Zeroth, first, and second-order phase transitions in deep neural networks

Liu Ziyin[1] and Masahito Ueda[1,2,3]

[1]*Department of Physics, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan*

[2]*RIKEN Center for Emergent Matter Science (CEMS), Wako, Saitama 351-0198, Japan*

[3]*Institute for Physics of Intelligence, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan*

(Received 8 November 2022; revised 31 October 2023; accepted 6 November 2023; published 14 December 2023)

We investigate deep-learning-unique first-order and second-order phase transitions, whose phenomenology closely follows that in statistical physics. In particular, we prove that the competition between prediction error and model complexity in the training loss leads to the second-order phase transition for deep linear nets with one hidden layer and the first-order phase transition for nets with more than one hidden layer. We also prove the linear origin theorem, which states that common deep nonlinear models are equivalent to a linear network of the same depth and connection structure close to the origin. Therefore, the proposed theory is directly relevant to understanding the optimization and initialization of neural networks and serves as a minimal model of the ubiquitous collapse phenomenon in deep learning.

DOI: 10.1103/PhysRevResearch.5.043243

I. INTRODUCTION

Understanding neural networks is a fundamental problem in both theoretical deep learning and neuroscience. In deep learning, learning proceeds as the parameters of different layers become structured so the model outputs correlate meaningfully to inputs. This is reminiscent of an ordered phase in physics, where the microscopic degrees of freedom respond to external perturbations collectively. Meanwhile, regularization prevents model overfitting by limiting the correlation between model output and input, like an entropic force in physics that leads to disorder. One thus expects a phase transition from the regime where the regularization is negligible to a regime where it is dominant. In the field of statistical physics of learning [1–12], a series of works studied the under-to-overparametrization (UO) phase transition in the context of linear regression [4,13–15]. Recently, this type of phase transition has seen a resurgence of interest [16,17]. However, the UO phase transition is not unique to deep learning because it appears in both shallow and deep models as well as in non-neural-network models [18]. To understand deep learning, we need to identify what is unique about deep neural networks.

In this paper, we study the loss landscape of a deep neural network and prove that there exist phase transitions that can be described as the first- and second-order phase transitions with a striking similarity to statistical physics. We argue that these phase transitions can have profound implications for deep learning, such as the role of symmetry breaking, the qualitative distinction between shallow and deep architectures, and why collapses occur so frequently in deep learning. For a multilayer linear net with stochastic neurons and trained with L_2 regularization:

(1) We identify an order parameter and effective landscape that describes the phase transition between a trivial phase and a feature learning phase.

(2) We prove that

(a) depth-0 nets (linear regression) do not have a phase transition,

(b) depth-1 nets have the second-order phase transitions,

(c) depth-D nets have the first-order phase transition for $D > 1$, and

(d) infinite-depth nets have the zeroth-order phase transition.

(3) We prove that such networks approximate commonly used nonlinear networks of the same depth and connectivity structure.

See Fig. 1 for an illustration of this phenomenology in comparison with the proposed theory. Our result implies that one can precisely classify the landscape of deep neural models according to the Ehrenfest classification of phase transitions. Lastly, we discuss the relevance of the theory towards understanding optimization and initialization of neural networks.

This paper is organized as follows. Section II introduces the theoretical formulation of the problem. Section III presents the main results. Section IV discusses the implications of the theory. We present all the proofs in the Appendices.

II. FORMAL FRAMEWORK

Let $\ell(w, \gamma)$ be a differentiable loss function that is dependent on the model parameter w and a hyperparameter γ. The loss function ℓ can be decomposed into a data-dependent feature-learning term ℓ_0 and a data-independent term $\gamma R(w)$

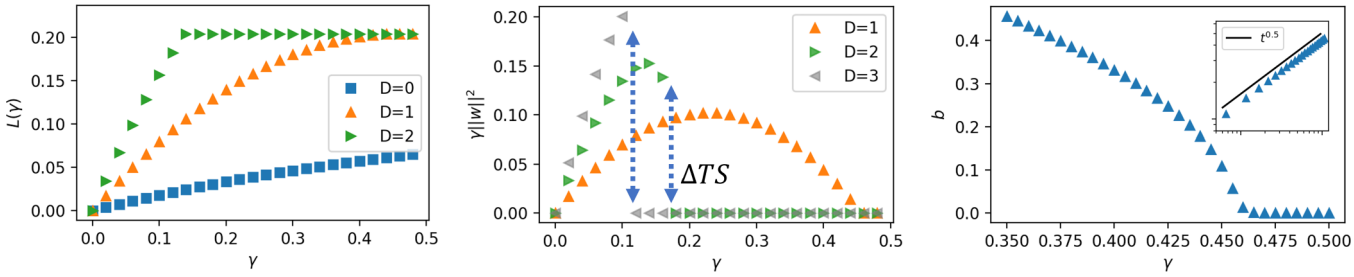

FIG. 1. Phase transitions in a linear net. In agreement with the theory, a depth-0 net has no phase transition. A depth-1 net has a second-order phase transition at approximately $\gamma = 0.45$, which is close to the theoretical value of $||\mathbb{E}[xy]||$, and a depth-2 net has a first-order phase transition at roughly $\gamma = 0.15$. The qualitative distinctions between networks of different depths are clearly seen in the data. Left: Training loss of a network with zero (linear regression), one, and two hidden layers. Middle: Magnitude of the regularization term at convergence. We see that for $D > 1$, there is a discontinuous jump from a nonzero value to 0. These jumps correspond to the latent heat ΔTS of the first-order phase transition process. Right: Order parameter b as a function of γ. The inset shows that b scales as $t^{0.5}$ with $t := -(\gamma - \gamma^*)$ in the vicinity of the phase transition, in agreement with the Landau theory of phase transitions.

that regularizes the model at strength γ:

$$\ell(w, \gamma) = \mathbb{E}_x[\ell_0(w, x)] + \gamma R(w). \tag{1}$$

Learning amounts to finding the global minimizer of the loss function:

$$L(\gamma) := \min_w \ell(w, \gamma)$$
$$w_* := \arg\min_w \ell(w, \gamma). \tag{2}$$

When γ is large, $\gamma R(w)$ causes w to stay close to zero. If L changes drastically against a small variation of γ, it is hard to tune γ to optimize the model performance. Thus, that $L(\gamma)$ is well-behaved is equivalent to γ being an easy-to-tune hyperparameter. We are thus interested in the case where the tuning of γ is difficult, which occurs where a phase transition comes into play. This formalism can be seen as a zero-temperature theory that ignores the stochastic effects due to the noise in the stochastic gradient Langevin dynamics. When the training proceeds with gradient flow and an injected isotropic Gaussian noise (namely, with the stochastic gradient Langevin dynamics algorithm), the stationary distribution of the model parameters obeys the Gibbs distribution,

$$p(w) \propto \exp[-\ell(w, \gamma)/T], \tag{3}$$

where T is proportional to the variance of the injected Gaussian noise in the gradient, and the partition function Z is given by the integral of $\exp[-\ell(w, \gamma)/T]$ over w, and the free energy is given by $-T \log Z$. In the limit $T \to 0_+$, the partition function approaches the global minimizer of the loss function ℓ:

$$F(T = 0_+, \gamma) = -\lim_{T \to 0_+} T \log Z = L(\gamma). \tag{4}$$

Therefore, γ is something like a nontemperature macroscopic thermodynamic variable, a little bit analogous to pressure. In fact, this identification of the free energy is common in the replica-symmetry treatment of the learning of neural networks [19]. Also, see Ref. [20] for a Bayesian setup of the partition function. In this view, optimization of the objective thus involves balancing the prediction error and the model complexity [15,21–23].

Since we will be considering the phase transitions at zero temperature, it is worthwhile to remark that in the Ehrenfest framework, a zero-temperature phase transition for a finite-size system is not forbidden. For a finite-temperature system, it is well-known that phase transitions can only happen when the system size tends to infinity. This is because the Gibbs measure $e^{-E/T}$ is analytic for any finite T and finite system. However, functions of the Gibbs measure (such as the free energy) are not guaranteed to be analytic in the limit $T \to 0$, which will serve as the mathematical basis behind the finite-size zero-temperature phase transitions we study in this paper.

We formally define the order parameter and the effective loss as follows.

Definition 1. $b = b(w) \in \mathbb{R}$ is said to be an order parameter of $\ell(w, \gamma)$ if there exists a function $\bar{\ell}$ such that for all γ, $\min_w \bar{\ell}(b(w), \gamma) = L(\gamma)$, where $\bar{\ell}$ is said to be an effective loss function of ℓ.

In other words, an order parameter is a one-dimensional quantity whose minimization on $\bar{\ell}$ gives $L(\gamma)$ [24]. Physical examples include the average magnetization in the Ising model and the average density of molecules in a gas-liquid transition. We establish the correspondence between key quantities in the learning process and those in statistical physics as listed in Table I.

The primary minimal model for deep learning is a deep linear network [25–27]. The most general type of deep linear nets, with L_2 regularization and stochastic neurons, has the following loss:

$$\mathbb{E}_{x,\epsilon}\left(\sum_{i_0,i_1,i_2,\ldots,i_D}^{d,d_0,d_0,\ldots d_0} \prod_{j=0}^{D} \epsilon_{i_j}^{(j)} W_{i_{j+1}i_j}^{(j)} x_{i_0} - y\right)^2 + \sum_{i=0}^{D} \gamma ||W^{(i)}||_F^2, \tag{5}$$

TABLE I. The correspondence between a learning process and phase transition.

Norm of model ($\|b\|$)	Magnitude of order parameter
Feature learning regime	Ordered phase
Trivial regime	Disordered phase
Noise required for learning	Latent heat

where $x \in \mathbb{R}^d$ is the input data, $y = y(x)$ the label, $W^{(i)}$ the model parameters, D the number of hidden layers, ϵ the randomness in the hidden layer (e.g., due to training techniques such as dropout [28]), d_0 the width of the model, and γ the weight decay strength. As is common practice, we let $\epsilon^{(D)} = 1$ deterministically. We also denote the weight matrix of the last layer by $U, U := W^{(D)}$. Comparing this equation with Eq. (1), one can identify the first term as $\mathbb{E}[\ell_0]$ and the second term as $\gamma R(w)$.

Let s denote the sign of the first element of U, $b = s||U||/d_0$ [29], $A_0 := \mathbb{E}[xx^T]$, and a_i be the ith eigenvalue of A_0. We note that the incorporation of the sign is not essential in the Ehrenfest framework but will help our discussion of symmetry breaking later. When $D > 0$, we can use these conditions in Eq. (5) to reduce it to a one-dimensional effective loss function [30],

$$\bar{\ell}(b, \gamma) := -\sum_i \frac{d_0^{2D} b^{2D} \mathbb{E}[x'y]_i^2}{d_0^D(\sigma^2 + d_0)^D a_i b^{2D} + \gamma} + \mathbb{E}_x[y^2] + \gamma D d_0^2 b^2, \tag{6}$$

where x' is a rotation of x: $x' = Rx$, where R is the eigenvectors of A_0. By Definition 1, b is the order parameter. See Figs. 2(a) and 2(b). The complicated landscape for $D > 1$ implies that neural networks are susceptible to initialization schemes and that entrapment in metastable states is common [31].

It is clear that $b = 0$ is a special point of the effective loss, and we are interested in the case when $b = 0$ is the global minimum of the landscape and how it makes a transition away from $b = 0$. The two phases also have a clear meaning in the context of machine learning: $b = 0$ is a trivial phase where no learning happens, and $b > 0$ is the nontrivial phase where learning should occur for the model to reach the global minimum.

Before we present our main results that are exact, we first provide a perturbative analysis for better understanding. When γ is large, one can expand the loss function around the origin:

$$\bar{\ell} \propto \gamma^{-2} c_0' \mathbb{E}[x^2] b^{4D} - \gamma^{-1} c_1' b^{2D} + \gamma c_2' b^2 + \text{const.} \tag{7}$$

Here, c_0', c_1', and c_2' are positive structural constants, depending on both the model (depth, width, etc.) and the data distribution. The first and third terms monotonically increase with b, thus suppressing b. The second term monotonically decreases in b^{2D}, which tends to build a positive correlation between b and the feature $\mathbb{E}[xy]$. The leading and lowest-order terms regularize the model, while the second term characterizes learning. For $D = 1$, the perturbative loss is identical to the Landau free energy, and a phase transition occurs when the second-order term flips the sign: $c_2'\gamma^2 = c_1'$. For $D > 1$, the origin is always a local minimum, dominated by the quadratic term. This leads to a first-order phase transition. When $D \to \infty$, the leading terms become discontinuous in b, and one obtains a zeroth-order phase transition. This simple analysis highlights one important distinction between physics and machine learning: in physics, the most common type of interaction is a two-body interaction, whereas, for machine learning, typical interactions are many body and tend to become infinite body as D increases [32]. Now, we present our main results. We stress that the proof of the main results is nonperturbative.

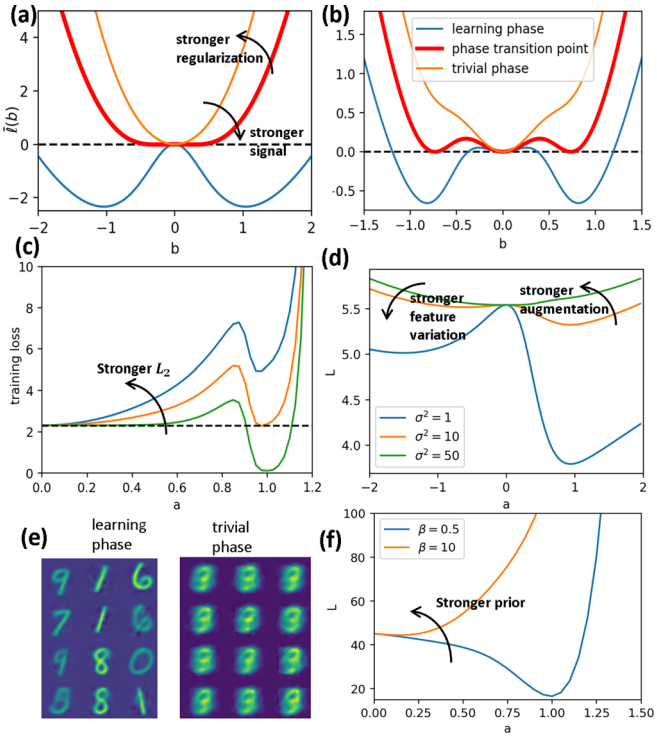

FIG. 2. Effective landscape given in Eq. (6) for (a) $D = 1$ and (b) $D = 2$. For $D = 1$, zero is either the global minimum or a local maximum. Note that the shape of the loss resembles that of the Landau free energy for the second-order phase transition. For $D = 2$, the landscape becomes more complicated, featuring the emergence of local minima. In particular, zero is always a local minimum. (c) Landscape of ResNet18 on CIFAR10 with cross-entropy loss in supervised learning. (d) The landscape of self-supervised learning for a ResNet18. A strong data augmentation leads to the trivial phase, while a stronger data variation leads to the learning phase. (e) The trivial phase and learning phase also emerge in generative models such as variational autoencoders (VAEs). (f) The landscape of VAE changes qualitatively as we change the strength, β, of the prior term of β-VAE [33].

III. PHASE TRANSITIONS

A. No-phase-transition theorems

Before we discuss Eq. (6) in detail, it is worth discussing the case of $D = 0$, which serves as a basis of comparison. Here, the global minimum as a function of γ is given by the ridge linear regression solution: $w^* = [A_0 + \gamma I]^{-1}\mathbb{E}[xy]$. L is then everywhere analytic. Thus, there is no phase transition in any hyperparameter (γ, $\mathbb{E}[xx^T]$, $\mathbb{E}[xy]$). In the parlance of physics, a linear regressor operates within the linear-response regime. Formally, one can state this result in the following way.

Theorem 1. There is no phase transition in any hyperparameter (γ, A_0, $E[xy]$, $E[y^2]$) in a simple ridge linear regression for any $\gamma \in (0, \infty)$.

We also prove a theorem showing that a finite-depth net cannot have zeroth-order phase transitions. Note that this theorem allows the weight decay parameter to be 0, and so our results also extend to the case when there is no weight decay.

Theorem 2. For any finite $D > 0$ and $\gamma \in [0, \infty)$, $L(\gamma)$ has no zeroth-order phase transition with respect to γ.

This theorem can be seen as a worst-case guarantee: the training loss needs to change continuously as one changes the hyperparameter. We note that this general theorem applies to standard nonlinear networks as well. Indeed, if we only consider the global minimum of the training loss, the training loss cannot jump. However, in practice, one can often observe jumps because the gradient-based algorithms can be trapped in local minima.

B. Phase transitions in deeper networks

The following theorem proves that a second-order phase transition from the trivial regime to the learning regime happens at a precise critical point.

Theorem 3. Equation (5) has the second-order phase transition between the trivial and feature learning phases at [34]

$$\gamma = ||\mathbb{E}[xy]||. \tag{8}$$

This critical point is surprisingly clean and invariant to minor details of the problem. It is independent of the width of the model, the feature variation A_0 of the data, or even the stochasticity σ^2 in the neurons. In fact, it is easy to show that close to the critical point $b^* \propto \sqrt{||\mathbb{E}[xy]|| - \gamma}$, featuring the typical expectation of Landau theory. In machine learning, γ is the regularization strength and $||E[xy]||$ is the signal. The phase transition occurs precisely when the regularization dominates the signal. Also, the phase transition for a depth-1 linear net is independent of the number of parameters of the model. For $D > 1$, the model size plays a role in determining the nature of the phase transition. However, γ remains the dominant variable controlling this phase transition. This independence of the model size is an advantage of the proposed theory because our result becomes directly relevant for all model sizes, not just the infinitely large ones often adopted by previous works.

For $D \geqslant 2$, we show that a first-order phase transition between the two phases at some $\gamma > 0$ exists, as the following theorem shows.

Theorem 4. Let $D \geqslant 2$. There exists a $\gamma^* > 0$ such that the loss function Eq. (5) has the first-order phase transition between the trivial and feature learning phases at $\gamma = \gamma^*$.

However, an analytical expression for the critical point is not known. In physics, first-order phase transitions are accompanied by latent heat. Our theory implies that this heat is equivalent to the amount of random noise needed for the model parameters to escape from a local to the global minimum of a deep model. This perspective suggests that noise can play an indispensable role for successful learning.

While infinite-depth networks are not used in practice, they are important from a theoretical point of view [35] because they can be used for understanding a (very) deep network that often appears in deep learning practice. Our result shows that the limiting landscape has a zeroth-order phase transition at $\gamma = 0$. See Appendix B 6 for a detailed discussion. The zeroth-order phase transition does not occur in physics and is a unique feature of deep learning.

C. Linear origin theorem

In this section, we show that all nonlinear networks of arbitrary connection and structure can only be approximated to the first two nonvanishing orders by a linear model with the same connection patterns at the origin. This directly establishes the connection of our theory based on linear networks to nonlinear networks.

To start, let us define a nonlinear network. For concision, we ignore the stochasticity of the neurons. We consider the type of elementwise nonlinearity $h(x)$ that is differentiable and vanishes at the origin: $h(x) = h'(0)x + O(x^2)$. To make our theoretical statement clearer, we define an interpolation of this nonlinearity under consideration with the linear activation,

$$g_a(x) = (1 - a)h(x) + ax, \tag{9}$$

namely, the model is nonlinear at $a = 0$ and linear at $a = 1$.

A fully connected network with trainable weights $W^{(i)}$ can be written as

$$\sum_{i_D} W^{(D)}_{i_D i_{D-1}} g_a\left(... \sum_{i_1} W^{(1)}_{i_2 i_1} g_a\left(\sum_i W^{(0)}_{i_1 i_0} x_{i_0} \right) ... \right). \tag{10}$$

However, one can define a more general version to account for structured connectivities such as a convolutional neural network. To achieve this, we introduce a fixed masking matrix M such that M_{ij} is fixed to be 1 or 0 throughout training. Thus, any generic type of feedforward network can be defined as

$$f(x; a, \mathbf{W}) = \sum_{i_{D-1}} M^{(D)}_{i_D i_{D-1}} W^{(D)}_{i_D i_{D-1}} g_a\left(... \sum_{i_1} M^{(1)}_{i_2 i_1} \times W^{(1)}_{i_2 i_1} g_a\left(\sum_{i_0} M^{(0)}_{i_1 i_0} W^{(0)}_{i_1 i_0} x_{i_0} \right) ... \right), \tag{11}$$

where we have used the notation $\mathbf{W}$ to denote all the weight matrices combined. The following proposition shows that the leading order expansion of f is equivalent to its linear counterpart.

Proposition 1. For any x,

$$f(x; 0, \mathbf{W}) = [h'(0)]^D f(x; 1, \mathbf{W}) + o\left(\prod_i^D ||W^{(i)}|| \right). \tag{12}$$

Proof. By definition of g, we have that for the first layer,

$$g_0\left(\sum_i M^{(0)}_{i_1 i} W^{(0)}_{i_1 i} x_i \right) = h\left(\sum_i M^{(0)}_{i_1 i} W^{(0)}_{i_1 i} x_i \right) = h'(0) \sum_i M^{(0)}_{i_1 i} W^{(0)}_{i_1 i} x_i + o(||W^{(0)}||). \tag{13}$$

Since a similar relation holds for every layer, one can deduce the leading order expansion of f at the origin:

$$f(x; 0, \mathbf{W}) = \sum_{i_D, ..., i_1, i} M^{(D)}_{i_D i_{D-1}} W^{(D)}_{i_D i_{D-1}} h\big(... M^{(2)}_{i_2 i_1} W^{(2)}_{i_2 i_1} h\big(M^{(1)}_{i_1 i} W^{(1)}_{i_1 i} x_i \big) ... \big) = [h'(0)]^D f\left(x; 1, \mathbf{W} \right) + o(\prod_i ||W^{(i)}||). \tag{14}$$

This finishes the proof. ■

While the proof is a straightforward Taylor expansion, we note that the insight it implies is rather extraordinary: close

to the origin, any nonlinear model can be approximated by a linear model of the *same* architecture, up to a linear rescaling factor $[h'(0)]^D$. Also note that, in general, the origin is the only solution that satisfies this special property. An immediate corollary of this proposition is that any loss function, with weight decay, for a neural network with at least one hidden layer is approximated by the loss function of a linear network close to the origin:

$$\begin{aligned} L\gamma(\mathbf{W}) &= \sum_x \ell(f(x;0,\mathbf{W})) + \gamma||\mathbf{W}||^2 \\ &= \sum_x \ell\left(f(x;1,\mathbf{W}) + o(\prod_i^D ||W^{(i)}||)\right) + \gamma||\mathbf{W}||^2 \\ &= \sum_x \ell(f(x;1,\mathbf{W})) + \gamma||\mathbf{W}||^2 + o\left(\prod_i^D ||W^{(i)}||\right). \end{aligned} \tag{15}$$

Therefore, this explains why different types of models (convolutional or fully connected) under different loss functions, such as MSE or cross entropy, all exhibit a phenomenology similar to a deep linear model of similar depth in practice [36,37]. Because the term $\ell(f(x;1,\mathbf{W}))$ is often of order $O(\prod_i^D ||W^{(i)}||)$, when $D = 1$, the first nonvanishing order term of the loss function agrees with that of a linear model. When $D > 1$, the first two nonvanishing terms agree. We numerically demonstrate these phase transitions and related phenomenology in nonlinear networks in Appendix A 3.

IV. ALGORITHMIC IMPLICATIONS

In this section, we discuss the implication of our theory for understanding the phenomenology in deep learning and for designing algorithms to improve it.

A. Layer structure and collapses

One major phenomenon in deep learning that can be explained by the proposed theory is the phenomenon of collapses. The collapse problem refers to the case when the learned representation of a neural network spans a low-rank subspace of the entire available space. The extreme case of collapse happens when the learning completely fails and the learned representation becomes a constant. In the past, collapses in different scenarios are often treated differently. Our result, in contrast, provides a unified perspective on the posterior collapse in Bayesian deep learning [38–41], the neural collapse problem in supervised learning [42–45], and the dimensional collapse in contrastive learning [46,47]. For neural collapse, our result agrees with the recent works that identify weight decay as a main cause [44,45]. For Bayesian deep learning, Ref. [41] identified the cause of the posterior collapse in a two-layer VAE structure as the regularization of the mean of the latent variable z being too strong, which then causes a change in the stability of the Hessian matrix at the origin. Importantly, it is shown that the norm of the model obeys the square-root scaling, where the norm of the learned classifier scales as $\sqrt{a_0 - \beta}$ in the vicinity of a collapse, where a_0 is the data variance and β is the strength of the prior. More recently, the origin and its stability have also been identified as a crucial feature for the dimensional collapse in self-supervised learning [46]. Again, the authors showed the existence of the square-root scaling in the vicinity of the collapse. Here, the norm of the classifier scales as $\sqrt{a_0 - c_0}$, where c_0 is the strength of the data augmentation.

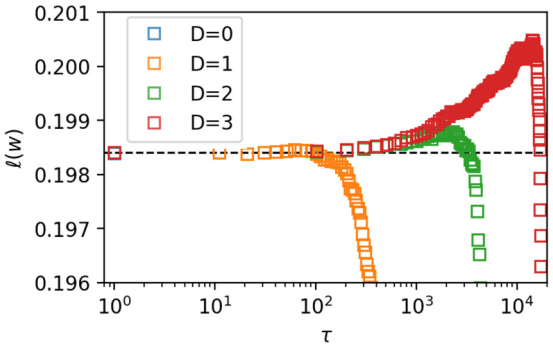

FIG. 3. Time evolution of the training loss when L is close to the initialized value (≈0.2). For $D = 1$, the loss decreases monotonically. For $D > 1$, in sharp contrast, the loss first increases slowly and then decreases precipitously, indicating a signature of escaping from a local minimum: the height of the peak may be interpreted as the latent heat of the phase transition since this is the energy barrier for the system to overcome to undergo the first-order phase transition.

See Figs. 2(c) and 2(d). In the setting of supervised learning, we consider ResNet18 [48], a modern large-scale convolutional neural network, trained on the CIFAR-10 data set [49], a standard image classification data set in machine learning. For all experiments, we train the model in a nontrivial phase to convergence and rescale the relevant set of parameters by a constant a. Figure 2(c) shows the landscape of a ResNet18 trained on the CIFAR-10 data set for different values of weight decay when we rescale all the parameters. We see that the landscape is qualitatively similar to a deep linear net with $D > 1$, as expected from our theory [50]. Figure 2(d) showcases a change in the landscape of a ResNet18 in self-supervised learning (SSL) when we rescale the output layer. Here, the strength of data augmentation plays the role of an effective regularization and the data variation plays the role of the data signal. The effective two-layer structure arises from the coupling of the last layer matrix with itself in SSL. We also perform simulations for generative models in Figs. 2(d) and 2(e). Here, the regularization effect is due to the KL divergence from the prior term, and the two-layer structure comes from the simultaneous use of an encoder and decoder. We see that the phase transition directly impacts the quality of the generated images [51] and that in all scenarios, the behaviors of deep learning models are similar to those of critical systems, exhibiting qualitative transitions in the landscape that cannot be understood in the linear response regime.

B. Sensitivity to the initial condition

Our result suggests that the learning of a deeper network is quite sensitive to the initialization schemes we use. In particular, for $D > 1$, some initialization schemes converge to the trivial solutions more easily, while others converge to the nontrivial solution more easily. Figure 4 plots the converged loss of a $D = 2$ model for two types of initialization: (a) large

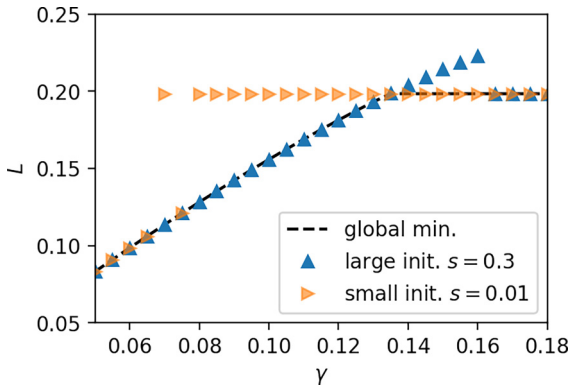

FIG. 4. Sensitivity of the obtained solution to the initialization of the model. We initialize the model around zero with the standard deviation s. The experiment shows that a large initialization variance ($s = 0.3$) favors the nontrivial solution over the trivial one, while a small initialization variance ($s = 0.01$) leads to the opposite tendency.

initialization, where the parameters are initialized around zero with the standard deviation $s = 0.3$ and (b) small initialization with $s = 0.01$. The value of s is thus equal to the expected norm of the model at initialization—a small s means that it is initialized closer to the trivial phase and a large s means that it is initialized closer to the learning phase. We see that over a wide range of γ, one of the initialization schemes gets stuck in a local minimum and does not converge to the global minimum. In light of the latent heat picture, the reason for the sensitivity to initial states is clear: one needs to inject additional energy for the system to leave the metastable state; otherwise, the system may be stuck for a very long time. See Fig. 3 for an illustration. The existing initialization methods are predominantly data dependent. However, our result (also see Ref. [52]) suggests that the size of the trivial minimum is data dependent, and our result thus highlights the importance of designing data-dependent initialization methods in deep learning.

C. Removing the trivial phase

We also explore our suggested fix to the trivial learning problem. Here, we regularize the model by $\gamma||w||_2^{D+2}$ rather than $\gamma||w||_2^2$. The training loss $\ell(b)$ and the model norm b are plotted in Fig. 5. We find that the trivial phase now completely disappears even if we go to very large γ. However, we note that this fix only removes the local maximum at zero, but zero remains a saddle point from which it takes the system a long time to escape.

V. CONCLUSION

The similarity between phase transitions in neural networks and statistical physics lends a great impetus to a more thorough investigation of deep learning through the lens of thermodynamics and statistical physics. Our theory also serves as a bridge between conventional machine learning and the statistical mechanics approaches to deep learning. One interesting future problem is to investigate whether we

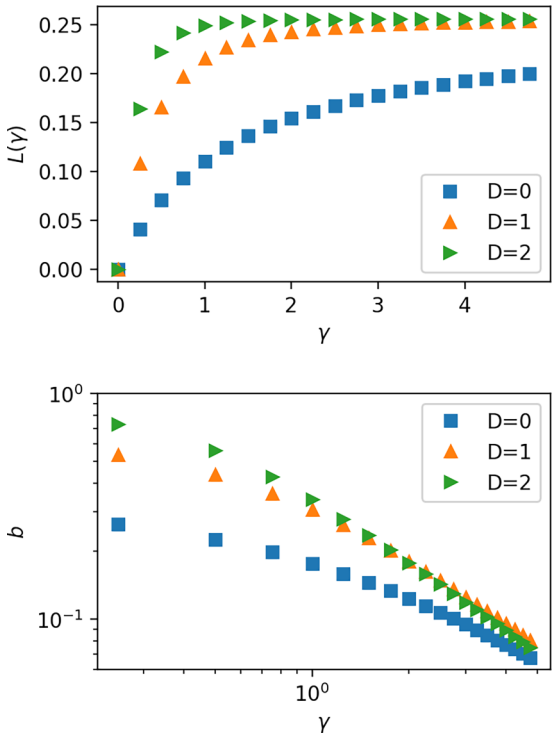

FIG. 5. Training loss $L(\gamma)$ (upper) and the model norm b (lower) when we train with a regularization term of the form $\gamma||w||^{D+2}$, which is a theoretically justified fix to the trivial learning problem. We see that the trivial phase disappears under this regularization.

can classify neural networks by symmetry and topological invariants instead of by the order of nonanalyticity.

ACKNOWLEDGMENT

This work was supported by KAKENHI Grant No. JP18H01145 from the Japan Society for the Promotion of Science.

APPENDIX A: ADDITIONAL EXPERIMENTS

1. Empirical validation of the theory

In Fig. 1, we show phase transitions in a linear net, where we validate the existence of the phase transitions discussed in the main text. In agreement with the theory, a depth-0 net has no phase transition, a depth-1 net has a second-order phase transition at approximately $\gamma = 0.45$, close to the theoretical value of $||\mathbb{E}[xy]||$, and a depth-2 net has a first-order phase transition at roughly $\gamma = 0.15$. The qualitative distinctions between networks of different depths are clearly seen in the data.

2. Experimental details

a. Supervised learning

We train a standard ResNet18 with roughly 10^7 parameters under the standard procedure, with a batch size of 256 for 100 epochs [53]. For the linear models, we use a hidden width of 32 without any bias term. The training proceeds with stochastic gradient descent with batch size 256 for 100

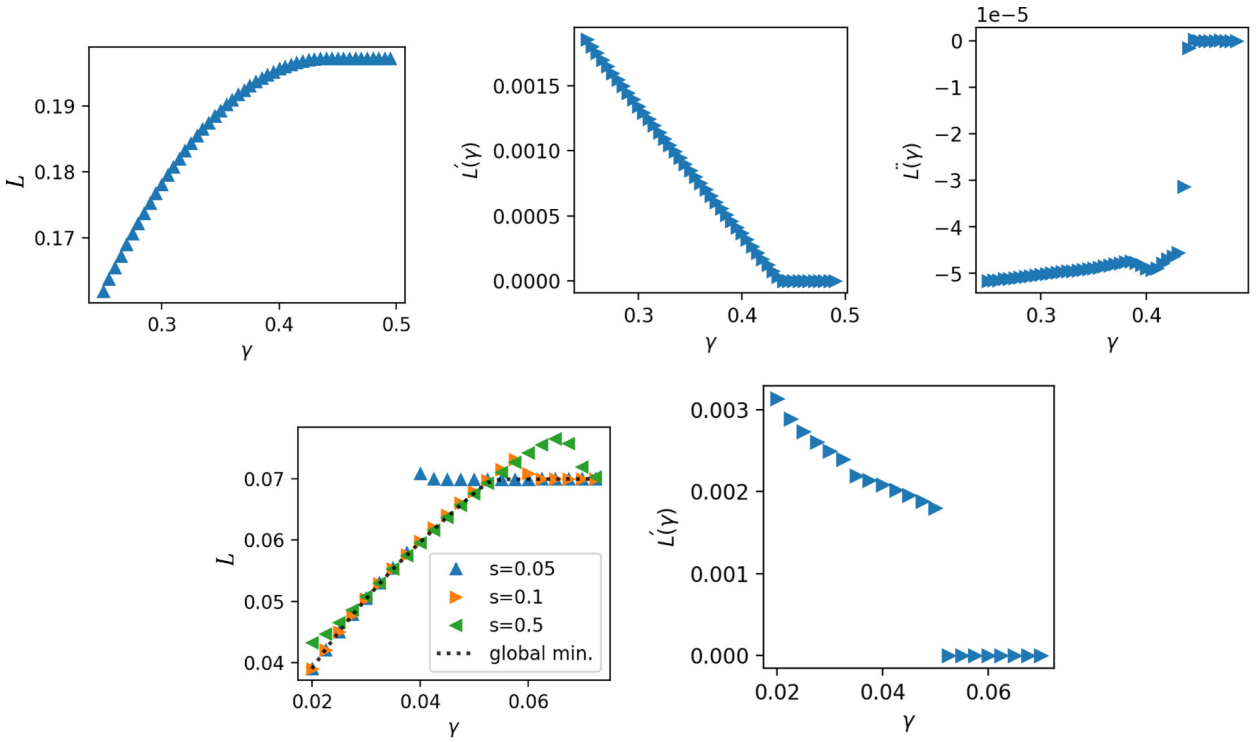

FIG. 6. Phase transition of a fully connected tanh network. Top: $D = 1$, which exhibits a second-order phase transition: the training loss $L(\gamma)$ (left), first derivative (middle), and the second derivative (right). Bottom: $D = 2$, which exhibits a first-order phase transition: the training loss $L(\gamma)$ (left) and first derivative $L'(\gamma)$ (middle). For $D = 2$, we initialize the model with three initialization at different scales and use the minimum of the respective loss values as an empirical estimate of the actual global minimum.

epochs with a momentum of 0.9. The learning rate is 0.002, chosen as the best learning rate from a grid search over $[0.001, 0.002, \ldots, 0.01]$.

b. Self-supervised learning

We train a Resnet18 on CIFAR10 with the SimCLR loss with a normalization of the last layer output [54] and with a weight decay strength of 10^{-3}. The training proceeds until the converged weights W^* are obtained. The representation has a dimension of 128. We rescale the weight matrix of the last layer W^*_{last} by a factor of a and compute the loss as a function of a.

c. Variational autoencoder

The experiment is performed on the MNIST data set. We use two-layer fully connected neural networks for the encoder and decoder with the ReLU activation functions and hidden dimension d_h. The dimension of the hidden layer is $d_h = 2048$. The model is optimized by Adam with a learning rate of 10^{-3}. The reported results are the converged values.

3. Nonlinear networks

We expect our theory to also apply to deep nonlinear networks that can be locally approximated by a linear net at the origin, e.g., a network with tanh activations. As shown in Fig. 6, the data shows that a tanh net also features a second-order phase transition for $D = 1$ and a first-order phase transition for $D = 2$.

One notable exception that our theory may not apply is the networks with the ReLU activation because these networks are not differentiable at the origin (i.e., in the trivial phase). However, there are smoother (and empirically better) alternatives to ReLU, such as the swish activation function, to which the present theory should also be relevant.

APPENDIX B: DERIVATION AND PROOF

Here, we present detailed and rigorous derivations of the main equations presented in the main text. We first note that the main loss in Eq. (3) can be written as

$$\mathbb{E}_x \mathbb{E}_{\epsilon^{(1)},\epsilon^{(2)},\ldots,\epsilon^{(D)}} \left(\sum_{i,i_1,i_2,\ldots,i_D}^{d_0,d_0,d_0,\ldots d_0} \prod_{j=1}^{D} U_{i_D} \epsilon^{(D)}_{i_D} \ldots \epsilon^{(2)}_{i_2} W^{(2)}_{i_2 i_1} \right.$$
$$\left. \times \epsilon^{(1)}_{i_1} W^{(1)}_{i_1 i} x_i - y \right)^2 + \gamma ||U||_2^2 + \sum_{i=1}^{D} \gamma ||W^{(i)}||_F^2. \quad \text{(B1)}$$

1. Linear regression

Here, the loss function becomes

$$\ell(W) = \mathbb{E}_x \left(\sum_i W_i x_i - y \right)^2 + \gamma ||W||^2, \quad \text{(B2)}$$

from which we prove the theorem.

Proof of Theorem 1. The global minimum of Eq. (B2) is

$$W_* = (A_0 + \gamma I)^{-1} E[xy]. \quad \text{(B3)}$$

The loss of the global minimum is thus

$$\begin{aligned} L &= \mathbb{E}_x\left(\sum_i W_i x_i - y\right)^2 + \gamma||W||^2 \\ &= W^T A_0 W - 2W^T\mathbb{E}[xy] + \mathbb{E}[y^2] + \gamma||W||^2 \\ &= \mathbb{E}[xy]^T \frac{A_0}{(A_0+\gamma I)^2}\mathbb{E}[xy] - 2\mathbb{E}[xy]^T\frac{1}{A_0+\gamma I}\mathbb{E}[xy] \\ &\quad + \mathbb{E}[y^2] + \gamma\mathbb{E}[xy]^T\frac{1}{(A_0+\gamma I)^2}\mathbb{E}[xy] \\ &= -\mathbb{E}[xy]^T(A_0+\gamma I)^{-1}\mathbb{E}[xy] + \mathbb{E}[y^2], \end{aligned} \tag{B4}$$

which is infinitely differentiable for any $\gamma \in (0,\infty)$ (note that A_0 is always positive semidefinite by definition). ■

2. Finite depth cannot have zeroth-order phase transitions

As discussed in the main text, the theorem shows that for a finite depth, $L(\gamma)$ must be continuous in γ.

Proof of Theorem 2. For any fixed and bounded w, $\ell(w,\gamma)$ is continuous in γ. Moreover, $\ell(w,\gamma)$ is a monotonically increasing function of γ. This implies that $L(\gamma)$ is also an increasing function of γ (but may not be strictly increasing).

We now prove by contradiction. We first show that $L(\gamma)$ is left-continuous. Suppose that for some D, $L(\gamma)$ is not left-continuous in γ at some γ^*. By definition, we have

$$L(\gamma^*-\epsilon) = \min_w \ell(w,\gamma^*-\epsilon) := \ell(w',\gamma^*-\epsilon), \tag{B5}$$

where w' is one of the (potentially many) global minima of $L(\gamma^*-\epsilon)$. Since $L(\gamma)$ is not left-continuous by assumption, there exists $\delta>0$ such that for any $\epsilon>0$,

$$L(\gamma^*-\epsilon) < L(\gamma^*) - \delta, \tag{B6}$$

which implies that

$$\ell(w',\gamma^*-\epsilon) = L(\gamma^*-\epsilon) < L(\gamma^*) - \delta \leqslant \ell(w',\gamma^*) - \delta, \tag{B7}$$

namely, the left discontinuity implies that for all $\epsilon>0$:

$$\ell(w',\gamma^*-\epsilon) \leqslant \ell(w',\gamma^*) - \delta. \tag{B8}$$

However, by definition of $\ell(w,\gamma)$, we have

$$\ell(w,\gamma) - \ell(w,\gamma-\epsilon) = \epsilon||w||^2. \tag{B9}$$

Thus, by choosing $\epsilon < \delta/||w||^2$, the relation in Eq. (B7) is violated. Thus, $L(\gamma)$ must be left-continuous.

In a similar manner, we can prove that L is right-continuous. Suppose that for some D, $L(\gamma)$ is not right-continuous in γ at some γ^*. Let $\gamma>0$. By definition, we have

$$L(\gamma^*+\epsilon) = \min_w \ell(w,\gamma^*+\epsilon) := \ell(w',\gamma^*+\epsilon), \tag{B10}$$

where w' is one of the (potentially many) global minima of $L(\gamma^*+\epsilon)$. Since $L(\gamma)$ is not right-continuous by assumption, there exists $\delta>0$ such that for any $\epsilon>0$,

$$L(\gamma^*+\epsilon) > L(\gamma^*) + \delta, \tag{B11}$$

which implies that

$$\ell(w',\gamma^*+\epsilon) = L(\gamma^*+\epsilon) > L(\gamma^*) + \delta \geqslant \ell(w',\gamma^*) + \delta, \tag{B12}$$

namely, the right discontinuity implies that for all $\epsilon>0$,

$$\ell(w',\gamma^*+\epsilon) \geqslant \ell(w',\gamma^*) + \delta. \tag{B13}$$

However, by definition of $\ell(w,\gamma)$, we have

$$\ell(w,\gamma+\epsilon) - \ell(w,\gamma) = \epsilon||w||^2. \tag{B14}$$

Thus, by choosing $\epsilon < \delta/||w||^2$, the relation in Eq. (B12) is violated. Thus, $L(\gamma)$ must be right-continuous.

Therefore, $L(\gamma)$ is continuous for all $\gamma>0$. By definition, this means that there is no zeroth-order phase transition in γ for L. Additionally, note that the above proof does not require $\gamma\neq 0$, and so we have also shown that $L(\gamma)$ is right-continuous at $\gamma=0$. ■

3. Order parameter and the effective loss

Theorem 5. Let $b = s||U||/d_0$, and let

$$\begin{aligned} \bar{\ell}(b,\gamma) := &-\sum_i \frac{d_0^{2D}b^{2D}\mathbb{E}[x'y]_i^2}{d_0^D(\sigma^2+d_0)^D a_i b^{2D}+\gamma} \\ &+ \mathbb{E}_x[y^2] + \gamma D d_0^2 b^2. \end{aligned} \tag{B15}$$

Then, b is an order parameter of Eq. (3) for the effective loss $\bar{\ell}$.

Proof. By Theorem 3 of Ref. [52], any global minimum of Eq. (3) is given by the following set of equations:

$$\begin{aligned} U &= \sqrt{d_0}|b|\mathbf{r}_D \\ W^{(i)} &= |b|\mathbf{r}_i\mathbf{r}_{i-1}^T \\ W^{(1)} &= \mathbf{r}_1\mathbb{E}[xy]^T d_0^{D-\frac{1}{2}}|b|^D\left[d_0^D(\sigma^2+d_0)^D b^{2D}A_0+\gamma\right]^{-1}, \end{aligned} \tag{B16}$$

where $\mathbf{r}_i = (\pm 1,\ldots,\pm 1)$ is an arbitrary vertex of a d_i-dimensional hypercube for all i. Therefore, the global minimum must lie on a one-dimensional space indexed by $b\in[0,\infty)$. Let $f(x)$ specify the model as

$$f(x) := \sum_{i,i_1,i_2,\ldots,i_D}^{d,d_1,d_2,\ldots d_D} U_{i_D}\epsilon_{i_D}^{(D)}\ldots\epsilon_{i_2}^{(2)}W_{i_2 i_1}^{(2)}\epsilon_{i_1}^{(1)}W_{i_1 i}^{(1)}x, \tag{B17}$$

and let η denote the set of all random noises ϵ_i.

Substituting Eq. (B16) in Eq. (3), one finds that within this subspace, the loss function can be written as

$$\begin{aligned} \ell(w,\gamma) &= \mathbb{E}_x\mathbb{E}_\eta(f(x)-y)^2 + L_2\text{ reg} \\ &= \mathbb{E}_{x,\eta}[f(x)^2] - 2\mathbb{E}_{x,\eta}[yf(x)] + \mathbb{E}_x[y^2] + L_2\text{ reg} \\ &= \sum_i \frac{d_0^{3D}(\sigma^2+d_0)^D b^{4D}a_i\mathbb{E}[x'y]_i^2}{\left[d_0^D(\sigma^2+d_0)^D a_i b^{2D}+\gamma\right]^2} \\ &\quad - 2\sum_i \frac{d_0^{2D}b^{2D}\mathbb{E}[x'y]_i^2}{d_0^D(\sigma^2+d_0)^D a_i b^{2D}+\gamma} + \mathbb{E}_x[y^2] + L_2\text{ reg}, \end{aligned} \tag{B18}$$

where the L_2 reg term is given by

$$L_2 \text{ reg} = \gamma D d_0^2 b^2 + \gamma \sum_i \frac{d_0^{2D} b^{2D} \mathbb{E}[x'y]_i^2}{\left[d_0^D (\sigma^2 + d_0)^D b^{2D} a_i + \gamma\right]^2}. \tag{B19}$$

Combining terms, we can simplify the expression for the loss function to be

$$-\sum_i \frac{d_0^{2D} b^{2D} \mathbb{E}[x'y]_i^2}{\left[d_0^D (\sigma^2 + d_0)^D a_i b^{2D} + \gamma\right]} + \mathbb{E}_x[y^2] + \gamma D d_0^2 b^2. \tag{B20}$$

We can now define the effective loss by

$$\bar{\ell}(b, \gamma) := -\sum_i \frac{d_0^{2D} b^{2D} \mathbb{E}[x'y]_i^2}{\left[d_0^D (\sigma^2 + d_0)^D a_i b^{2D} + \gamma\right]} + \mathbb{E}_x[y^2] + \gamma D d_0^2 b^2. \tag{B21}$$

Then, the above argument shows that for all γ,

$$\min_b \bar{\ell}(b, \gamma) = \min_w \ell(w, \gamma). \tag{B22}$$

By Definition 2 in the main text, b is an order parameter of ℓ with respect to the effective loss $\bar{\ell}(b, \gamma)$. This completes the proof. ■

4. $D = 1$

In this section, we prove Theorem 3

We first prove a lemma that will simplify the proof of Theorem 3 significantly.

Lemma 1. If $L(\gamma)$ is differentiable, then for at least one of the global minima b_*,

$$\frac{d}{d\gamma} L(\gamma) = \sum_i \frac{d_0^{2D} b_*^{2D} \mathbb{E}[x'y]_i^2}{[d_0^D (\sigma^2 + d_0)^D a_i b_*^{2D} + \gamma]^2} + D d_0^2 b_*^2 \geqslant 0. \tag{B23}$$

Proof of Lemma 1. Because L is differentiable in γ, there exist at least one global minima b^* such that

$$\frac{d}{d\gamma} L(\gamma) = \frac{d}{d\gamma} \bar{\ell}(b^*(\gamma), \gamma) \tag{B24}$$

$$= \frac{\partial}{\partial b^*} \bar{\ell}(b^*, \gamma) \frac{\partial b^*}{\partial \gamma} + \frac{\partial}{\partial \gamma} \bar{\ell}(b^*, \gamma) \tag{B25}$$

$$= \frac{\partial}{\partial \gamma} \bar{\ell}(b^*, \gamma) \tag{B26}$$

$$= \sum_i \frac{d_0^{2D} b_*^{2D} \mathbb{E}[x'y]_i^2}{\left[d_0^D (\sigma^2 + d_0)^D a_i b_*^{2D} + \gamma\right]^2} + D d_0^2 b_*^2 \geqslant 0, \tag{B27}$$

where we have used the optimality condition $\frac{\partial}{\partial b^*} \bar{\ell}(b^*(\gamma), \gamma) = 0$ in the second equality. ■

Proof of Theorem 3. By Definition 1, it suffices only to prove the existence of phase transitions on the effective loss. For $D = 1$, the effective loss is

$$\bar{\ell}(b, \gamma) = -d_1 b^2 E[xy]^T [b^2(\sigma^2 + d_1)A + \gamma I]^{-1} E[xy] + E[y^2] + \gamma d_1 b^2. \tag{B28}$$

By Theorem 1 of Ref. [52], the phase transition, if it exists, must occur precisely at $\gamma = ||\mathbb{E}[xy]||$. To prove that there is a second-order phase transition at $\gamma = ||\mathbb{E}[xy]||$, we must check both its first and second derivatives.

When $\gamma \to ||E[xy]||$ from the right, all derivatives of $L(\gamma)$ vanish because the loss is identically equal to $\mathbb{E}[y^2]$. We now consider the derivative of L when $\gamma \to ||E[xy]||$ from the left. We first need to find the minimizer of Eq. (B28). Because Eq. (B28) is differentiable, its derivative in b must be equal to 0 at the global minimum:

$$-2\gamma d_1 b \mathbb{E}[xy]^T [b^2(\sigma^2 + d_1)^2 A + \gamma I]^{-2} \mathbb{E}[xy] + 2\gamma d_1 b = 0. \tag{B29}$$

Finding the minimizer b is thus equivalent to finding the real roots of a high-order polynomial in b. When $\gamma \geqslant ||\mathbb{E}[xy]||$, the solution is unique [52],

$$b_0^2 = 0, \tag{B30}$$

where we label the solution with the subscript 0 to emphasize that this solution is also the zeroth-order term of the solution in a perturbatively small neighborhood of $\gamma = ||E[xy]||$. From this point, we define a shifted regularization strength: $\Delta := \gamma - ||\mathbb{E}[xy]||$. When $\Delta < 0$, the condition Eq. (B29) simplifies to

$$\mathbb{E}[xy]^T [b^2(\sigma^2 + d_1)A + \gamma I]^{-2} \mathbb{E}[xy] = 1. \tag{B31}$$

Because the polynomial is not singular in Δ, one can Taylor expand the (squared) solution b^2 in Δ:

$$b(\gamma)^2 = \beta_0 + \beta_1 \Delta + O(\Delta^2). \tag{B32}$$

We first substitute Eq. (B32) in (B29) to find [55]

$$\beta_0 = 0. \tag{B33}$$

One can then again substitute Eq. (B32) in Eq. (B29) to find β_1. To the first order in b^2, Eq. (B29) reads

$$\frac{1}{\gamma^2} ||\mathbb{E}[xy]||^2 - 2b^2 \frac{(\sigma^2 + d_1)}{\gamma^3} ||\mathbb{E}[xy]||_{A_0}^2 = 1 \tag{B34}$$

$$\iff -2\beta_1 \Delta \frac{(\sigma^2 + d_1)}{\gamma^3} ||\mathbb{E}[xy]||_{A_0}^2 = 2\frac{\Delta}{||\mathbb{E}[xy]||} \tag{B35}$$

$$\iff \beta_1 = -\frac{1}{(\sigma^2 + d_1)} \frac{||\mathbb{E}[xy]||^2}{||\mathbb{E}[xy]||_{A_0}^2}. \tag{B36}$$

Substituting this first-order solution to Lemma 1, we obtain

$$\frac{d}{d\gamma} L(\gamma)|_{\gamma = ||E[xy]||_-} \sim b_*^2 = 0 = \frac{d}{d\gamma} L(\gamma)|_{\gamma = ||E[xy]||_+}. \tag{B37}$$

Thus, the first-order derivative of $L(\gamma)$ is continuous at the phase transition point.

We now find the second-order derivative of $L(\gamma)$. To achieve this, we also need to find the second-order term of b^2 in γ. We expand b^2 as

$$b(\gamma)^2 = 0 + \beta_1 \Delta + \beta_2 \Delta^2 + O(\Delta^3). \tag{B38}$$

To the second order in b^2, Eq. (B29) reads

$$\frac{1}{\gamma^2}||\mathbb{E}[xy]||^2 - 2b^2\frac{(\sigma^2+d_1)}{\gamma^3}||\mathbb{E}[xy]||^2_{A_0} + 3b^4\frac{(\sigma^2+d_1)^2}{\gamma^4}||\mathbb{E}[xy]||^2_{A_0^2} = 1 \tag{B39}$$

$$\iff \gamma^2||\mathbb{E}[xy]||^2 - 2b^2(\sigma^2+d_1)\gamma||\mathbb{E}[xy]||^2_{A_0} + 3b^4(\sigma^2+d_1)^2||\mathbb{E}[xy]||^2_{A_0^2} = \gamma^4 \tag{B40}$$

$$\iff \Delta^2E_0^2 - 2\beta_2\Delta^2(\sigma^2+d_1)E_0E_1^2 - 2\beta_1\Delta^2(\sigma^2+d_1)E_1^2 + 3\beta_1^2\Delta^2(\sigma^2+d_1)^2E_2^2 = 6E_0^2\Delta^2 \tag{B41}$$

$$\iff \beta_2 = \frac{3\beta_1^2(\sigma^2+d_1)^2E_2^2 - 5E_0^2 - 2\beta_1(\sigma^2+d_1)E_1^2}{2(\sigma^2+d_1)E_0E_1^2}, \tag{B42}$$

where, from the third line, we have used the shorthand notations $E_0 := ||\mathbb{E}[xy]||$, $E_1 := ||\mathbb{E}[xy]||_{A_0}$, and $E_2 := ||\mathbb{E}[xy]||_{A_0^2}$. Substituting in β_1, we obtain

$$\beta_2 = \frac{3E_0\left(E_2^2 - E_1^2\right)}{2(\sigma^2+d_1)E_1^4}. \tag{B43}$$

This allows us to find the second derivative of $L(\gamma)$. Substituting β_1 and β_2 into Eq. (B28) and expanding the result up to the second order in Δ, we obtain

$$L(\gamma) = -d_1b^2E[xy]^T[b^2(\sigma^2+d_1)A + \gamma I]^{-1}E[xy] + E[y^2] + \gamma d_1b^2 \tag{B44}$$

$$= -d_1(\beta_1\Delta + \beta_2\Delta^2)\mathbb{E}[xy]^T[(\beta_1\Delta + \beta_2\Delta^2)(\sigma^2+d_1)A_0 + \gamma I]^{-1}\mathbb{E}[xy] + \gamma d_1(\beta_1\Delta + \beta_2\Delta). \tag{B45}$$

At the critical point,

$$\begin{aligned}
&\frac{d^2}{d\gamma^2}L(\gamma)|_{\gamma||\mathbb{E}[xy]||_-} \\
&= -d_1\beta_2E_0 + d_1\beta_1^2(\sigma^2+d_1)\frac{E_1^2}{E_0^2} + d_1\beta_1 + d_1\beta_1 + d_1\beta_2E_0 \\
&= 2d_1\beta_1 + d_1\beta_1^2(\sigma^2+d_1)\frac{E_1^2}{E_0^2} \\
&= d_1\beta_1 \\
&= -\frac{d_1}{\sigma^2+d_1}\frac{||\mathbb{E}[xy]||^2}{||\mathbb{E}[xy]||^2_{A_0}}.
\end{aligned} \tag{B46}$$

Notably, the second left derivative of L is only dependent on β_1 and not on β_2:

$$\frac{d^2}{d\gamma^2}L(\gamma)|_{\gamma=||\mathbb{E}[xy]||_-} = -\frac{d_1}{\sigma^2+d_1}\frac{||\mathbb{E}[xy]||^2}{||\mathbb{E}[xy]||^2_{A_0}} < 0. \tag{B47}$$

Thus, the second derivative of $L(\gamma)$ is discontinuous at $\gamma = ||\mathbb{E}[xy]||$. This completes the proof. ■

Remark. Note that the above proof suggests that close to the critical point, $b \sim \sqrt{\Delta}$, in agreement with the Landau theory.

5. $D > 1$

Proof of Theorem 4. It suffices to show that $\frac{d}{d\gamma}L(\gamma)$ is not continuous. We prove the statement by contradiction. Suppose that $\frac{d}{d\gamma}L(\gamma)$ is everywhere continuous on $\gamma \in (0,\infty)$. Then, by Lemma 1, one can find the derivative for at least one of the global minima b^*:

$$\frac{d}{d\gamma}L(\gamma) = \sum_i \frac{d_0^{2D}b_*^{2D}\mathbb{E}[x'y]_i^2}{\left[d_0^D(\sigma^2+d_0)^Da_ib_*^{2D} + \gamma\right]^2} + \gamma Dd_0^2b_*^2 \geqslant 0. \tag{B48}$$

Both terms on the right-hand side are nonnegative, and so one necessary condition for $\frac{d}{d\gamma}L(\gamma)$ to be continuous is that both of these two terms are continuous in γ.

In particular, one necessary condition is that $\gamma Dd_0^2b_*^2$ is continuous in γ. By Proposition 3 of Ref. [52], there exist constants c_0, c_1 such that $0 < c_0 \leqslant c_1$, and

$$\begin{aligned}
&b_* = 0 \quad \text{if } \gamma < c_0 \\
&b_* > 0, \quad \text{if } \gamma > c_1.
\end{aligned} \tag{B49}$$

Additionally, if $b_* > 0$, b_* must be lower-bounded by some nonzero value [52]:

$$b_* \geqslant \frac{1}{d_0}\left(\frac{\gamma}{||\mathbb{E}[xy]||}\right)^{\frac{1}{D-1}} > \frac{1}{d_0}\left(\frac{c_1}{||\mathbb{E}[xy]||}\right)^{\frac{1}{D-1}} > 0. \tag{B50}$$

Therefore, for any $D > 1$, $b_*(\gamma)$ must have a discontinuous jump from 0 to a value larger than $\frac{1}{d_0}(\frac{c_0}{||\mathbb{E}[xy]||})^{\frac{1}{D-1}}$, and cannot be continuous. This, in turn, implies that $\frac{d}{d\gamma}L(\gamma)$ jumps from zero to a nonzero value and cannot be continuous. This completes the proof. ■

6. $D \to \infty$

The following theorem formally studies the case of $D \to \infty$.

Theorem 6. Let $L^{(D)}(\gamma)$ denote the loss function for a fixed depth D as a function of γ. Then, for $\gamma \in [0,\infty)$ and some constant r,

$$L^{(D)}(\gamma) \to \begin{cases} r & \text{if } \gamma = 0 \\ \mathbb{E}[y^2] & \text{otherwise.} \end{cases} \tag{B51}$$

Proof. It suffices to show that a nonzero global minimum cannot exist at a sufficiently large D, when one fixes γ. By

Proposition 3 of Ref. [52], when $\gamma > 0$, any nonzero global minimum must obey the following two inequalities:

$$\frac{1}{d_0}\left[\frac{\gamma}{||\mathbb{E}[xy]||}\right]^{\frac{1}{D-1}} \leqslant b^* \leqslant \left[\frac{||\mathbb{E}[xy]||}{d_0(\sigma^2+d_0)^D a_{\max}}\right]^{\frac{1}{D+1}}, \tag{B52}$$

where $a_{\max}$ is the largest eigenvalue of A_0. In the limit $D \to \infty$, the lower bound becomes

$$\frac{1}{d_0}\left[\frac{\gamma}{||\mathbb{E}[xy]||}\right]^{\frac{1}{D-1}} \to \frac{1}{d_0}. \tag{B53}$$

The upper bound becomes

$$\left[\frac{||\mathbb{E}[xy]||}{d_0(\sigma^2+d_0)^D a_{\max}}\right]^{\frac{1}{D+1}} \to \frac{1}{\sigma^2+d_0}. \tag{B54}$$

But for any $\sigma^2 > 0$, $\frac{1}{d_0} < \frac{1}{\sigma^2+d_0}$. Thus, the set of such b^* is empty.

On the other hand, when $\gamma = 0$, the global minimizer has been found in Ref. [56] and is nonzero, which implies that $L(0) < \mathbb{E}[y^2]$. This means that $L(\gamma)$ is not continuous at 0. This completes the proof. ■

[1] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, Proc. Natl. Acad. Sci. USA **79**, 2554 (1982).

[2] M. Biehl, E. Schlösser, and M. Ahr, Phase transitions in soft-committee machines, Europhys. Lett. **44**, 261 (1998).

[3] M. Ahr, M. Biehl, and R. Urbanczik, Statistical physics and practical training of soft-committee machines, Eur. Phys. J. B **10**, 583 (1999).

[4] T. L. H. Watkin, A. Rau, and M. Biehl, The statistical mechanics of learning a rule, Rev. Mod. Phys. **65**, 499 (1993).

[5] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, The loss surfaces of multilayer networks, in *Artificial Intelligence and Statistics* (PMLR, 2015), pp. 192–204.

[6] C. H. Martin and M. W. Mahoney, Rethinking generalization requires revisiting old ideas: Statistical mechanics approaches and complex learning behavior, arXiv:1710.09553.

[7] Y. Bahri, J. Kadmon, J. Pennington, S. S. Schoenholz, J. Sohl-Dickstein, and S. Ganguli, Statistical mechanics of deep learning, Annu. Rev. Condens. Matter Phys. **11**, 501 (2020).

[8] L. Dabelow and M. Ueda, Three learning stages and accuracy–efficiency tradeoff of restricted Boltzmann machines, Nat. Commun. **13**, 5474 (2022).

[9] S. Goldt and U. Seifert, Stochastic thermodynamics of learning, Phys. Rev. Lett. **118**, 010601 (2017).

[10] N. P. Baskerville, J. P. Keating, F. Mezzadri, and J. Najnudel, The loss surfaces of neural networks with general activation functions, J. Stat. Mech. (2021) 064001.

[11] C. Baldassi, C. Lauditi, E. M. Malatesta, G. Perugini, and R. Zecchina, Unveiling the structure of wide flat minima in neural networks, Phys. Rev. Lett. **127**, 278301 (2021).

[12] C. Baldassi, C. Lauditi, E. M. Malatesta, R. Pacelli, G. Perugini, and R. Zecchina, Learning through atypical phase transitions in overparameterized neural networks, Phys. Rev. E **106**, 014116 (2022).

[13] A. Krogh and J. A. Hertz, Generalization in a linear perceptron in the presence of noise, J. Phys. A: Math. Gen. **25**, 1135 (1992).

[14] A. Krogh and J. A. Hertz, A simple weight decay can improve generalization, Adv. Neural Inform. Process. Syst. **4**, 950 (1992).

[15] D. Haussler, M. Kearns, H. S. Seung, and N. Tishby, Rigorous learning curve bounds from statistical mechanics, Mach. Learn. **25**, 195 (1996).

[16] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, Surprises in high-dimensional ridgeless least squares interpolation, arXiv:1903.08560.

[17] Z. Liao, R. Couillet, and M. W. Mahoney, A random matrix analysis of random fourier features: beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent, J. Stat. Mech. (2021) 124006.

[18] M. Belkin, D. Hsu, and J. Xu, Two models of double descent for weak features, SIAM J. Math. Data Sci. **2**, 1167 (2020).

[19] H. S. Seung, H. Sompolinsky, and N. Tishby, Statistical mechanics of learning from examples, Phys. Rev. A **45**, 6056 (1992).

[20] H. Cui, F. Krzakala, and L. Zdeborová, Optimal learning of deep random networks of extensive-width, PMLR **202**, 6468 (2023).

[21] V. Vapnik, *Estimation of Dependences Based on Empirical Data* (Springer Science & Business Media, New York, 2006).

[22] G. M. Benedek and A. Itai, Learnability with respect to fixed distributions, Theor. Comput. Sci. **86**, 377 (1991).

[23] Q. Li and H. Sompolinsky, Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization, Phys. Rev. X **11**, 031059 (2021).

[24] We restrict to one-dimensional order parameters in this paper. One can extend the theory to a multidimensional case as in the Landau–de Gennes theory.

[25] K. Kawaguchi, Deep learning without poor local minima, Adv. Neural Inform. Process. Syst. **29**, 586 (2016).

[26] M. Hardt and T. Ma, Identity matters in deep learning, arXiv:1611.04231.

[27] T. Laurent and J. Brecht, Deep linear networks with arbitrary loss: All local minima are global, in *International Conference on Machine Learning* (PMLR, 2018), pp. 2902–2907.

[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. **15**, 1929 (2014).

[29] Note that it does not matter from which element we are extracting the sign; the sign of any element gives the same result. This is because the global minima of the loss function have a local discrete sign flip symmetry. This comes from the arbitrariness of the **r** vector in Theorem 2 of Ref. [52].

[30] See Appendix B.

[31] See Appendix A 2.

[32] One practical implication is that the L_2 regularization may be too strong for deep learning because it creates a trivial phase. Our result also suggests a way to avoid the trivial phase. Instead of regularizing by $\gamma||w||_2^2$, one might consider $\gamma||w||_2^{2D+1}$,

which is the lowest-order regularization that does not lead to a trivial phase. The effectiveness of this suggested method is confirmed in Appendix A 3.

[33] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, beta-VAE: Learning basic visual concepts with a constrained variational framework, in *International Conference on Learning Representations* (PMLR, 2016).

[34] When the two layers have different regularization strengths γ_u and γ_w, one can show that the phase transition occurs precisely at $\sqrt{\gamma_u \gamma_w} = ||\mathbb{E}[xy]||$.

[35] S. Sonoda and N. Murata, Transport analysis of infinitely deep neural network, J. Mach. Learn. Res. **20**, 31 (2019).

[36] J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington, Wide neural networks of any depth evolve as linear models under gradient descent, J. Stat. Mech. (2020) 124002.

[37] D. Kalimeris, G. Kaplun, P. Nakkiran, B. Edelman, T. Yang, B. Barak, and H. Zhang, SGD on neural networks learns functions of increasing complexity, in *Advances in Neural Information Processing Systems 32* (PMLR, 2019).

[38] B. Dai and D. Wipf, Diagnosing and enhancing VAE models, arXiv:1903.05789.

[39] A. Alemi, B. Poole, I. Fischer, J. Dillon, R. A. Saurous, and K. Murphy, Fixing a broken ELBO, PMLR, 159 (2018).

[40] J. Lucas, G. Tucker, R. B. Grosse, and M. Norouzi, Don't blame the ELBO! A linear VAE perspective on posterior collapse, in *Advances in Neural Information Processing Systems 32* (PMLR, 2019).

[41] Z. Wang and L. Ziyin, Posterior collapse of a linear latent variable model, arXiv:2205.04009.

[42] V. Papyan, X. Y. Han, and D. L. Donoho, Prevalence of neural collapse during the terminal phase of deep learning training, Proc. Natl. Acad. Sci. USA **117**, 24652 (2020).

[43] T. Galanti, A. György, and M. Hutter, On the role of neural collapse in transfer learning, arXiv:2112.15121.

[44] A. Rangamani and A. Banburski-Fahey, Neural collapse in deep homogeneous classifiers and the role of weight decay, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, Piscataway, NJ, 2022), pp. 4243–4247.

[45] A. Rangamani, M. Xu, A. Banburski, Q. Liao, and T. Poggio, Dynamics and neural collapse in deep classifiers trained with the square loss, CBMM Memos, 117 (2021).

[46] L. Ziyin, E. S. Lubana, M. Ueda, and H. Tanaka, What shapes the loss landscape of self-supervised learning? arXiv:2210.00638.

[47] Y. Tian, Understanding deep contrastive learning via coordinate-wise optimization, arXiv:2201.12680.

[48] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2016), pp. 770–778.

[49] A. Krizhevsky and G. Hinton, Learning multiple layers of features from tiny images, (2009).

[50] Using ResNet, one needs to change the dimension of the hidden layer after every bottleneck, and a learnable linear transformation is applied here. Thus, the "effective depth" of a ResNet would be roughly between the number of its bottlenecks and its total number of blocks. For example, a ResNet18 applied to CIFAR10 often has five bottlenecks and 18 layers in total. We thus expect it to have qualitatively similar behavior to a deep linear net with a depth in between.

[51] For all experiments, the experimental details are provided in Appendix A.

[52] L. Ziyin, B. Li, and X. Meng, Exact solutions of a deep linear network, arXiv:2202.04777.

[53] Specifically, we use the implementation and training procedure of https://github.com/kuangliu/pytorch-cifar, with standard augmentations such as random crop, etc.

[54] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, A simple framework for contrastive learning of visual representations, in *International Conference on Machine Learning* (PMLR, 2020), pp. 1597–1607.

[55] Note that, alternatively, $\beta_0 = 0$ is implied by the no-zeroth-order transition theorem.

[56] P. Mianjy and R. Arora, On dropout and nuclear norm regularization, in *International Conference on Machine Learning* (PMLR, 2019), pp. 4575–4584.