4.3)) as Gaussian variables (for $d >> 1$) in the partition function, as,

$$\begin{aligned} \langle\langle Z^r\rangle\rangle_{\mathbf{x},\mathbf{W}^*} = \int \prod_{a=1}^{r} \mathrm{d}\mu\left(\mathbf{W}^a\right) \prod_{a,\mu} \mathrm{d}y^{a,\mu} \prod_{\mu} \mathrm{d}y^{*\mu} \prod_{a,\mu} \exp^{-n\beta E_{\mathcal{T}}(y^{*\mu}, y^{a,\mu})} \\ \times \prod_{a,\mu} \left\langle\left\langle \delta\left(y^{*\mu} - \frac{1}{\sqrt{d}} \sum_{i=1}^{d} W_i^* x_i^{\mu}\right) \delta\left(y^{a,\mu} - \frac{1}{\sqrt{d}} \sum_{i=1}^{d} \gamma_i W_i^a x_i^{\mu}\right) \right\rangle\right\rangle_{\mathbf{x},\mathbf{W}^*} \end{aligned} \tag{C.5}$$

with a being the replica index. We next express the above δ functions using their integral representations, to get,

$$\begin{aligned} \langle\langle Z^r\rangle\rangle_{\mathbf{x},\mathbf{W}^*} = \int \prod_{a=1}^{r} \mathrm{d}\mu\left(\mathbf{W}^a\right) \prod_{a,\mu} \mathrm{d}y^{a,\mu} \prod_{\mu} \mathrm{d}y^{*\mu} \prod_{a,\mu} \exp^{-n\beta E_{\mathcal{T}}(y^{*\mu}, y^{a,\mu})} \prod_{a,\mu} \int \frac{\mathrm{d}\hat{y}^{*\mu}}{2\pi} \int \frac{\mathrm{d}\hat{y}^{a,\mu}}{2\pi} \\ \times e^{i(y^{*\mu}\hat{y}^{*\mu} + y^{a,\mu}\hat{y}^{a,\mu})} \left\langle\left\langle e^{-\frac{i}{\sqrt{d}}\left(\left(\sum_i W_i^* x_i^{\mu}\right)\hat{y}^{*\mu} + \left(\sum_i \gamma_i W_i^a x_i^{\mu}\right)\hat{y}^{a,\mu}\right)} \right\rangle\right\rangle_{\mathbf{x},\mathbf{W}^*} \end{aligned} \tag{C.6}$$

Now, as $d \to \infty$, one can write,

$$\begin{aligned}
&\prod_{a,\mu} \left\langle \left\langle e^{-\frac{i}{\sqrt{d}}\left(\left(\sum_i W_i^* x_i^\mu\right)\hat{y}^{*\mu} + \left(\sum_i \gamma_i W_i^a x_i^\mu\right)\hat{y}^{a,\mu}\right)} \right\rangle \right\rangle_{\mathbf{x},\mathbf{W}^*} \\
&\qquad \approx \left\langle \left\langle 1 - \frac{1}{2d} \sum_{i,\mu} \left(W_i^* \hat{y}^{*\mu} + \sum_a \gamma_i W_i^a \hat{y}^{a,\mu} \right)^2 \right\rangle \right\rangle_{\mathbf{W}^*} \\
&\qquad = \left\langle \left\langle e^{\ln\left(1 - \frac{1}{2d}\sum\limits_{i,\mu}\left(W_i^*\hat{y}^{*\mu} + \sum\limits_a \gamma_i W_i^a \hat{y}^{a,\mu}\right)^2\right)} \right\rangle \right\rangle_{\mathbf{W}^*}
\end{aligned} \tag{C.7}$$

where we have made use of the fact that $\langle x_i \rangle = 0$ and $\langle x_i x_j \rangle = \delta_{ij}$. Now, since $\ln(1+x) \approx x$ for $x << 1$, we can furthermore write the above as,

$$\begin{aligned}
&\prod_{a,\mu} \left\langle \left\langle e^{-\frac{i}{\sqrt{d}}\left(\left(\sum_i W_i^* x_i^\mu\right)\hat{y}^{*\mu} + \left(\sum_i \gamma_i W_i^a x_i^\mu\right)\hat{y}^{a,\mu}\right)} \right\rangle \right\rangle_{\mathbf{x},\mathbf{W}^*} \\
&\qquad \approx \left\langle \left\langle e^{-\frac{1}{2d}\sum\limits_{i,\mu}\left(W_i^*\hat{y}^{*\mu} + \sum\limits_a \gamma_i W_i^a \hat{y}^{a,\mu}\right)^2} \right\rangle \right\rangle_{\mathbf{W}^*}
\end{aligned} \tag{C.8}$$

We can additionally expand the argument of the above exponential as,

$$\begin{aligned}
&-\frac{1}{2d} \sum_{i,\mu} \left(W_i^* \hat{y}^{*\mu} + \sum_a \gamma_i W_i^a \hat{y}^{a,\mu} \right)^2 \\
&= -\frac{1}{2d} \sum_{i,\mu} \left(W_i^* \hat{y}^{*\mu} + \sum_a \gamma_i W_i^a \hat{y}^{a,\mu} \right) \left(W_i^* \hat{y}^{*\mu} + \sum_b \gamma_i W_i^b \hat{y}^{b,\mu} \right) \\
&= -\frac{1}{2} \sum_{\mu,a,b} \hat{y}^{a,\mu}\hat{y}^{b,\mu} \cdot \frac{1}{d} \sum_i \gamma_i^2 W_i^a W_i^b - \sum_{\mu,a} \hat{y}^{a,\mu}\hat{y}^{*\mu} \cdot \frac{1}{d} \sum_i \gamma_i^2 W_i^a W_i^* \\
&\quad - \frac{1}{2} \sum_\mu (\hat{y}^{a,\mu})^2 \cdot \frac{1}{d} \sum_i \gamma_i^2 (W_i^*)^2
\end{aligned} \tag{C.9}$$

If next set,

$$\frac{1}{d}\sum_{i=1}^d (W_i^a)^2 \gamma_i^2 \equiv Q_0, \quad \frac{1}{d}\sum_{i=1}^d W_i^a W_i^b \gamma_i^2 \equiv Q^{ab}, \text{ for } a \neq b \tag{C.10}$$

$$\frac{1}{d}\sum_{i=1}^d \gamma_i W_i^a W_i^* = R^a, \quad \frac{1}{d}\sum_{i=1}^d \gamma_i^2 (W_i^*)^2 = 1, \tag{C.11}$$

then we can consequently write,

$$\begin{aligned}\langle\langle Z^r\rangle\rangle_{\mathbf{x},\mathbf{W}^*} = &\int \prod_{a=1}^{r} \mathrm{d}\mu\left(\mathbf{W}^a\right) \prod_{a,\mu} \mathrm{d}y^{a,\mu} \prod_{\mu} \mathrm{d}y^{*\mu} \prod_{a,\mu} \exp^{-n\beta E_{\mathcal{T}}(y^{*\mu},y^{a,\mu})} \\ &\times \prod_{a,\mu} \int \frac{\mathrm{d}\hat{y}^{*\mu}}{2\pi} \int \frac{\mathrm{d}\hat{y}^{a,\mu}}{2\pi} e^{i(y^{*\mu}\hat{y}^{*\mu}+y^{a,\mu}\hat{y}^{a,\mu})} \times \\ &\left\langle\left\langle \exp^{\left(-\frac{1}{2}\sum_{\mu,a}(\hat{y}^{a,\mu})^2 Q_0 - \frac{1}{2}\sum_{\mu,a\neq b}\hat{y}^{a,\mu}\hat{y}^{b,\mu}Q^{ab} - \sum_{\mu,a}\hat{y}^{a,\mu}\hat{y}^{*\mu}R^a - \frac{1}{2}\sum_{\mu}(\hat{y}^{*\mu})^2\right)} \right\rangle\right\rangle_{\mathbf{W}^t}\end{aligned} \tag{C.12}$$

We can now additionally incorporate integrals over dQ^{ab} and dR^a as,

$$\begin{aligned}\langle\langle Z^r\rangle\rangle_{\mathbf{x},\mathbf{W}^*} = &\int \prod_{a=1}^{r} \mathrm{d}\mu\left(\mathbf{W}^a\right) \prod_{a,\mu} \mathrm{d}y^{a,\mu} \prod_{\mu} \mathrm{d}y^{*\mu} \prod_{a,\mu} \exp^{-n\beta E_{\mathcal{T}}(y^{*\mu},y^{a,\mu})} \\ &\times \prod_{a,\mu} \int \frac{\mathrm{d}\hat{y}^{*\mu}}{2\pi} \int \frac{\mathrm{d}\hat{y}^{a,\mu}}{2\pi} e^{i(y^{*\mu}\hat{y}^{*\mu}+y^{a,\mu}\hat{y}^{a,\mu})} \prod_{a<b} \mathrm{d}\left(dQ^{ab}\right) \prod_{a} \mathrm{d}\left(dR^a\right) \\ &\times \left\langle\left\langle \prod_a \delta\left(\sum_i \gamma_i W_i^a W_i^* - dR^a\right)\right\rangle\right\rangle_{\mathbf{W}^*} \\ &\times \left\langle\left\langle \prod_{a<b} \delta\left(\sum_i \gamma_i^2 W^a W^b - dQ^{ab}\right)\right\rangle\right\rangle_{\mathbf{W}^*} \\ &\times e^{-\frac{1}{2}\sum_{\mu,a}(\hat{y}^{a,\mu})^2 Q_0 - \frac{1}{2}\sum_{\mu,a\neq b}\hat{y}^{a,\mu}\hat{y}^{b,\mu}Q^{ab} - \sum_{\mu,a}\hat{y}^{a,\mu}\hat{y}^{*\mu}R^a - \frac{1}{2}\sum_{\mu}(\hat{y}^{*\mu})^2}\end{aligned} \tag{C.13}$$

$$\begin{aligned}\Rightarrow \langle\langle Z^r\rangle\rangle_{\mathbf{x},\mathbf{W}^*} = &\int \prod_{i,a} \frac{\mathrm{d}W_i^a}{\left(\sqrt{2\pi}\right)^d} \delta\left(\sum_{i=1}^{d} \gamma_i^2 \left(W_i^a\right)^2 - dQ_0\right) \prod_{a,\mu} \mathrm{d}y^{a,\mu} \prod_{\mu} \mathrm{d}y^{*\mu} \\ &\times \prod_{a,\mu} \exp^{-n\beta E_{\mathcal{T}}(y^{*\mu},y^{a,\mu})} \prod_{a,\mu} \int \frac{\mathrm{d}\hat{y}^{*\mu}}{2\pi} \int \frac{\mathrm{d}\hat{y}^{a,\mu}}{2\pi} e^{i(y^{*\mu}\hat{y}^{*\mu}+y^{a,\mu}\hat{y}^{a,\mu})} \\ &\times \prod_{a<b} \mathrm{d}\left(dQ^{ab}\right) \prod_{a} \mathrm{d}\left(dR^a\right) \left\langle\left\langle \prod_a \delta\left(\sum_i \gamma_i W_i^a W_i^* - dR^a\right)\right\rangle\right\rangle_{\mathbf{W}^*} \\ &\times \left\langle\left\langle \prod_{a<b} \delta\left(\sum_i \gamma_i^2 W_i^a W_i^b - dQ^{ab}\right)\right\rangle\right\rangle_{\mathbf{W}^*} \\ &\times e^{-\frac{1}{2}\sum_{\mu,a}(\hat{y}^{a,\mu})^2 Q_0 - \frac{1}{2}\sum_{\mu,a\neq b}\hat{y}^{a,\mu}\hat{y}^{b,\mu}Q^{ab} - \sum_{\mu,a}\hat{y}^{a,\mu}\hat{y}^{*\mu}R^a - \frac{1}{2}\sum_{\mu}(\hat{y}^{*\mu})^2}\end{aligned} \tag{C.14}$$

where in the above we have re-expressed $\mathrm{d}\mu\left(\mathbf{W}^a\right)$ using Eq. (C.3). Now, integrating over $\hat{y}^{*\mu}$ and $\mathbf{W}_i^*$, then yields,

$$\begin{aligned}
\langle\langle Z^r\rangle\rangle_{\mathbf{x},\mathbf{W}^*} = & \int \prod_{i,a} \frac{\mathrm{d}W_i^a}{\left(\sqrt{2\pi}\right)^d} \int \prod_a \frac{\mathrm{d}\hat{k}_a}{4\pi} \prod_{a,\mu} \frac{\mathrm{d}y^{a,\mu}\mathrm{d}\hat{y}^{a,\mu}}{2\pi} \prod_\mu \frac{\mathrm{d}y^{*\mu}}{\sqrt{2\pi}} \prod_{a,\mu} \exp^{-n\beta E_{\mathcal{T}}(y^{*\mu},y^{a,\mu})} \\
& \times \prod_{a<b} \frac{\mathrm{d}Q^{ab}\mathrm{d}\hat{Q}^{ab}}{2\pi/d} \prod_a \frac{\mathrm{d}R^a\mathrm{d}\hat{R}^a}{2\pi/d} \prod_{a,\mu} e^{\frac{idQ_0}{2}\sum_a \hat{k}_a + id\sum_a R^a\hat{R}^a + id\sum_{a<b} Q^{ab}\hat{Q}^{ab}} \\
& \times e^{-i\sum_{i,a}\frac{\hat{k}^a}{2}\gamma_i^2\left(W_i^a\right)^2 - i\sum_{i,a<b}\gamma_i^2 W_i^a W_i^b\hat{Q}^{ab} - i\sum_i \gamma_i W_i^a\hat{R}^a} \\
& \times e^{-\frac{1}{2}\sum_\mu (y^{*\mu})^2 - \frac{1}{2}\sum_a\left(Q_0-(R^a)^2\right)\sum_\mu(\hat{y}^{a,\mu})^2 - \frac{1}{2}\sum_{\mu,a\neq b}\hat{y}^{a,\mu}\hat{y}^{b,\mu}\left(Q^{ab}-R^aR^b\right)} \\
& \times e^{i\sum_{\mu,a} y^{a,\mu}\hat{y}^{a,\mu} - i\sum_{\mu,a} y^{*\mu}\hat{y}^{a,\mu}R^a}
\end{aligned} \tag{C.15}$$

C.3 Experimental Details

C.3.1 ResNet-18 on CIFAR-10

We train a ResNet-18 on the CIFAR-10 [189], following the setup of [254] for two different width of $k = 20$ and $k = 60$ which according to [254], are considered moderately and highly overparameterized, respectively. For the experiments a 15% label noise is added.

C.3.2 Decomposition of the Generalization Error

We would like to draw readers' attention to Eq. 11, in which each of R and Q are consisted of two terms which together results in the expression for the generalization error as in Eq. 9. Let us denote these two terms as following,

$$\mathcal{L}_{\mathcal{G}}^{(1)} := \frac{1}{2}(H_1 - 2R_1H_1 + Q1),$$
$$\mathcal{L}_{\mathcal{G}}^{(2)} := \frac{1}{2}(H_2 - 2R_2H_2 + Q2),$$

where,

$$
\begin{aligned}
R_1 &:= \frac{\alpha_1}{a_1 p} H_1 \\
R_2 &:= \frac{\alpha_2}{a_2 (d-p)} H_2, \\
Q_1 &:= \frac{\alpha_1}{a_1^2 - \alpha_1} \left(G_1 - \frac{2 - a_1}{a_1} \alpha_1 H_1 \right), \\
Q_2 &:= \frac{\alpha_2}{a_2^2 - \alpha_2} \left(G_2 - \frac{2 - a_2}{a_2} \alpha_2 H_2 \right),
\end{aligned}
$$

in which, the terms with a super or subscript 1 evolve at a faster time-scale compared to those with 2 as we have assumed that $\gamma_1 \gg \gamma_2$.

In Fig. C.1, we separately plot $\mathcal{L}_{\mathcal{G}}^{(1)}$ and $\mathcal{L}_{\mathcal{G}}^{(2)}$ for a fixed $\gamma_1 = 1$ while varying γ_2 from 0.0 to 0.1.

C.3.3 Extra Experiments Varying n/d

In the framework of statistical mechanics, the derivations are valid in the limit of high dimensions. Hence, it is reasonable to study the case where $n, d, p \to \infty$ while their ratio remains finite. In this section, we consider the case where $\frac{p}{d} = 0.5$ remains fixed while we vary $\frac{n}{d}$ from 0.1 to 3. In all the experiments we fix $\gamma_1 = 1$ and $\gamma_2 = 0.1$. Fig. C.2 shows a plot of this experiment.

C.3.4 Computational Resources

For the experiments, an approximate number of 100 GPU-hours has been used. GPUs used for the experiments are NVIDIA-V100 mostly on internal cluster and partly on public cloud clusters. Numerical simulations of the analytical expressions are performed on

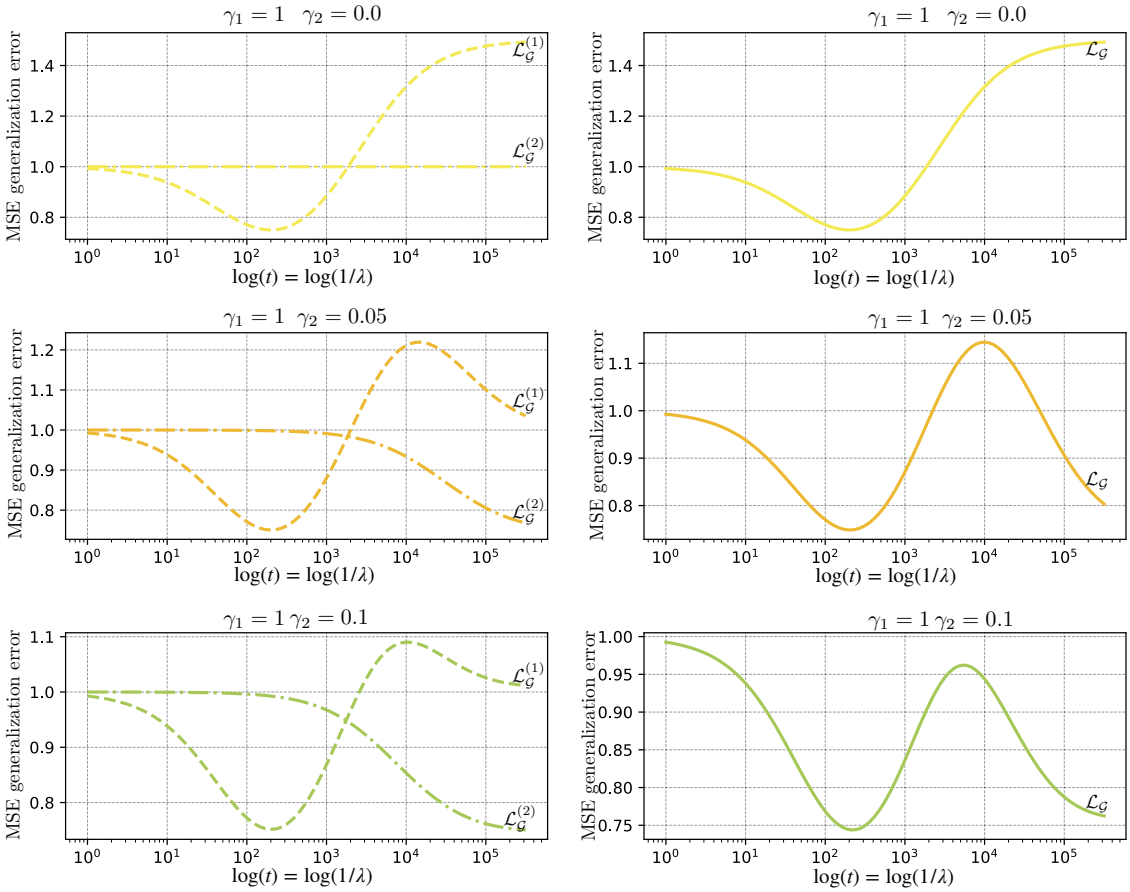

Figure C.1: The plot shows the decomposition of the generalization error into fast and slow components. The double descent curve results from overlapping of these two components.

authors' personal computer with CPU.

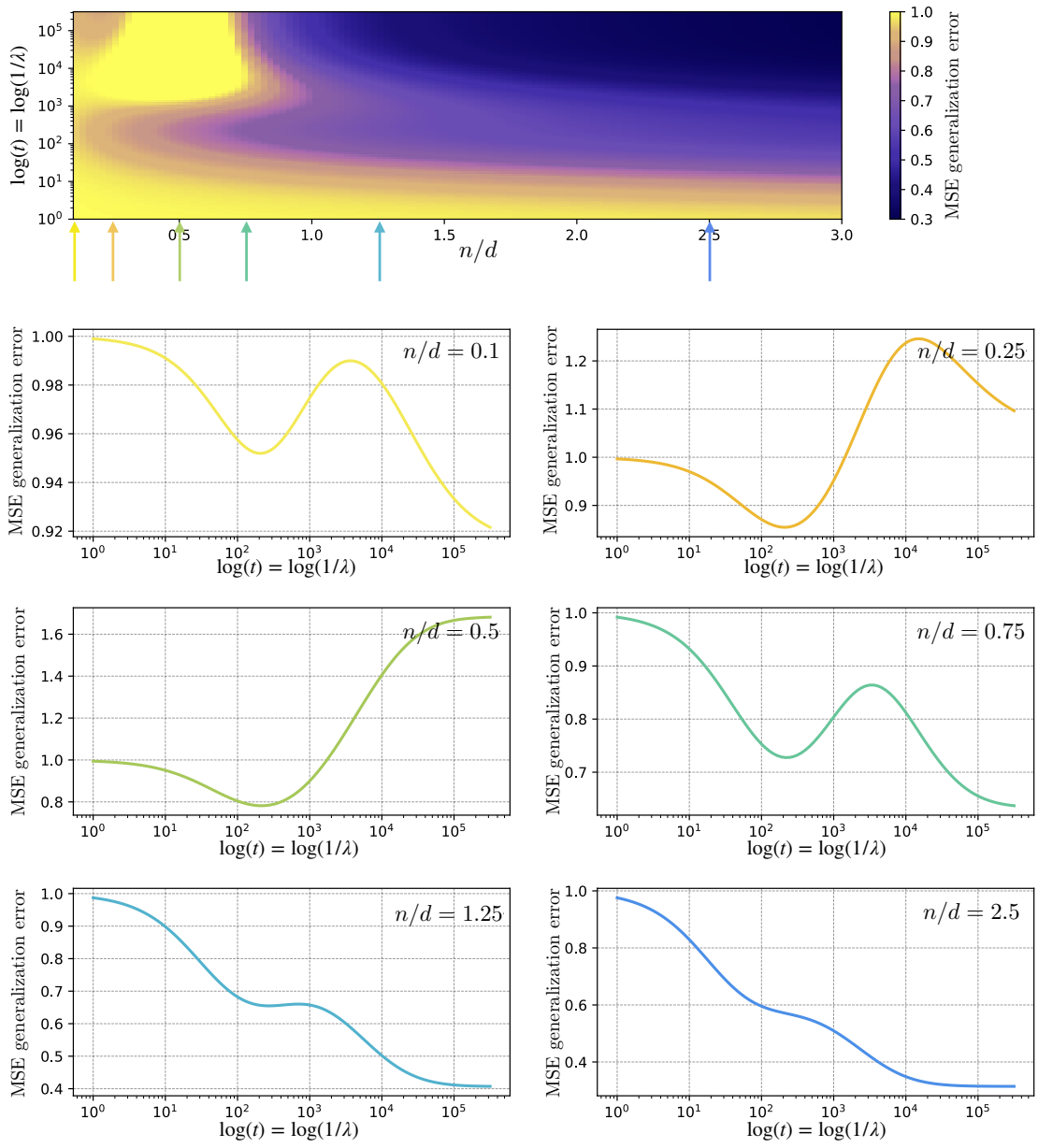

Figure C.2: The plot shows the dynamics of the generalization error as $\frac{n}{d}$ is varied from 0.1 to 3. The ratio $\frac{p}{d} = 0.5$, $\gamma_1 = 1$, and $\gamma_2 = 0.1$ are fixed.

Bibliography

[1] SIS. O. M. a. I. Radiology. Covid-19 database, 2020.

[2] Jacob Abernethy, Kevin A Lai, and Andre Wibisono. Last-iterate convergence rates for min-max optimization. *arXiv preprint arXiv:1906.02027*, 2019.

[3] Madhu Advani, Subhaneil Lahiri, and Surya Ganguli. Statistical mechanics of complex neural systems and high dimensional data. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(03):P03014, 2013.

[4] Madhu S Advani and Andrew M Saxe. High-dimensional dynamics of generalization error in neural networks. *arXiv preprint arXiv:1710.03667*, 2017.

[5] Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

[6] Faruk Ahmed and Aaron Courville. Detecting semantic anomalies. *arXiv preprint arXiv:1908.04388*, 2019.

[7] Kartik Ahuja, Karthikeyan Shanmugam, and Amit Dhurandhar. Linear regression games: Convergence guarantees to approximate out-of-distribution solutions. *arXiv preprint arXiv:2010.15234*, 2020.

[8] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. *arXiv preprint arXiv:2002.04692*, 2020.

[9] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018.

[10] Alnur Ali, Edgar Dobriban, and Ryan Tibshirani. The implicit regularization of stochastic gradient flow for least squares. In *International Conference on Machine Learning*, pages 233–244. PMLR, 2020.

[11] Alnur Ali, J Zico Kolter, and Ryan J Tibshirani. A continuous-time view of early stopping for least squares regression. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1370–1378. PMLR, 2019.

[12] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in neural information processing systems*, pages 6158–6169, 2019.

[13] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

[14] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252, 2019.

[15] Shun-ichi Amari, Jimmy Ba, Roger Grosse, Xuechen Li, Atsushi Nitanda, Taiji Suzuki, Denny Wu, and Ji Xu. When does preconditioning help or hurt generalization? *arXiv preprint arXiv:2006.10732*, 2020.

[16] Alexandr Andoni, Piotr Indyk, Thijs Laarhoven, Ilya Razenshteyn, and Ludwig Schmidt. Practical and optimal lsh for angular distance. In *Advances in neural information processing systems*, pages 1225–1233, 2015.

[17] Thomas Appelquist, Daniel Nash, and L. C. R. Wijewardhana. Critical behavior in (2+1)-dimensional qed. *Phys. Rev. Lett.*, 60:2575–2578, Jun 1988.

[18] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[19] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, pages 7413–7424, 2019.

[20] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, pages 8139–8148, 2019.

[21] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*, 2019.

[22] Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. *arXiv preprint arXiv:1706.05394*, 2017.

[23] Benjamin Aubin, Antoine Maillard, Florent Krzakala, Nicolas Macris, Lenka Zdeborová, et al. The committee machine: computational to statistical gaps in learning a two-layers neural network. In *Advances in Neural Information Processing Systems*, pages 3223–3234, 2018.

[24] N. N. Author. Suppressed for anonymity, 2020.

[25] Waïss Azizian, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games. In *International Conference on Artificial Intelligence and Statistics*, pages 2863–2873, 2020.

[26] Waïss Azizian, Damien Scieur, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. Accelerating smooth games by manipulating spectral shapes. *International Conference on Artificial Intelligence and Statistics*, 2020.

[27] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

[28] James P Bailey and Georgios Piliouras. Multi-agent learning in network zero-sum games is a hamiltonian system. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 233–241. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[29] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, 14(12):e1006613, 2018.

[30] Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

[31] D Balduzzi, S Racaniere, J Martens, J Foerster, K Tuyls, and T Graepel. *deepmind-symplectic-gradient-adjustment*, 2018. `https://github.com/deepmind/symplectic-gradient-adjustment/blob/master/Symplectic_Gradient_Adjustment.ipynb`.

[32] D Balduzzi, S Racaniere, J Martens, J Foerster, K Tuyls, and T Graepel. The mechanics of n-player differentiable games. In *ICML*, volume 80, pages 363–372. JMLR. org, 2018.

[33] Ajit C Balram, Csaba Tőke, and Jainendra K Jain. Luttinger theorem for the strongly correlated fermi liquid of composite fermions. *Physical review letters*, 115(18):186805, 2015.

[34] Aristide Baratin, Thomas George, César Laurent, R Devon Hjelm, Guillaume Lajoie, Pascal Vincent, and Simon Lacoste-Julien. Implicit regularization in deep learning: A view from function space. *arXiv preprint arXiv:2008.00938*, 2020.

[35] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

[36] Maissam Barkeshli, Michael Mulligan, and Matthew PA Fisher. Particle-hole symmetry and the composite fermi liquid. *Physical Review B*, 92(16):165125, 2015.

[37] Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

[38] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[39] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[40] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018.

[41] Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173*, 2017.

[42] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[43] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[44] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

[45] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. *arXiv preprint arXiv:1802.01396*, 2018.

[46] Yoshua Bengio and Yann LeCun. Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*. MIT Press, 2007.

[47] Yoshua Bengio, Patrice Simard, Paolo Frasconi, et al. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

[48] MV Berry and Pragya Shukla. Curl force dynamics: symmetries, chaos and constants of motion. *New Journal of Physics*, 18(6):063018, 2016.

[49] Dimitri P Bertsekas. *Nonlinear programming*. Athena scientific Belmont, 1999.

[50] Michael Betancourt, Michael I Jordan, and Ashia C Wilson. On symplectic optimization. *arXiv preprint arXiv:1802.03653*, 2018.

[51] Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. In *Advances in Neural Information Processing Systems*, pages 12893–12904, 2019.

[52] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Occam's razor. *Information processing letters*, 24(6):377–380, 1987.

[53] Mariusz Bojarski, Anna Choromanska, Krzysztof Choromanski, Francois Fagan, Cedric Gouy-Pailler, Anne Morvan, Nourhan Sakr, Tamas Sarlos, and Jamal Atif. Structured adaptive and random spinners for fast machine learning computations. *arXiv preprint arXiv:1610.06209*, 2016.

[54] S Bös, W Kinzel, and M Opper. Generalization ability of perceptrons with continuous outputs. *Physical Review E*, 47(2):1384, 1993.

[55] Siegfried Bös. Statistical mechanics approach to early stopping and weight decay. *Physical Review E*, 58(1):833, 1998.

[56] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[57] Leo Breiman. Reflections after refereeing papers for nips. *The Mathematics of Generalization*, pages 11–15, 1995.

[58] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019.

[59] Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[60] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[61] Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. Sgd learns over-parameterized networks that provably generalize on linearly separable data. *arXiv preprint arXiv:1710.10174*, 2017.

[62] Lucian Bu, Robert Babu, Bart De Schutter, et al. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

[63] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[64] Rhonald Burgos and Caio Lewenkopf. Weiss oscillations in graphene with a modulated height profile, 2016.

[65] Emmanuel J. Candès and Pragya Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *Ann. Statist.*, 48(1):27–42, 02 2020.

[66] Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning. *arXiv preprint arXiv:1912.01198*, 2019.

[67] Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*, pages 10836–10846, 2019.

[68] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[69] M. Charbonneau, K. M. van Vliet, and P. Vasilopoulos. Linear response theory revisited iii: One-body response formulas and generalized boltzmann equations. *Journal of Mathematical Physics*, 23(2):318–336, February 1982.

[70] Tatjana Chavdarova, Matteo Pagliardini, Martin Jaggi, and Francois Fleuret. Taming gans with lookahead. *arXiv preprint arXiv:2006.14567*, 2020.

[71] Lin Chen, Yifei Min, Mikhail Belkin, and Amin Karbasi. Multiple descent: Design your own generalization curve. *arXiv preprint arXiv:2008.01036*, 2020.

[72] Xiuyuan Cheng and Amit Singer. The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, 02(04):1350010, 2013.

[73] Alfred KC Cheung, S Raghu, and Michael Mulligan. Weiss oscillations and particle-hole symmetry at the half-filled landau level. *Physical Review B*, 95(23):235424, 2017.

[74] Lenaic Chizat and Francis Bach. A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956*, 1, 2018.

[75] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, pages 192–204, 2015.

[76] Krzysztof M Choromanski, Mark Rowland, and Adrian Weller. The unreasonable effectiveness of structured random orthogonal embeddings. In *Advances in Neural Information Processing Systems*, pages 219–228, 2017.

[77] Joseph Paul Cohen, Paul Morrison, and Lan Dao. Covid-19 image data collection. *arXiv preprint arXiv:2003.11597*, 2020.

[78] Remi Tachet des Combes, Mohammad Pezeshki, Samira Shabanian, Aaron Courville, and Yoshua Bengio. On the learning dynamics of deep neural networks. *arXiv preprint arXiv:1809.06848*, 2018.

[79] Thomas M Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, EC-14(3):326–334, 1965.

[80] André Belotto da Silva and Maxime Gazeau. A general system of differential equations to model first order adaptive algorithms. *arXiv preprint arXiv:1810.13108*, 2018.

[81] Amit Daniely. Sgd learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems*, pages 2422–2430, 2017.

[82] Stéphane d'Ascoli, Levent Sagun, and Giulio Biroli. Triple descent and the two kinds of overfitting: Where & why do they appear? *arXiv preprint arXiv:2006.03509*, 2020.

[83] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. In *International Conference on Learning Representations*, 2018.

[84] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1019–1028. JMLR. org, 2017.

[85] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685, 2019.

[86] Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems*, pages 384–395, 2018.

[87] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

[88] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, 2nd edition, 2000.

[89] G. V. Dunne. Course 3: Aspects of Chern-Simons Theory. In A. Comtet, T. Jolicoeur, S. Ouvry, and F. David, editors, *Topological Aspects of Low Dimensional Systems*, volume 69, page 177, January 1999.

[90] Stéphane d'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance (s) in the lazy regime. In *International Conference on Machine Learning*, pages 2280–2290. PMLR, 2020.

[91] Andreas Engel and Christian Van den Broeck. *Statistical mechanics of learning*. Cambridge University Press, 2001.

[92] Zhou Fan and Andrea Montanari. The spectral norm of random inner-product kernel matrices. *Probability Theory and Related Fields*, 173(1):27–85, Feb 2019.

[93] Mohammad Farazmand. Slow manifold analysis of accelerated gradient methods. *arXiv preprint arXiv:1807.11354*, 2018.

[94] Neil Fenichel. Geometric singular perturbation theory for ordinary differential equations. *Journal of differential equations*, 31(1):53–98, 1979.

[95] Tanner Fiez and Lillian Ratliff. Gradient descent-ascent provably converges to strict local minmax equilibria with a finite timescale separation. *arXiv preprint arXiv:2009.14820*, 2020.

[96] Jakob Foerster, Richard Y Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 122–130. International Foundation for Autonomous Agents and Multiagent Systems, 2018.

[97] Eduardo Fradkin. *Field theories of condensed matter physics*. Cambridge University Press, 2013.

[98] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[99] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*, 2014.

[100] Elizabeth Gardner. The space of interactions in neural network models. *Journal of physics A: Mathematical and general*, 21(1):257, 1988.

[101] Elizabeth Gardner and Bernard Derrida. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and general*, 21(1):271, 1988.

[102] Elizabeth Gardner and Bernard Derrida. Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A: Mathematical and General*, 22(12):1983, 1989.

[103] Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, Levent Sagun, Stéphane d'Ascoli, Giulio Biroli, Clément Hongler, and Matthieu Wyart. Scaling description of generalization with number of parameters in deep learning. *arXiv preprint arXiv:1901.01608*, 2019.

[104] Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart. Disentangling feature and lazy learning in deep neural networks: an empirical study. *arXiv preprint arXiv:1906.08034*, 2019.

[105] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020.

[106] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[107] Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

[108] Ian Gemp and Sridhar Mahadevan. Global convergence to the equilibrium of gans using variational inequalities. *arXiv preprint arXiv:1808.01531*, 2018.

[109] Thomas George. Nngeometry: Easy and fast fisher information matrices and neural tangent kernels in pytorch, 2020.

[110] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pages 3452–3462. PMLR, 2020.

[111] Scott D Geraedts, Michael P Zaletel, Roger SK Mong, Max A Metlitski, Ashvin Vishwanath, and Olexei I Motrunich. The half-filled landau level: The case for dirac composite fermions. *Science*, 352(6282):197–201, 2016.

[112] Rolf R Gerhardts. Quasiclassical calculation of magnetoresistance oscillations of a two-dimensional electron gas in spatially periodic magnetic and electrostatic fields. *Physical Review B*, 53(16):11064, 1996.

[113] RR Gerhardts, Dieter Weiss, and K v Klitzing. Novel magnetoresistance oscillations in a periodically modulated two-dimensional electron gas. *Physical review letters*, 62(10):1173, 1989.

[114] Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In *Advances in Neural Information Processing Systems*, pages 3202–3211, 2019.

[115] Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *International Conference on Learning Representations*, 2018.

[116] Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Rémi Le Priol, Gabriel Huang, Simon Lacoste-Julien, and Ioannis Mitliagkas. Negative momentum for improved game dynamics. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1802–1811, 2019.

[117] SM Girvin. Particle-hole symmetry in the anomalous quantum hall effect. *Physical Review B*, 29(10):6012, 1984.

[118] Herbert Goldstein, Charles Poole, and John Safko. Classical mechanics, 2002.

[119] 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. In *Advances in Neural Information Processing Systems*, pages 6981–6991, 2019.

[120] Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modelling the influence of data structure on learning in neural networks. *arXiv preprint arXiv:1909.11500*, 2019.

[121] Siavash Golkar, Dung Xuan Nguyen, Matthew M Roberts, and Dam Thanh Son. Higher-spin theory of the magnetorotons. *Physical review letters*, 117(21):216403, 2016.

[122] Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. *arXiv preprint arXiv:2002.00057*, 2020.

[123] Xinyu Gong. *sngan.pytorch*, 2019. `https://github.com/GongXinyuu/sngan.pytorch`.

[124] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*. MIT press Cambridge, 2016.

[125] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[126] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[127] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[128] E. V. Gorbar, V. A. Miransky, I. A. Shovkovy, and Xinyang Wang. Radiative corrections to chiral separation effect in qed. *Phys. Rev. D*, 88:025025, Jul 2013.

[129] Pallab Goswami, Xun Jia, and Sudip Chakravarty. Quantum oscillations in graphene in the presence of disorder and interactions. *Phys. Rev. B*, 78:245406, Dec 2008.

[130] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[131] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. *arXiv preprint arXiv:1802.08246*, 2018.

[132] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pages 9461–9471, 2018.

[133] Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6151–6159, 2017.

[134] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*, 2018.

[135] V. P. Gusynin and S. G. Sharapov. Magnetic oscillations in planar systems with the dirac-like spectrum of quasiparticle excitations. ii. transport properties. *Phys. Rev. B*, 71:125124, Mar 2005.

[136] VP Gusynin, VA Miransky, and IA Shovkovy. Dimensional reduction and catalysis of dynamical symmetry breaking by a magnetic field. *Nuclear Physics B*, 462(2-3):249–290, 1996.

[137] Walid Hachem, Philippe Loubaton, and Jamal Najim. Deterministic equivalents for certain functionals of large random matrices. *Ann. Appl. Probab.*, 17(3):875–930, 06 2007.

[138] Wolfgang Hahn, Hans H Hosenthien, and H Lehnigk. *Theory and application of Liapunov's direct method*. Prentice-Hall Englewood Cliffs, NJ, 1963.

[139] F. D. M. Haldane. Berry curvature on the fermi surface: Anomalous hall effect as a topological fermi-liquid property. *Phys. Rev. Lett.*, 93:206602, Nov 2004.

[140] Bertrand I Halperin, Patrick A Lee, and Nicholas Read. Theory of the half-filled landau level. *Physical Review B*, 47(12):7312, 1993.

[141] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[142] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[143] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[144] Reinhard Heckel and Fatih Furkan Yilmaz. Early stopping in deep networks: Double descent and how to eliminate it. *arXiv preprint arXiv:2007.10099*, 2020.

[145] Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *arXiv preprint arXiv:1710.11469*, 2017.

[146] Reyhane Askari Hemmat, Amartya Mitra, Guillaume Lajoie, and Ioannis Mitliagkas. Lead: Least-action dynamics for min-max optimization. *arXiv preprint arXiv:2010.13846*, 2020.

[147] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[148] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[149] Katherine L Hermann and Andrew K Lampinen. What shapes feature representations? exploring datasets, architectures, and training. *arXiv preprint arXiv:2006.12433*, 2020.

[150] Tom M Heskes and Bert Kappen. On-line learning processes in artificial neural networks. In *North-Holland Mathematical Library*, volume 51, pages 199–233. Elsevier, 1993.

[151] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017.

[152] Geoffrey E. Hinton, Simon Osindero, and Yee Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

[153] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

[154] T. Holstein, R. E. Norton, and P. Pincus. de haas-van alphen effect and the specific heat of an electron gas. *Phys. Rev. B*, 8:2649–2656, Sep 1973.

[155] Ya-Ping Hsieh, Panayotis Mertikopoulos, and Volkan Cevher. The limits of min-max optimization algorithms: convergence to spurious non-critical sets. *arXiv preprint arXiv:2006.09065*, 2020.

[156] Jiaoyang Huang and Horng-Tzer Yau. Dynamics of deep neural networks and neural tangent hierarchy. *arXiv preprint arXiv:1909.08156*, 2019.

[157] Aaron Hui, Eun-Ah Kim, and Michael Mulligan. Non-abelian bosonization and modular transformation approach to superuniversality. *Phys. Rev. B*, 99:125135, Mar 2019.

[158] Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. *arXiv preprint arXiv:2006.07322*, 2020.

[159] Adam Ibrahim, Waïss Azizian, Gauthier Gidel, and Ioannis Mitliagkas. Linear lower bounds and conditioning of differentiable games. In *International conference on machine learning*, 2020.

[160] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. *arXiv preprint arXiv:1804.08598*, 2018.

[161] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pages 125–136, 2019.

[162] Muhammad Ilyas, Hina Rehman, and Amine Nait-ali. Detection of covid-19 from chest x-ray images using artificial intelligence: An early review. *arXiv preprint arXiv:2004.05436*, 2020.

[163] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[164] Claude Itzykson and Jean-Bernard Zuber. *Quantum field theory*. Courier Corporation, 2012.

[165] Tommi S Jaakkola and David Haussler. Probabilistic kernel regression models. In *AISTATS*, 1999.

[166] Jörn-Henrik Jacobsen, Jens Behrmann, Richard Zemel, and Matthias Bethge. Excessive invariance causes adversarial vulnerability. *arXiv preprint arXiv:1811.00401*, 2018.

[167] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[168] Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Implicit regularization of random feature models. In *International Conference on Machine Learning*, pages 4631–4640. PMLR, 2020.

[169] Aukosh Jagannath, Patrick Lopatto, and Leo Miolane. Statistical thresholds for tensor PCA. *arXiv preprint arXiv:1812.03403*, 2018.

[170] Jainendra K Jain. *Composite fermions*. Cambridge University Press, 2007.

[171] Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pages 1772–1798, 2019.

[172] Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. *arXiv preprint arXiv:1909.12292*, 2019.

[173] Jason Jo and Yoshua Bengio. Measuring the tendency of cnns to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*, 2017.

[174] Alexia Jolicoeur-Martineau and Ioannis Mitliagkas. Connections between support vector machines, wasserstein distance and gradient-penalty gans. *arXiv preprint arXiv:1910.06922*, 2019.

[175] Yoshiyuki Kabashima, Tadashi Wadayama, and Toshiyuki Tanaka. A typical reconstruction limit for compressed sensing based on lp-norm minimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(09):L09003, 2009.

[176] Shamit Kachru, Michael Mulligan, Gonzalo Torroba, and Huajia Wang. Mirror symmetry and the half-filled landau level. *Phys. Rev. B*, 92:235105, Dec 2015.

[177] Dobromir Kamburov, Yang Liu, MA Mueed, Mansour Shayegan, LN Pfeiffer, KW West, and KW Baldwin. What determines the fermi wave vector of composite fermions? *Physical review letters*, 113(19):196801, 2014.

[178] Andreas Karch and David Tong. Particle-vortex duality from 3d bosonization. *Phys. Rev. X*, 6:031043, Sep 2016.

[179] Ilya Kavalerov, Wojciech Czaja, and Rama Chellappa. cgans with multi-hinge loss. *arXiv preprint arXiv:1912.04216*, 2019.

[180] M. J. Kearns. *Computational Complexity of Machine Learning*. PhD thesis, Department of Computer Science, Harvard University, 1989.

[181] V. R. Khalilov and I. V. Mamsurov. Polarization operator in the 2+1 dimensional quantum electrodynamics with a nonzero fermion density in a constant uniform magnetic field. *The European Physical Journal C*, 75(4):167, 2015.

[182] Yong Baek Kim, Akira Furusaki, Xiao-Gang Wen, and Patrick A. Lee. Gauge-invariant response functions of fermions coupled to a gauge field. *Phys. Rev. B*, 50:17917–17932, Dec 1994.

[183] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[184] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[185] SA Kivelson, DH Lee, Y Krotov, and J Gan. Composite-fermion hall conductance at $\nu=$. *Physical Review B*, 55(23):15552, 1997.

[186] Robert Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does sgd escape local minima? *arXiv preprint arXiv:1802.06175*, 2018.

[187] Simon Kornblith, Honglak Lee, Ting Chen, and Mohammad Norouzi. What's in a loss function for image classification? *arXiv preprint arXiv:2010.16402*, 2020.

[188] GM Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

[189] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[190] Anders Krogh and John A Hertz. Generalization in a linear perceptron in the presence of noise. *Journal of Physics A: Mathematical and General*, 25(5):1135, 1992.

[191] Anders Krogh and John A Hertz. A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pages 950–957, 1992.

[192] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688*, 2020.

[193] Florent Krzakala, Marc Mézard, François Sausset, YF Sun, and Lenka Zdeborová. Statistical-physics-based reconstruction in compressed sensing. *Physical Review X*, 2(2):021005, 2012.

[194] Brian Kulis, Kate Saenko, and Trevor Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR 2011*, pages 1785–1792. IEEE, 2011.

[195] Prashant Kumar, Michael Mulligan, and S Raghu. Topological phase transition underpinning particle-hole symmetry in the halperin-lee-read theory. *Physical Review B*, 98(11):115105, 2018.

[196] Prashant Kumar, Michael Mulligan, and S Raghu. Emergent reflection symmetry from nonrelativistic composite fermions. *Physical Review B*, 99(20):205151, 2019.

[197] Prashant Kumar, Srinivas Raghu, and Michael Mulligan. Composite fermion hall conductivity and the half-filled landau level. *Physical Review B*, 99(23):235114, 2019.

[198] Daniel Kunin, Javier Sagastuy-Brena, Surya Ganguli, Daniel LK Yamins, and Hidenori Tanaka. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. *arXiv preprint arXiv:2012.04728*, 2020.

[199] Andrew K Lampinen and Surya Ganguli. An analytic theory of generalization dynamics and transfer learning in deep linear networks. *arXiv preprint arXiv:1809.10374*, 2018.

[200] LD Landau and EM Lifshitz. *Course of theoretical physics. vol. 1: Mechanics.* Oxford, 1960.

[201] Lev Davidovich Landau, JS Bell, MJ Kearsley, LP Pitaevskii, EM Lifshitz, and JB Sykes. *Electrodynamics of continuous media*, volume 8. elsevier, 2013.

[202] P. Langley. Crafting papers on machine learning. In Pat Langley, editor, *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pages 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

[203] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1–8, 2019.

[204] Quoc Le, Tamás Sarlós, and Alex Smola. Fastfood-approximating kernel expansions in loglinear time. In *Proceedings of the international conference on machine learning*, volume 85, 2013.

[205] Yann Le Cun, Ido Kanter, and Sara A Solla. Eigenvalues of covariance matrices: Application to neural-network learning. *Physical Review Letters*, 66(18):2396, 1991.

[206] 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 *Advances in neural information processing systems*, pages 8570–8581, 2019.

[207] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pages 7167–7177, 2018.

[208] Sung-Sik Lee. Low-energy effective theory of fermi surface coupled with u(1) gauge field in $2+1$ dimensions. *Phys. Rev. B*, 80:165102, Oct 2009.

[209] Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

[210] Alistair Letcher, David Balduzzi, Sébastien Racaniere, James Martens, Jakob N Foerster, Karl Tuyls, and Thore Graepel. Differentiable game mechanics. *Journal of Machine Learning Research*, 20(84):1–40, 2019.

[211] Alistair Letcher, Jakob Foerster, David Balduzzi, Tim Rocktäschel, and Shimon Whiteson. Stable opponent shaping in differentiable games. *arXiv preprint arXiv:1811.08469*, 2018.

[212] Michael Levin and Dam Thanh Son. Particle-hole symmetry and electromagnetic response of a half-filled landau level. *Physical Review B*, 95(12):125120, 2017.

[213] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8157–8166, 2018.

[214] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[215] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[216] Nicolas Loizou, Hugo Berard, Alexia Jolicoeur-Martineau, Pascal Vincent, Simon Lacoste-Julien, and Ioannis Mitliagkas. Stochastic hamiltonian gradient methods for smooth games. *ICML*, 2020.

[217] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1640–1650, 2018.

[218] Edward N Lorenz. On the existence of a slow manifold. *Journal of the atmospheric sciences*, 43(15):1547–1558, 1986.

[219] Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, Apr 2018.

[220] Cosme Louart, Zhenyu Liao, Romain Couillet, et al. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

[221] Aleksandr Mikhailovich Lyapunov. The general problem of the stability of motion. *International journal of control*, 55(3):531–534, 1992.

[222] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354. PMLR, 2018.

[223] Chris J Maddison, Daniel Paulin, Yee Whye Teh, Brendan O'Donoghue, and Arnaud Doucet. Hamiltonian descent methods. *arXiv preprint arXiv:1809.05042*, 2018.

[224] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[225] H. C. Manoharan, M. Shayegan, and S. J. Klepper. Signatures of a novel fermi liquid in a two-dimensional composite particle metal. *Phys. Rev. Lett.*, 73:3270–3273, Dec 1994.

[226] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2017.

[227] A. Matulis and F. M. Peeters. Appearance of enhanced weiss oscillations in graphene: Theory. *Phys. Rev. B*, 75:125429, Mar 2007.

[228] Eric V Mazumdar, Michael I Jordan, and S Shankar Sastry. On finding local nash equilibria (and only local nash equilibria) in zero-sum games. *arXiv preprint arXiv:1901.00838*, 2019.

[229] R Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*, 2019.

[230] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

[231] Song Mei and Andrea Montanari. The generalization error of random features regression: precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

[232] Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. *arXiv preprint arXiv:1807.02629*, 2018.

[233] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of gans. In *Advances in Neural Information Processing Systems*, pages 1825–1835, 2017.

[234] Max A. Metlitski and Subir Sachdev. Quantum phase transitions of metals in two spatial dimensions. i. ising-nematic order. *Phys. Rev. B*, 82:075127, Aug 2010.

[235] Max A Metlitski and Ashvin Vishwanath. Particle-vortex duality of two-dimensional dirac fermion from electric-magnetic duality of three-dimensional topological insulators. *Physical Review B*, 93(24):245151, 2016.

[236] Marc Mézard. Mean-field message-passing equations in the hopfield model and its generalizations. *Physical Review E*, 95(2):022117, 2017.

[237] Marc Mézard, Giorgio Parisi, and Miguel Virasoro. *Spin glass theory and beyond: an introduction to the Replica Method and its applications*, volume 9. World Scientific Publishing Company, 1987.

[238] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, editors. *Machine Learning: An Artificial Intelligence Approach, Vol. I.* Tioga, Palo Alto, CA, 1983.

[239] V. A. Miransky, G. W. Semenoff, I. A. Shovkovy, and L. C. R. Wijewardhana. Color superconductivity and nondecoupling phenomena in (2+1)-dimensional qcd. *Phys. Rev. D*, 64:025005, Jun 2001.

[240] Vladimir A Miransky and Igor A Shovkovy. Quantum field theory in a magnetic field: From quantum chromodynamics to graphene and dirac semimetals. *Physics Reports*, 576:1–209, 2015.

[241] T. M. Mitchell. The need for biases in learning generalizations. Technical report, Computer Science Department, Rutgers University, New Brunswick, MA, 1980.

[242] Amartya Mitra and Michael Mulligan. Fluctuations and magnetoresistance oscillations near the half-filled landau level. *Phys. Rev. B*, 100:165122, Oct 2019.

[243] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[244] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[245] Marcin Moczulski, Misha Denil, Jeremy Appleyard, and Nando de Freitas. ACDC: A structured efficient linear layer. *arXiv preprint arXiv:1511.05946*, 2015.

[246] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: high-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

[247] Paul Mooney. Chest x-ray images (pneumonia). *Online], https://www. kaggle. com/paultimothymooney/chest-xray-pneumonia, tanggal akses*, 2018.

[248] David F. Mross, Jason Alicea, and Olexei I. Motrunich. Explicit derivation of duality between a free dirac cone and quantum electrodynamics in $(2 + 1)$ dimensions. *Phys. Rev. Lett.*, 117:016802, Jun 2016.

[249] David F. Mross, John McGreevy, Hong Liu, and T. Senthil. Controlled expansion for certain non-fermi-liquid metals. *Phys. Rev. B*, 82:045121, Jul 2010.

[250] Michael Muehlebach and Michael Jordan. A dynamical systems perspective on nesterov acceleration. In *International Conference on Machine Learning*, pages 4656–4662, 2019.

[251] Ganpathy Murthy and R. Shankar. $\nu = \frac{1}{2}$ landau level: Half-empty versus half-full. *Phys. Rev. B*, 93:085405, Feb 2016.

[252] Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83, 2020.

[253] Vaishnavh Nagarajan and J Zico Kolter. Gradient descent gan optimization is locally stable. In *Advances in neural information processing systems*, pages 5585–5595, 2017.

[254] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: where bigger models and more data hurt. *ICLR 2020, arXiv preprint arXiv:1912.02292*, 2019.

[255] Preetum Nakkiran, Gal Kaplun, Dimitris Kalimeris, Tristan Yang, Benjamin L Edelman, Fred Zhang, and Boaz Barak. Sgd on neural networks learns functions of increasing complexity. *arXiv preprint arXiv:1905.11604*, 2019.

[256] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. *arXiv preprint arXiv:2007.02561*, 2020.

[257] Kamil Nar, Orhan Ocal, S Shankar Sastry, and Kannan Ramchandran. Cross-entropy loss and low-rank features have responsibility for adversarial examples. *arXiv preprint arXiv:1901.08360*, 2019.

[258] Kamil Nar and S Shankar Sastry. Persistency of excitation for robustness of neural networks. *arXiv preprint arXiv:1911.01043*, 2019.

[259] Ali Narin, Ceren Kaya, and Ziynet Pamuk. Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. *arXiv preprint arXiv:2003.10849*, 2020.

[260] Daniel Nash. Higher-order corrections in (2+1)-dimensional qed. *Phys. Rev. Lett.*, 62:3024–3026, Jun 1989.

[261] Brady Neal, Sarthak Mittal, Aristide Baratin, Vinayak Tantia, Matthew Scicluna, Simon Lacoste-Julien, and Ioannis Mitliagkas. A modern take on the bias-variance tradeoff in neural networks. *arXiv preprint arXiv:1810.08591*, 2018.

[262] Arkadii Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

[263] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[264] A. Newell and P. S. Rosenbloom. Mechanisms of skill acquisition and the law of practice. In J. R. Anderson, editor, *Cognitive Skills and Their Acquisition*, chapter 1, pages 1–51. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 1981.

[265] Behnam Neyshabur. Implicit regularization in deep learning. *arXiv preprint arXiv:1709.01953*, 2017.

[266] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 5947–5956, 2017.

[267] Behnam Neyshabur, Ryota Tomioka, Ruslan Salakhutdinov, and Nathan Srebro. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071*, 2017.

[268] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

[269] Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. *arXiv preprint arXiv:1907.07355*, 2019.

[270] Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems*, pages 14934–14942, 2019.

[271] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 151–159, 2020.

[272] Ruben Ohana, Jonas Wacker, Jonathan Dong, Sébastien Marmin, Florent Krzakala, Maurizio Filippone, and Laurent Daudet. Kernel computations from large-scale random features obtained by optical processing units. *arXiv preprint arXiv:1910.09880*, 2019.

[273] Manfred Opper. Statistical mechanics of learning: Generalization. *The handbook of brain theory and neural networks*, pages 922–925, 1995.

[274] Manfred Opper and Wolfgang Kinzel. Statistical mechanics of generalization. In *Models of neural networks III*, pages 151–209. Springer, 1996.

[275] Samet Oymak, Zalan Fabian, Mingchen Li, and Mahdi Soltanolkotabi. Generalization guarantees for neural networks via harnessing the low-rank structure of the jacobian. *arXiv preprint arXiv:1906.05392*, 2019.

[276] Brendan O'donoghue and Emmanuel Candes. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.

[277] W Pan, W Kang, MP Lilly, JL Reno, KW Baldwin, KW West, LN Pfeiffer, and DC Tsui. Particle-hole symmetry and the fractional quantum hall effect in the lowest landau level. *Physical review letters*, 124(15):156801, 2020.

[278] Giambattista Parascandolo, Alexander Neitz, Antonio Orvieto, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary. *arXiv preprint arXiv:2009.00329*, 2020.

[279] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. s. 2017.

[280] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: An overview of recent advances. *IEEE Signal Processing Magazine*, 2014.

[281] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[282] FM Peeters and P Vasilopoulos. Electrical and thermal properties of a two-dimensional electron gas in a one-dimensional periodic potential. *Physical Review B*, 46(8):4667, 1992.

[283] FM Peeters and P Vasilopoulos. Quantum transport of a two-dimensional electron gas in a spatially modulated magnetic field. *Physical Review B*, 47(3):1466, 1993.

[284] Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems 30*, pages 2637–2646. 2017.

[285] Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *arXiv preprint arXiv:2011.09468*, 2020.

[286] Oskar Pfungst. *Clever Hans:(the horse of Mr. Von Osten.) a contribution to experimental animal and human psychology*. Holt, Rinehart and Winston, 1911.

[287] Robert D. Pisarski. Chiral-symmetry breaking in three-dimensional electrodynamics. *Phys. Rev. D*, 29:2423–2426, May 1984.

[288] Tomaso Poggio, Kenji Kawaguchi, Qianli Liao, Brando Miranda, Lorenzo Rosasco, Xavier Boix, Jack Hidary, and Hrushikesh Mhaskar. Theory of deep learning iii: explaining the non-overfitting puzzle. *arXiv preprint arXiv:1801.00173*, 2017.

[289] Tomaso Poggio, Ryan Rifkin, Sayan Mukherjee, and Partha Niyogi. General conditions for predictivity in learning theory. *Nature*, 428(6981):419, 2004.

[290] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[291] Chongli Qin, Yan Wu, Jost Tobias Springenberg, Andrew Brock, Jeff Donahue, Timothy P Lillicrap, and Pushmeet Kohli. Training generative adversarial networks by solving ordinary differential equations. *arXiv preprint arXiv:2010.15040*, 2020.

[292] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, pages 6076–6085, 2017.

[293] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019.

[294] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pages 1177–1184. 2008.

[295] Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems 21*, pages 1313–1320. 2009.

[296] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. *arXiv preprint arXiv:1707.04131*, 2017.

[297] N. Read. Recent progress in the theory of composite fermions near even-denominator filling factors. *Surface Science*, 361:7–12, July 1996.

[298] Frederick Reif. *Fundamentals of statistical and thermal physics*. Waveland Press, 2009.

[299] Edward H Rezayi and F Duncan M Haldane. Incompressible paired hall state, stripe order, and the composite fermion liquid phase in half-filled landau levels. *Physical Review Letters*, 84(20):4685, 2000.

[300] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[301] V. I. Ritus. Eigenfunction method and mass operator in the quantum electrodynamics of a constant field. *Zhurnal Ehksperimental'noj i Teoreticheskoj Fiziki*, 17(5):1560–1583, 1978.

[302] Nicholas Rombes and Sudip Chakravarty. Specific heat and pairing of dirac composite fermions in the half-filled landau level. *Annals of Physics*, 409:167915, 2019.

[303] Basri Ronen, David Jacobs, Yoni Kasten, and Shira Kritchman. The convergence rate of neural networks for learned functions of different frequencies. In *Advances in Neural Information Processing Systems*, pages 4761–4771, 2019.

[304] Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

[305] Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. Falkon: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems*, pages 3888–3898, 2017.

[306] Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3215–3225. Curran Associates, Inc., 2017.

[307] Ernest K Ryu, Kun Yuan, and Wotao Yin. Ode analysis of stochastic gradient methods with optimism and anchoring for minimax problems and gans. *arXiv preprint arXiv:1905.10899*, 2019.

[308] Alaa Saade, Francesco Caltagirone, Igor Carron, Laurent Daudet, Angélique Drémeau, Sylvain Gigan, and Florent Krzakala. Random projections through multiple optical scattering: approximating kernels at the speed of light. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6215–6219. IEEE, 2016.

[309] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[310] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. *arXiv preprint arXiv:2005.04345*, 2020.

[311] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

[312] A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):211–229, 1959.

[313] 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.

[314] Andrew M Saxe, James L McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, 2019.

[315] Florian Schäfer and Anima Anandkumar. Competitive gradient descent. In *Advances in Neural Information Processing Systems*, pages 7623–7633, 2019.

[316] Florian Schäfer, Hongkai Zheng, and Anima Anandkumar. Implicit competitive regularization in gans. *arXiv preprint arXiv:1910.05852*, 2019.

[317] Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[318] Julian Schwinger. On gauge invariance and vacuum polarization. *Phys. Rev.*, 82:664–679, Jun 1951.

[319] Damien Scieur, Vincent Roulet, Francis Bach, and Alexandre d'Aspremont. Integration methods and optimization algorithms. In *Advances in Neural Information Processing Systems*, pages 1109–1118, 2017.

[320] Mohamed El Amine Seddik, Cosme Louart, Mohamed Tamaazousti, and Romain Couillet. Random matrix theory proves that deep learning representations of gan-data behave as gaussian mixtures. *arXiv preprint arXiv:2001.08370*, 2020.

[321] Mohamed El Amine Seddik, Mohamed Tamaazousti, and Romain Couillet. Kernel random matrices of large concentrated data: the example of gan-generated images. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7480–7484. IEEE, 2019.

[322] Nathan Seiberg, T. Senthil, Chong Wang, and Edward Witten. A duality web in 2+1 dimensions and condensed matter physics. *Annals of Physics*, 374:395–433, 2016.

[323] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, pages 527–538, 2018.

[324] T Senthil, Dam Thanh Son, Chong Wang, and Cenke Xu. Duality between (2+ 1) d quantum critical points. *Physics Reports*, 827:1–48, 2019.

[325] Hyunjune Sebastian Seung, Haim Sompolinsky, and Naftali Tishby. Statistical mechanics of learning from examples. *Physical review A*, 45(8):6056, 1992.

[326] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. *arXiv preprint arXiv:2006.07710*, 2020.

[327] Dan Shahar, DC Tsui, M Shayegan, RN Bhatt, and JE Cunningham. Universal conductivity at the quantum hall liquid to insulator transition. *Physical review letters*, 74(22):4511, 1995.

[328] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[329] Mansour Shayegan. Probing composite fermions near half-filled landau levels. In *Fractional Quantum Hall Effects: New Developments*, pages 133–181. World Scientific, 2020.

[330] Bin Shi, Simon S Du, Michael I Jordan, and Weijie J Su. Understanding the acceleration phenomenon via high-resolution differential equations. *arXiv preprint arXiv:1810.08907*, 2018.

[331] Bin Shi, Simon S Du, Weijie Su, and Michael I Jordan. Acceleration via symplectic discretization of high-resolution differential equations. In *Advances in Neural Information Processing Systems*, pages 5745–5753, 2019.

[332] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[333] Steve Smale. The fundamental theorem of algebra and complexity theory. *Bulletin of the American Mathematical Society*, 4(1):1–36, 1981.

[334] Dam Thanh Son. Is the composite fermion a dirac particle? *Physical Review X*, 5(3):031027, 2015.

[335] Jun Ho Son, Jing-Yuan Chen, and S. Raghu. Duality web on a 3d euclidean lattice and manifestation of hidden symmetries. *Journal of High Energy Physics*, 2019(6):38, 2019.

[336] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[337] S Spigler, M Geiger, S d'Ascoli, L Sagun, G Biroli, and M Wyart. A jamming transition from under-to over-parametrization affects generalization in deep learning. *Journal of Physics A: Mathematical and Theoretical*, 52(47):474001, 2019.

[338] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[339] Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.

[340] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.

[341] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[342] M. Tahir and K. Sabeeh. Quantum transport of dirac electrons in graphene in the presence of a spatially modulated magnetic field. *Phys. Rev. B*, 77:195421, May 2008.

[343] Michel Talagrand. The Parisi formula. *Annals of mathematics*, 163:221–263, 2006.

[344] Shichang Tang. Lessons learned from the training of gans on artificial datasets. *IEEE Access*, 8:165044–165055, 2020.

[345] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[346] Guillermo Valle-Pérez, Chico Q Camargo, and Ard A Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. *arXiv preprint arXiv:1805.08522*, 2018.

[347] Vladimir N. Vapnik. *The nature of statistical learning theory*. Wiley, New York, 1st edition, September 1998.

[348] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[349] Santosh Vempala and John Wilmes. Gradient descent for one-hidden-layer neural networks: Polynomial convergence and sq lower bounds. In *Conference on Learning Theory*, pages 3115–3117, 2019.

[350] Joseph D Viviano, Becks Simpson, Francis Dutil, Yoshua Bengio, and Joseph Paul Cohen. Underwhelming generalization improvements from controlling feature attribution. *arXiv preprint arXiv:1910.00199*, 2019.

[351] Chong Wang, Nigel R Cooper, Bertrand I Halperin, and Ady Stern. Particle-hole symmetry in the fermion-chern-simons and dirac descriptions of a half-filled landau level. *Physical Review X*, 7(3):031029, 2017.

[352] Chong Wang and T. Senthil. Dual dirac liquid on the surface of the electron topological insulator. *Phys. Rev. X*, 5:041031, Nov 2015.

[353] Chong Wang and T Senthil. Composite fermi liquids in the lowest landau level. *Physical Review B*, 94(24):245107, 2016.

[354] Haohan Wang, Zexue He, Zachary C Lipton, and Eric P Xing. Learning robust representations by projecting superficial statistics out. *arXiv preprint arXiv:1903.06256*, 2019.

[355] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[356] Yuanhao Wang, Guodong Zhang, and Jimmy Ba. On solving minimax optimization locally: A follow-the-ridge approach. In *International Conference on Learning Representations*, 2019.

[357] Timothy LH Watkin, Albrecht Rau, and Michael Biehl. The statistical mechanics of learning a rule. *Reviews of Modern Physics*, 65(2):499, 1993.

[358] P. Watson and H. Reinhardt. Quark gap equation in an external magnetic field. *Phys. Rev. D*, 89:045008, Feb 2014.

[359] Dieter Weiss. Magnetoquantum oscillations in a lateral superlattice. In *Electronic properties of multilayers and low-dimensional semiconductor structures*, pages 133–150. Springer, 1990.

[360] Dieter Weiss, KV Klitzing, K Ploog, and G Weimann. Magnetoresistance oscillations in a two-dimensional electron gas induced by a submicrometer periodic potential. *EPL (Europhysics Letters)*, 8(2):179, 1989.

[361] Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.

[362] Robert L Willett. Experimental evidence for composite fermions. *Advances in physics*, 46(5):447–544, 1997.

[363] Ashia Wilson, Lester Mackey, and Andre Wibisono. Accelerating rescaled gradient descent. *arXiv preprint arXiv:1902.08825*, 2019.

[364] Ashia C Wilson, Benjamin Recht, and Michael I Jordan. A lyapunov analysis of momentum methods in optimization. *arXiv preprint arXiv:1611.02635*, 2016.

[365] RW Winkler, JP Kotthaus, and K Ploog. Landau band conductivity in a two-dimensional electron system modulated by an artificial one-dimensional superlattice potential. *Physical review letters*, 62(10):1177, 1989.

[366] David H Wolpert. The lack of a priori distinctions between learning algorithms. *Neural computation*, 8(7):1341–1390, 1996.

[367] LW Wong, HW Jiang, and WJ Schaff. Universality and phase diagram around half-filled landau levels. *Physical Review B*, 54(24):R17323, 1996.

[368] Blake Woodworth, Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Kernel and deep regimes in overparametrized models. *arXiv preprint arXiv:1906.05827*, 2019.

[369] Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523*, 2019.

[370] Zhi-Qin John Xu, Yaoyu Zhang, and Yanyang Xiao. Training behavior of deep neural network in frequency domain. In *International Conference on Neural Information Processing*, pages 264–274. Springer, 2019.

[371] Greg Yang and Hadi Salman. A fine-grained spectral perspective on neural networks. *arXiv preprint arXiv:1907.10599*, 2019.

[372] Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning*, pages 10767–10777. PMLR, 2020.

[373] Yasin Yazıcı, Chuan-Sheng Foo, Stefan Winkler, Kim-Hui Yap, Georgios Piliouras, and Vijay Chandrasekhar. The unusual effectiveness of averaging in gan training, 2019.

[374] D Yoshioka, Bertrand I Halperin, and PA Lee. Ground state of two-dimensional electrons in strong magnetic fields and 1 3 quantized hall effect. *Physical review letters*, 50(16):1219, 1983.

[375] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[376] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

[377] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric K Oermann. Confounding variables can degrade generalization performance of radiological deep learning models. *arXiv preprint arXiv:1807.00431*, 2018.

[378] Chao Zhang and Rolf R Gerhardts. Theory of magnetotransport in two-dimensional electron systems with unidirectional periodic modulation. *Physical Review B*, 41(18):12850, 1990.

[379] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[380] Marvin Zhang, Henrik Marklund, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: A meta-learning approach for tackling group shift. *arXiv preprint arXiv:2007.02931*, 2020.

[381] Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *The Journal of Machine Learning Research*, 16(1):3299–3340, 2015.

[382] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

[383] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109(3):467–492, 2020.