Exact Phase Transitions in Deep Learning

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

[1] *Department of Physics, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033*
[2] *Institute for Physics of Intelligence, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033*
[3] *RIKEN Center for Emergent Matter Science (CEMS), Wako, Saitama 351-0198, Japan*

May 26, 2022

Abstract

This work reports 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 nets with one hidden layer and the first-order phase transition for nets with more than one hidden layer. The proposed theory is directly relevant to the optimization of neural networks and points to an origin of the posterior collapse problem in Bayesian deep learning.

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 correlated so that the model responds to an input in a meaningful way. This is reminiscent of an ordered phase in physics, where the microscopic degrees of freedom behave collectively and coherently. Meanwhile, regularization effectively prevents the overfitting of the model by reducing the correlation between model parameters in a manner similar to the effect of an entropic force in physics. One thus expects a phase transition in the model behavior from the regime where the regularization is negligible to the regime where it is dominant. In the long history of statistical physics of learning [10, 23, 18, 2], a series of works studied the under-to-overparametrization (UO) phase transition in the context of linear regression [12, 13, 23, 9]. Recently, this type of phase transition has seen a resurgence of interest [8, 16]. One recent work by [15] deals with the UO transition in a deep linear model. However, the UO phase transition is not unique to deep learning because it appears in both shallow and deep models and also in non-neural-network models [3]. To understand deep learning, we need to identify what is unique about deep neural networks.

In this work, we address the fundamental problem of the loss landscape of a deep neural network and prove that there exist phase transitions in deep learning that can be described precisely as the first- and second-order phase transitions with a striking similarity to physics. We argue that these phase transitions can have profound implications for deep learning, such as the importance of symmetry breaking for learning and the qualitative difference between shallow and deep architectures. We also show that these phase transitions are unique to machine learning and deep learning. They are unique to machine learning because they are caused by the competition between the need to make predictions more accurate and the need to make the model simpler. These phase transitions are also deep-learning unique because they only appear in "deeper" models. For a multilayer linear net with stochastic neurons and trained with L_2 regularization,

1. we identify an order parameter and effective landscape that describe the phase transition between a trivial phase and a feature learning phase as the L_2 regularization hyperparameter is changed (Theorem 3);

2. we show that finite-depth networks cannot have the zeroth-order phase transition (Theorem 2);

3. we prove that:

(a) depth-0 nets (linear regression) do not have a phase transition (Theorem 1);

(b) depth-1 nets have the second-order phase transitions (Theorem 4);

(c) depth-D nets have the first-order phase transition (Theorem 5) for $D > 1$;

(d) infinite-depth nets have the zeroth-order phase transition (Theorem 6).

The theorem statements and proofs are presented in the Supplementary Section B. To the best of our knowledge, we are the first to identify second-order and first-order phase transitions in the context of deep learning. Our result implies that one can precisely classify the landscape of deep neural models according to the Ehrenfest classification of phase transitions.

Results

Formal framework. Let $\ell(w, a)$ be a differentiable loss function that is dependent on the model parameter w and a hyperparameters a. The loss function ℓ can be decomposed into a data-dependent feature learning term ℓ_0 and a data-independent term $aR(w)$ that regularizes the model at strength a:

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

Learning amounts to finding the global minimizer of the loss:

$$\begin{cases} L(a) := \min_w \ell(w, a); \\ w_* := \arg\min_w \ell(w, a). \end{cases} \tag{2}$$

Naively, one expects $L(a)$ to change smoothly as we change a. If L changes drastically or even discontinuously when one perturb a, it becomes hard to tune a to optimize the model performance. Thus, that $L(a)$ is well-behaved is equivalent to that a is an easy-to-tune hyperparameter. We are thus interested in the case where the tuning of a is difficult, which occurs when a phase transition comes into play.

It is standard to treat the first term in Eq. (1) as an energy. To formally identify the regularization term as an entropy, its coefficient must be proportional to the temperature:

$$aR(w) = \frac{T}{2\sigma^2} R(w), \tag{3}$$

where σ^2 controls the fluctuation of w at zero temperature. We note that this identification is consistent with many previous works, where the term that encourages a lower model complexity is identified as an "entropy" [9, 22, 4, 6, 15]. In this view, learning is a balancing process between the learning error and the model complexity. Intuitively, one expects phase transitions to happen when one term starts to dominate the other, just like thermodynamic phase transitions that take place when the entropy term starts to dominate the energy.

In this setting, the partition function is $Z(a) = \int dw \exp[-\ell(w, a)/T]$. We consider a special limit of the partition function, where both T and $2\sigma^2$ are made to vanish with their ratio held fixed at $T/2\sigma^2 = \gamma$. In this limit, one can find the free energy with the saddle point approximation, which is exact in the zero-temperature limit:

$$F(a) = \lim_{T\to 0,\ \sigma^2 \to 0,\ T/2\sigma^2 = \gamma} -T \log \int dw \exp[-\ell(w, a)/T] = \min_w \ell(w, a). \tag{4}$$

We thus treat L as the free energy.

Definition 1. *$L(a)$ is said to have the nth-order phase transition in a at $a = a^*$ if n is the smallest integer such that $\frac{d^n}{da^n} L(a)|_{a=a^*}$ is discontinuous.*

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

machine learning	statistical physics
training loss	free energy
prediction error	internal energy
regularization	negative entropy

learning process	symmetry breaking
norm of model (b)	order parameter
feature learning regime	ordered phase
trivial regime	disordered phase
noise required for learning	latent heat

Table 1: Left table: the correspondence between machine learning and statistical physics. Right table: the correspondence between a learning process and symmetry breaking.

Definition 2. *$b = b(w) \in \mathbb{R}$ is said to be an order parameter of $\ell(w, a)$ if there exists a function $\bar{\ell}$ such that for all a, $\min_w \bar{\ell}(b(w), a) = L(a)$, 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 $L(a)$. The existence of an order parameter suggests that the original problem $\ell(w, a)$ can effectively be reduced to a low-dimensional problem that is much easier to understand. Physical examples are the average magnetization in the Ising model and the average density of molecules in a water-to-vapor phase transition. A dictionary of the corresponding concepts between physics and deep learning is given in Table 1.

Our theory deals with deep linear nets, the primary minimal model for deep learning. It is well-established that the landscape of a deep linear net can be used to understand that of nonlinear networks [11, 7, 14]. The most general type of deep linear nets, with L_2 regularization and stochastic neurons, has the following loss:

$$\underbrace{\mathbb{E}_x \mathbb{E}_{\epsilon^{(1)}, \epsilon^{(2)}, \ldots, \epsilon^{(D)}} \left(\sum_{i, i_1, i_2, \ldots, i_D}^{d_0, d_0, d_0, \ldots d_0} U_{i_D} \epsilon_{i_D}^{(D)} \ldots \epsilon_{i_2}^{(2)} W_{i_2 i_1}^{(2)} \epsilon_{i_1}^{(1)} W_{i_1 i}^{(1)} x_i - y \right)^2}_{L_0} + \underbrace{\gamma ||U||_2^2 + \sum_{i=1}^{D} \gamma ||W^{(i)}||_F^2}_{L_2 \text{ reg.}}, \tag{5}$$

where x is the input data, y the label, U and $W^{(i)}$ the model parameters, D the network depth, ϵ the noise in the hidden layer (e.g., dropout), d_0 the width of the model, and γ the weight decay strength. We build on the recent results established in [24]. Let $b := ||U||/d_0$. Ref. [24] shows that all the global minima of Eq. (5) must take the form $U = f(b)$ and $W_i = f_i(b)$, where f and f_i are explicit functions of the hyperparameters. Ref. [24] further shows that there are two regimes of learning, where, for some range of γ, the global minimum is uniquely given by $b = 0$, and for some other range of γ, some $b > 0$ gives the global minimum. When $b = 0$, the model outputs a constant 0, and so this regime is called the "trivial regime," and the regime where $b = 0$ is not the global minimum is called the "feature learning regime." In this work, we prove that the transition between these two regimes corresponds to a phase transition in the Ehrenfest sense (Definition 1), and therefore one can indeed refer to these two regimes as two different phases.

No-phase-transition theorems. The first result we prove is that there is no phase transition in any hyperparameter $(\gamma,\ E[xx^T],\ E[xy],\ E[y^2])$ for a simple linear regression problem. In our terminology, this corresponds to the case of $D = 0$. The fact that there is no phase transition in any of these hyperparameters means that the model's behavior is predictable as one tunes the hyperparameters. In the parlance of physics, a linear regressor operates within the linear-response regime.

Theorem 2 shows that a finite-depth net cannot have zeroth-order phase transitions. This theorem can be seen as a worst-case guarantee: the training loss needs to change continuously as one changes the hyperparameter. We also stress 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. The following theory offers a direct explanation for this phenomenon.

Phase Transitions in Deeper Networks. Theorem 4 shows that the quantity b is an order parameter describing any phase transition induced by the weight decay parameter in Eq. (5). Let $b = ||U||/d_u$, $A_0 := \mathbb{E}[xx^T]$, and a_i be the i-th eigenvalue of A_0. The effective loss landscape is

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

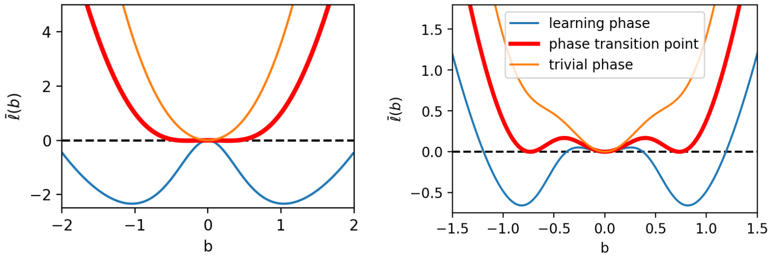

Figure 1: Effective landscape given in Eq. (6) for $D = 1$ (left) and $D = 2$ (right). For $D = 1$, zero is either the global minimum or a local maximum. Note that the shape of the loss resembles that of the Landau free energy for the second-order phase transition. For $D = 2$, the landscape becomes more complicated, featuring the emergence of local minima. In particular, zero is always a local minimum.

where x' is a rotation of x. See Figure 1 for an illustration. The complicated landscape for $D > 1$ implies that neural networks are susceptible to initialization schemes and entrapment in meta-stable states is common (see Supplementary Section A.1).

Theorem 5 shows that when $D = 1$ in Eq. (5), there is a second-order phase transition precisely at

$$\gamma = ||\mathbb{E}[xy]||. \tag{7}$$

In machine learning language, γ is the regularization strength and $||E[xy]||$ is the signal. The phase transition occurs precisely when the regularization dominates the signal. In physics, γ and $||\mathbb{E}[xy]||$ are proportional to the temperature and energy, respectively. The phase transition occurs exactly when the entropy dominates the energy. Also, the phase transition for a depth-1 linear net is independent of the number of parameters of the model. For $D > 1$, the size of the model does play a role in influencing 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 that the previous works often adopt.

For $D \geq 2$, we show that there is a first-order phase transition between the two phases at some $\gamma > 0$. 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 we have to inject into the model parameters to escape from a local to the global minimum for a deep model. We illustrate the phase transitions studied in Figure 2. We also experimentally demonstrate that the same phase transitions take place in deep nonlinear networks with the corresponding depths (Supplementary Section A.3). While infinite-depth networks are not used in practice, they are important from a theoretical point of view [20] because they can be used for understanding a (very) deep network that often appears in the deep learning practice. Our result shows that the limiting landscape has a zeroth-order phase transition at $\gamma = 0$. In fact, zeroth-order phase transitions do not occur in physics, and it is a unique feature of deep learning.

Relevance of symmetry breaking. The phase transitions we studied also involve symmetry breaking. This can be seen directly from the effective landscape in Eq. (6). The loss is unaltered as one flip the sign of b, and therefore the loss is symmetric in b. Figure 3 illustrates the effect and importance of symmetry breaking on the gradient descent dynamics. Additionally, this observation may also provide an alternative venue for studying general symmetry-breaking dynamics because the computation with neural networks is both accurate and efficient.

Mean-Field Analysis. The crux of our theory can be understood by applying a simplified "mean-field" analysis of the loss function in Eq. (5). Let each weight matrix be approximated by a scalar $U = b_{D+1}$, $W_i = b_i$, ignore the stochasticity due to ϵ_i, and let x be one-dimensional. One then obtains a simplified mean-field loss:

$$\mathbb{E}_x\left[\left(c_0 x \prod_{i=1}^{D+1} b_i - y\right)^2\right] + \gamma \sum_{i=1}^{D} c_i b_i^2, \tag{8}$$

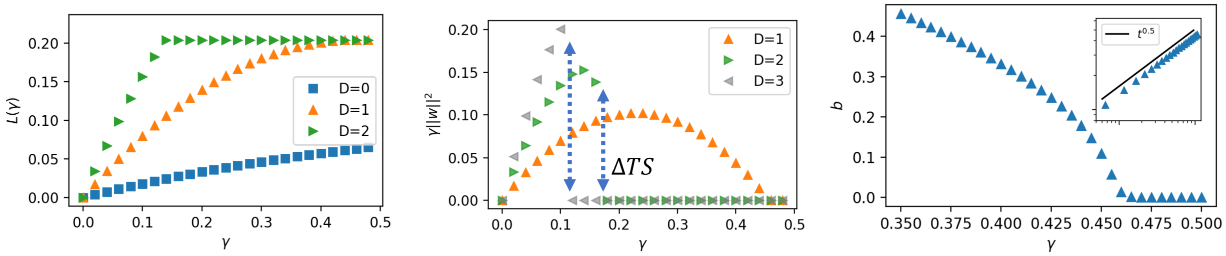

Figure 2: Phase transitions in a linear net. In agreement with the theory, a depth-0 net has no phase transition. A depth-1 net has a second-order phase transition at approximately $\gamma = 0.45$, close to the theoretical value of $||\mathbb{E}[xy]||$, and a depth-2 net has a first-order phase transition at roughly $\gamma = 0.15$. The qualitative differences between networks of different depths are clearly observed in the data. **Left**: Training loss of a network with 0 (linear regression), 1, and 2 hidden layers. Clearly, a depth-0 net shows no phase transition. A depth-1 net has a second-order phase transition at approximately $\gamma = 0.45$, and a depth-2 net has a first-order phase transition at roughly $\gamma = 0.15$. **Middle**: Magnitude of the regularization term at convergence. As discussed in the main text, this term corresponds to the entropy term TS. We see that for $D > 1$, there is a jump (discontinuity) in TS from a finite value to 0. This jump corresponds to the latent heat of the first-order phase transition process. **Right**: Order parameter as a function of γ. The inset shows that b precisely scales as $t^{0.5}$ with $t := -(\gamma - \gamma^*)$ in the vicinity of the phase transition, in agreement with the standard Landau theory.

where c_i's are constants. The first term can be interpreted as a special type of $(D+1)$-body interaction. We now perform a second mean-field approximation, where all the b_i take the same value b:

$$\ell \propto c_0'\mathbb{E}[x^2]b^{2D+2} - c_1'\mathbb{E}[xy]b^{D+1} + \gamma c_2' b^2 + const. \tag{9}$$

Here, c_0', c_1' and c_2' are structural constants, only depending on the model (depth, width, etc). The first and the third terms monotonically increase in b, encouraging a smaller b. The second term monotonically decreases in $b^{D+1}\mathbb{E}[xy]$, encouraging a positive correlation between b and the feature $\mathbb{E}[xy]$. The leading and lowest-order terms regularize the model, while the intermediate term characterizes learning. For $D = 0$, the loss is quadratic and has no transition. For $D = 1$, the loss is identical to the Landau free energy, and a phase transition occurs when the second-order term flips sign: $c_2'\gamma = c_1'\mathbb{E}[xy]$. 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, the common interaction is many-body and tends to infinite-body as D increases.

One implication is that 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 $\gamma||w||_2^{d+2}$, which is the lowest-order regularization that does not lead to a trivial phase. The effectiveness of this suggested method is confirmed in Supplementary Section A.2.

Posterior Collapse in Bayesian Deep Learning. Our results also identify an origin of the well-known problem posterior collapse problem in Bayesian deep learning. Posterior collapse refers to the learning failure where the learned posterior distribution coincides with the prior, and so no learning has happened even after training [5, 1, 17]. Our results offer a direct explanation for this posterior collapse problem. In the Bayesian interpretation, the training loss in Eq. (5) is the exact negative log posterior, and the trivial phase exactly corresponds to the posterior collapse: the global minimum of the loss is identical to the global maximum of the prior term. Our results thus imply that (1) posterior collapse is a unique problem of deep learning because it does not occur in shallow models, and (2) posterior collapse happens as a direct consequence of the competition between the prior and the likelihood. This means that it is not a good idea to assume a Gaussian prior for the deep neural network models. The suggested fix also leads to a clean and Bayesian-principled solution to the posterior collapse problem by using a prior $\log p(w) \propto -||w||_2^{D+2}$.

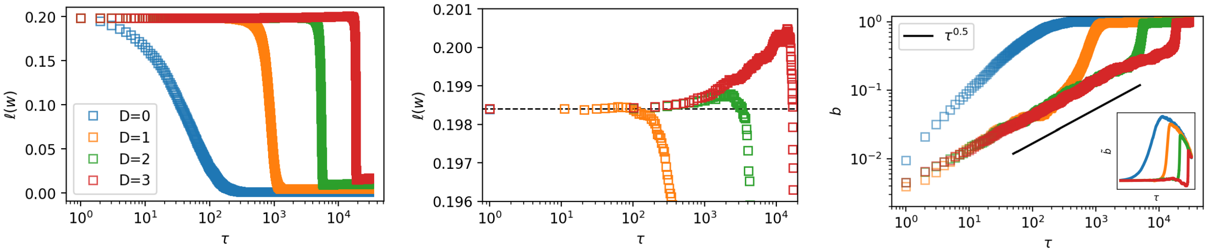

Figure 3: Dynamics of training, where the model is initialized at the origin and the learning proceeds under gradient descent with injected Gaussian noise. Before training, the models lie roughly in the trivial phase because the model is initialized close to the origin and has not learned anything yet. However, for the feature learning phase, any global minimum must choose a specific $b \neq 0$, and so the actual solution does not feature the symmetry in b: a symmetry breaking in b must take place for the learning to happen. The recent work of Ref. [21] showed that the symmetries in the loss could become difficult obstacles in the training of a neural network, and our result complements this view by identifying a precise deep-learning-relevant symmetry to be broken. **Left**: the training loss L; except for $D = 0$, where no symmetry breaking occurs, the dynamics exhibits a wide plateau that hinders learning emerging at initialization. **Middle**: a zoom-in of the left panel 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, 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 in order to undergo the first-order phase transition. **Right**: time evolution of the order parameter b: one sees that $D = 0$ shows a fast increase of b from the beginning. For $D \geq 1$, the initial stage is dominated by slow diffusion, where b increases as the square root of time. The diffusion phase only ends after a long period, before a fast learning period begins. One also notices that in the fast learning period, the slope of b versus time is different for different depths, with deeper models considerably faster than shallower ones. The inset shows the corrected order parameter $\tilde{b} := b - D\sqrt{\tau}$, where τ is the training step, and D is the diffusion constant of the noisy gradient descent. One sees that $\tilde{b}$ stays zero over an extended period of time for $D > 0$.

Discussion

The striking similarity between phase transitions in neural networks and statistical-physics phase transitions lends a great impetus to a more thorough investigation of deep learning through the lens of thermodynamics and statistical physics. We now outline a few major future steps:

1. Instead of classification by analyticity, can we classify neural networks by symmetry and topological invariants?

2. What are other possible phases for a nonlinear network? Does a new phase emerge?

3. Can we find any analogy of other thermodynamic quantities such as volume and pressure? More broadly, can we establish thermodynamics for deep learning?

4. Can we utilize the latent heat picture to devise better algorithms for escaping local minima in deep learning?

This work shows that the Ehrenfest classification of phase transitions aligns precisely with the number of layers in deep neural networks. We believe that the statistical-physics approach to deep learning will bring about fruitful developments in both fields of statistical physics and deep learning.

References

[1] Alemi, A., Poole, B., Fischer, I., Dillon, J., Saurous, R. A., and Murphy, K. (2018). Fixing a broken elbo. In *International Conference on Machine Learning*, pages 159–168. PMLR.

[2] Bahri, Y., Kadmon, J., Pennington, J., Schoenholz, S. S., Sohl-Dickstein, J., and Ganguli, S. (2020). Statistical mechanics of deep learning. *Annual Review of Condensed Matter Physics*, 11:501–528.

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

[4] Benedek, G. M. and Itai, A. (1991). Learnability with respect to fixed distributions. *Theoretical Computer Science*, 86(2):377–389.

[5] Dai, B. and Wipf, D. (2019). Diagnosing and enhancing vae models. *arXiv preprint arXiv:1903.05789*.

[6] Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in cognitive sciences*, 13(7):293–301.

[7] Hardt, M. and Ma, T. (2016). Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*.

[8] Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2019). Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*.

[9] Haussler, D., Kearns, M., Seung, H. S., and Tishby, N. (1996). Rigorous learning curve bounds from statistical mechanics. *Machine Learning*, 25(2):195–236.

[10] Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558.

[11] Kawaguchi, K. (2016). Deep learning without poor local minima. *Advances in Neural Information Processing Systems*, 29:586–594.

[12] Krogh, A. and Hertz, J. A. (1992a). Generalization in a linear perceptron in the presence of noise. *Journal of Physics A: Mathematical and General*, 25(5):1135.

[13] Krogh, A. and Hertz, J. A. (1992b). A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pages 950–957.

[14] Laurent, T. and Brecht, J. (2018). Deep linear networks with arbitrary loss: All local minima are global. In *International conference on machine learning*, pages 2902–2907. PMLR.

[15] Li, Q. and Sompolinsky, H. (2021). Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization. *Physical Review X*, 11(3):031059.

[16] Liao, Z., Couillet, R., and Mahoney, M. W. (2020). A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent. *Advances in Neural Information Processing Systems*, 33:13939–13950.

[17] Lucas, J., Tucker, G., Grosse, R., and Norouzi, M. (2019). Don't blame the elbo! a linear vae perspective on posterior collapse.

[18] Martin, C. H. and Mahoney, M. W. (2017). Rethinking generalization requires revisiting old ideas: statistical mechanics approaches and complex learning behavior. *arXiv preprint arXiv:1710.09553*.

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

[20] Sonoda, S. and Murata, N. (2019). Transport analysis of infinitely deep neural network. *The Journal of Machine Learning Research*, 20(1):31–82.

[21] Tanaka, H. and Kunin, D. (2021). Noether's learning dynamics: Role of symmetry breaking in neural networks. *Advances in Neural Information Processing Systems*, 34.

[22] Vapnik, V. (2006). *Estimation of dependences based on empirical data.* Springer Science & Business Media.

[23] Watkin, T. L., Rau, A., and Biehl, M. (1993). The statistical mechanics of learning a rule. *Reviews of Modern Physics*, 65(2):499.

[24] Ziyin, L., Li, B., and Meng, X. (2022). Exact solutions of a deep linear network. *arXiv preprint arXiv:2202.04777*.

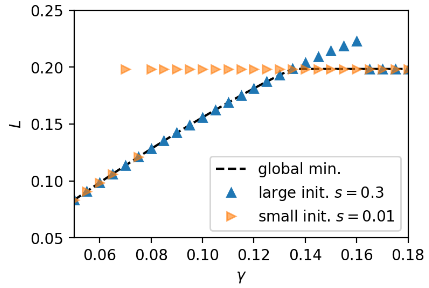

Figure 4: Sensitivity of the obtained solution to the initialization of the model. We initialize the model around zero with standard deviation s. The experiment shows that a larger initialization variance ($s = 0.3$) affords a preference of the nontrivial solution over the trivial one, while a smaller initialization leads to the opposite preference.

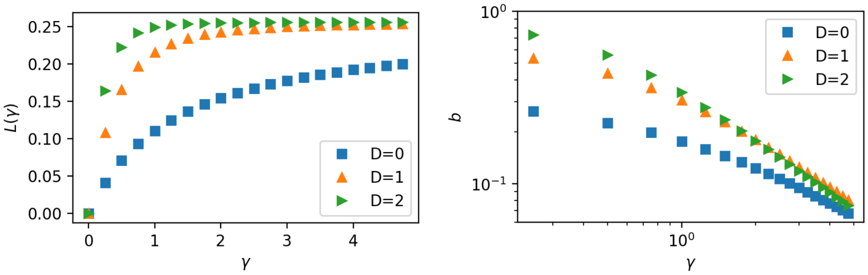

Figure 5: The training loss $L(\gamma)$ (left) and the model norm b (right) when we train with a regularization term of the form $\gamma||w||^{D+2}$, which is a theoretically justified fix to the trivial learning problem. We see that the trivial phase disappears under this regularization.

A Additional Experiments

A.1 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) larger 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, and a small s means that it is initialized closer to the trivial phase and a larger s means that it is initialized closer to the learning phase. We see that across 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 to the system to leave the meta-stable state; otherwise, the system may become stuck for a very long time. The existing initialization methods are predominantly data-dependent. However, our result (also see [24]) 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.

A.2 Removing the Trivial Phase

We also explore our suggested fix to the trivial learning problem. Here, instead of regularization the model by $\gamma||w||_2^2$, we regularize the model by $\gamma||w||_2^{D+2}$. The training loss and the model norm b are plotted in Figure 5. We find that the trivial phase now completely disappears even if we go to very high 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.

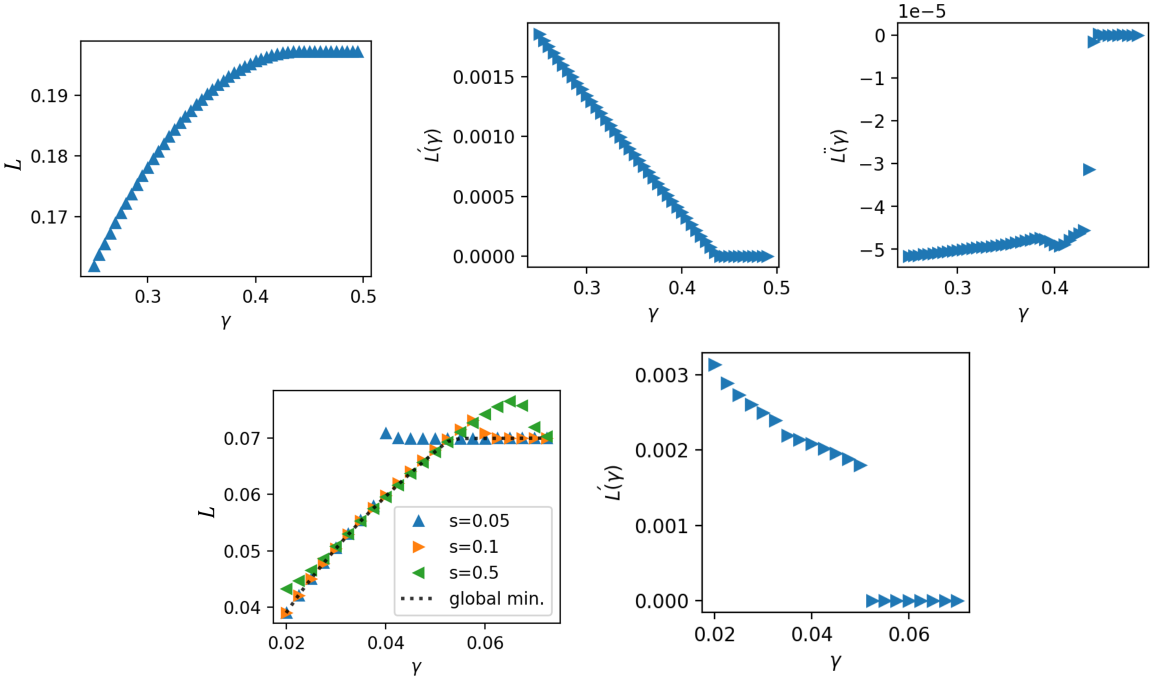

Figure 6: Phase transition of a fully connected tanh network. **Top** row shows the case of $D = 1$, exhibiting a second-order phase transition: the training loss $L(\gamma)$ (left), first derivative (middle), and the second derivative (right). **Bottom** row shows the case of $D = 2$, exhibiting a first-order 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.

A.3 Nonlinear Networks

We expect our theory to also apply to deep nonlinear networks that can be locally approximated by linear net at the origin, e.g., a network with tanh activations. As shown in Figure 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 zero (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.

B Main Results

B.1 Theorem Statements

For a simple ridge linear regression, the minimization objective is

$$\ell(W) = \mathbb{E}_x\left(\sum_i W_i x_i - y\right)^2 + \gamma||W||^2. \tag{10}$$

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 $\gamma \in (0,\infty)$.*

The following result shows that for a finite depth, $L(\gamma)$ must be continuous in γ.

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

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.

The following theorem shows that there exists order parameters describing any phase transition induced by the weight decay parameter in Eq. (5).

Theorem 3. *Let $b = ||U||/d_u$, and let*

$$\bar{\ell}(b,\gamma) := -\sum_i \frac{d_0^{2D} b^{2D} \mathbb{E}[x'y]_i^2}{d_0^D(\sigma^2 + d_0)^D a_i b^{2D} + \gamma} + \mathbb{E}_x[y^2] + \gamma D d_0^2 b^2. \tag{11}$$

Then, b is an order parameter of Eq. (5) for the effective loss $\bar{\ell}$.

Here, the norm of the last layer is referred to as the order parameter. The meaning of this choice should be clear. The norm of the last layer is zero if and only if all weights of the last layer is zero, and the model is a trivial model. The model can only learn something when the order parameter is nonzero. Additionally, we note that the choice of the order parameter is not unique and there are other choices for the order parameter (e.g., the norm of any other layer, or the sum of the norms of all layers).

The following theorem shows that when $D = 1$ in Eq. (5), there is a second-order phase transition with respect to γ.

Theorem 4. *Equation (5) has the second-order phase transition between the trivial and feature learning phases at*[1]

$$\gamma = ||\mathbb{E}[xy]||. \tag{12}$$

Now, we show that for $D \geq 2$, there is a first-order phase transition.

Theorem 5. *Let $D \geq 2$. 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 $\gamma = \gamma^*$.*

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,*

$$L^{(D)}(\gamma) \to \begin{cases} r & \textit{if } \gamma = 0; \\ \mathbb{E}[y^2] & \textit{otherwise.} \end{cases} \tag{13}$$

The constant r is, in general, not equal to $\mathbb{E}[y^2]$. For example, in the limit $\sigma \to 0$, r converges to the loss value of a simple linear regression, which is not equal to $\mathbb{E}[y^2]$ as long as $\mathbb{E}[xy] \neq 0$.

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

B.2 Proof of Theorem 1

Proof. The global minimum of Eq. (10) is

$$W_* = (A_0 + \gamma I)^{-1} E[xy]. \tag{14}$$

The loss of the global minimum is thus

$$\begin{aligned} L &= \mathbb{E}_x \left(\sum_i W_i x_i - y \right)^2 + \gamma ||W||^2 && (15) \\ &= W^T A_0 W - 2W^T \mathbb{E}[xy] + \mathbb{E}[y^2] + \gamma ||W||^2 && (16) \\ &= \mathbb{E}[xy]^T \frac{A_0}{(A_0 + \gamma I)^2} \mathbb{E}[xy] - 2\mathbb{E}[xy]^T \frac{1}{A_0 + \gamma I} \mathbb{E}[xy] + \mathbb{E}[y^2] + \gamma \mathbb{E}[xy]^T \frac{1}{(A_0 + \gamma I)^2} \mathbb{E}[xy] && (17) \\ &= -\mathbb{E}[xy]^T (A_0 + \gamma I)^{-1} \mathbb{E}[xy] + \mathbb{E}[y^2], && (18) \end{aligned}$$

which is infinitely differentiable for any $\gamma \in (0, \infty)$ (note that A_0 is always positive semi-definite by definition). □

B.3 Proof of Theorem 2

Proof. 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

$$L(\gamma^* - \epsilon) = \min_w \ell(w, \gamma^* - \epsilon) := \ell(w', \gamma^* - \epsilon), \tag{19}$$

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 $\epsilon > 0$,

$$L(\gamma^* - \epsilon) < L(\gamma^*) - \delta, \tag{20}$$

which implies that

$$\ell(w', \gamma^* - \epsilon) = L(\gamma^* - \epsilon) < L(\gamma^*) - \delta \le \ell(w', \gamma^*) - \delta. \tag{21}$$

Namely, the left-discontinuity implies that for all $\epsilon > 0$,

$$\ell(w', \gamma^* - \epsilon) \le \ell(w', \gamma^*) - \delta. \tag{22}$$

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

$$\ell(w, \gamma) - \ell(w, \gamma - \epsilon) = \epsilon ||w||^2. \tag{23}$$

Thus, by choosing $\epsilon < \delta / ||w||^2$, the relation in (21) 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

$$L(\gamma^* + \epsilon) = \min_w \ell(w, \gamma^* + \epsilon) := \ell(w', \gamma^* + \epsilon), \tag{24}$$

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 $\epsilon > 0$,

$$L(\gamma^* + \epsilon) > L(\gamma^*) + \delta, \tag{25}$$

which implies that

$$\ell(w', \gamma^* + \epsilon) = L(\gamma^* + \epsilon) > L(\gamma^*) + \delta \ge \ell(w', \gamma^*) + \delta. \tag{26}$$

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

$$\ell(w', \gamma^* + \epsilon) \geq \ell(w', \gamma^*) + \delta. \tag{27}$$

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

$$\ell(w, \gamma + \epsilon) - \ell(w, \gamma) = \epsilon ||w||^2. \tag{28}$$

Thus, by choosing $\epsilon < \delta/||w||^2$, the relation in (26) 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 $\gamma = 0$. □

B.4 Proof of Theorem 3

Proof. By Theorem 3 of Ref. [24], any global minimum of Eq. (5) is given by the following set of equations for some $b \geq 0$:

$$\begin{cases} U = \sqrt{d_0} b \mathbf{r}_D; \\ W^{(i)} = b \mathbf{r}_i \mathbf{r}_{i-1}^T; \\ W^{(1)} = \mathbf{r}_1 \mathbb{E}[xy]^T d_0^{D-\frac{1}{2}} b^D \left[d_0^D (\sigma^2 + d_0)^D b^{2D} A_0 + \gamma \right]^{-1}, \end{cases} \tag{29}$$

where $\mathbf{r}_i = (\pm 1, ..., \pm 1)$ 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

$$f(x) := \sum_{i, i_1, i_2, ..., i_D}^{d, d_1, d_2, ... d_D} U_{i_D} \epsilon_{i_D}^{(D)} ... \epsilon_{i_2}^{(2)} W_{i_2 i_1}^{(2)} \epsilon_{i_1}^{(1)} W_{i_1 i}^{(1)} x, \tag{30}$$

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

Substituting Eq. (29) in Eq. (5), one finds that within this subspace, the loss function can be written as

$$\begin{aligned} \ell(w, \gamma) &= \mathbb{E}_x \mathbb{E}_\eta (f(x) - y)^2 + L_2 \text{ reg.} & (31) \\ &= \mathbb{E}_{x,\eta}[f(x)^2] - 2\mathbb{E}_{x,\eta}[y f(x)] + \mathbb{E}_x[y^2] + L_2 \text{ reg.} & (32) \\ &= \sum_i \frac{d_0^{3D} (\sigma^2 + d_0)^D b^{4D} a_i \mathbb{E}[x'y]_i^2}{[d_0^D (\sigma^2 + d_0)^D a_i b^{2D} + \gamma]^2} - 2 \sum_i \frac{d_0^{2D} b^{2D} \mathbb{E}[x'y]_i^2}{d_0^D (\sigma^2 + d_0)^D a_i b^{2D} + \gamma} + \mathbb{E}_x[y^2] + L_2 \text{ reg.}, & (33) \end{aligned}$$

where the L_2 reg. term is

$$L_2 \text{ reg.} = \gamma D d_0^2 b^2 + \gamma \sum_i \frac{d_0^{2D} b^{2D} ||\mathbb{E}[x'y]_i||^2}{[d_0^D (\sigma^2 + d_0)^D b^{2D} a_i + \gamma]^2}. \tag{34}$$

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

$$-\sum_i \frac{d_0^{2D} b^{2D} \mathbb{E}[x'y]_i^2}{[d_0^D (\sigma^2 + d_0)^D a_i b^{2D} + \gamma]} + \mathbb{E}_x[y^2] + \gamma D d_0^2 b^2. \tag{35}$$

We can now define the effective loss by

$$\bar{\ell}(b, \gamma) := -\sum_i \frac{d_0^{2D} b^{2D} \mathbb{E}[x'y]_i^2}{[d_0^D (\sigma^2 + d_0)^D a_i b^{2D} + \gamma]} + \mathbb{E}_x[y^2] + \gamma D d_0^2 b^2. \tag{36}$$

Then, the above argument shows that, for all γ,

$$\min_b \bar{\ell}(b, \gamma) = \min_w \ell(w, \gamma). \tag{37}$$

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

B.5 Two Useful Lemmas

Before continuing the proofs, we first prove two lemmas that will simplify the following proofs significantly.

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

$$\frac{d}{d\gamma}L(\gamma) = \sum_i \frac{d_0^{2D} b_*^{2D} \mathbb{E}[x'y]_i^2}{[d_0^D(\sigma^2+d_0)^D a_i b_*^{2D} + \gamma]^2} + D d_0^2 b_*^2 \geq 0. \tag{38}$$

Proof. Because L is differentiable in γ, one can find the derivative for at least one of the global minima b^*

$$\frac{d}{d\gamma}L(\gamma) = \frac{d}{d\gamma}\bar{\ell}(b^*(\gamma),\gamma) \tag{39}$$

$$= \frac{\partial}{\partial b^*}\bar{\ell}(b^*,\gamma)\frac{\partial b^*}{\partial \gamma} + \frac{\partial}{\partial \gamma}\bar{\ell}(b^*,\gamma) \tag{40}$$

$$= \frac{\partial}{\partial \gamma}\bar{\ell}(b^*,\gamma) \tag{41}$$

$$= \sum_i \frac{d_0^{2D} b_*^{2D} \mathbb{E}[x'y]_i^2}{[d_0^D(\sigma^2+d_0)^D a_i b_*^{2D} + \gamma]^2} + D d_0^2 b_*^2 \geq 0, \tag{42}$$

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

B.6 Proof of Theorem 4

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

$$\bar{\ell}(b,\gamma) = -d_1 b^2 E[xy]^T [b^2(\sigma^2+d_1)A + \gamma I]^{-1} E[xy] + E[y^2] + \gamma d_1 b^2. \tag{43}$$

By Theorem 1 of Ref. [24], the phase transition, if exists, must occur precisely at $\gamma = ||\mathbb{E}[xy]||$. To prove that $\gamma = ||\mathbb{E}[xy]||$ has a second-order phase transition, we must check both its first derivative and second derivative.

When $\gamma \to ||E[xy]||$ from the right, we find that the all derivatives of $L(\gamma)$ are zero 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. (43). Because Eq.(43) is differentiable, its derivative in b must be equal to 0 at the global minimum

$$-2\gamma d_1 b \mathbb{E}[xy]^T [b^2(\sigma^2+d_1)^2 A + \gamma I]^{-2} \mathbb{E}[xy] + 2\gamma d_1 b = 0. \tag{44}$$

Finding the minimizer b is thus equivalent to finding the real roots of a high-order polynomial in b. When $\gamma \geq ||\mathbb{E}[xy]||$, the solution is unique [24]:

$$b_0^2 = 0, \tag{45}$$

where we labeled 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 (44) simplifies to

$$\mathbb{E}[xy]^T [b^2(\sigma^2+d_1)A + \gamma I]^{-2} \mathbb{E}[xy] = 1. \tag{46}$$

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

$$b(\gamma)^2 = \beta_0 + \beta_1 \Delta + O(\Delta^2). \tag{47}$$

We first Substitute (47) in (44) to find[2]

$$\beta_0 = 0. \tag{48}$$

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

One can then again substitute Eq. (47) in Eq. (44) to find β_1. To the first order in b^2, Eq. (44) reads

$$\frac{1}{\gamma^2}||\mathbb{E}[xy]||^2 - 2b^2\frac{(\sigma^2+d_1)}{\gamma^3}||\mathbb{E}[xy]||^2_{A_0} = 1 \tag{49}$$

$$\iff -2\beta_1\Delta\frac{(\sigma^2+d_1)}{\gamma^3}||\mathbb{E}[xy]||^2_{A_0} = 2\frac{\Delta}{||\mathbb{E}[xy]||} \tag{50}$$

$$\iff \beta_1 = -\frac{1}{(\sigma^2+d_1)}\frac{||\mathbb{E}[xy]||^2}{||\mathbb{E}[xy]||^2_{A_0}} \tag{51}$$

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

$$\frac{d}{d\gamma}L(\gamma)|_{\gamma=||E[xy]||_-} \sim b_*^2 = 0 = \frac{d}{d\gamma}L(\gamma)|_{\gamma=||E[xy]||_+}. \tag{52}$$

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

$$b(\gamma)^2 = 0 + \beta_1\Delta + \beta_2\Delta^2 + O(\Delta^3). \tag{53}$$

To the second order in b^2, (44) reads

$$\frac{1}{\gamma^2}||\mathbb{E}[xy]||^2 - 2b^2\frac{(\sigma^2+d_1)}{\gamma^3}||\mathbb{E}[xy]||^2_{A_0} + 3b^4\frac{(\sigma^2+d_1)^2}{\gamma^4}||\mathbb{E}[xy]||^2_{A_0^2} = 1 \tag{54}$$

$$\iff \gamma^2||\mathbb{E}[xy]||^2 - 2b^2(\sigma^2+d_1)\gamma||\mathbb{E}[xy]||^2_{A_0} + 3b^4(\sigma^2+d_1)^2||\mathbb{E}[xy]||^2_{A_0^2} = \gamma^4 \tag{55}$$

$$\iff \Delta^2E_0^2 - 2\beta_2\Delta^2(\sigma^2+d_1)E_0E_1^2 - 2\beta_1\Delta^2(\sigma^2+d_1)E_1^2 + 3\beta_1^2\Delta^2(\sigma^2+d_1)^2E_2^2 = 6E_0^2\Delta^2 \tag{56}$$

$$\iff \beta_2 = \frac{3\beta_1^2(\sigma^2+d_1)^2E_2^2 - 5E_0^2 - 2\beta_1(\sigma^2+d_1)E_1^2}{2(\sigma^2+d_1)E_0E_1^2}, \tag{57}$$

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

$$\beta_2 = \frac{3E_0(E_2^2 - E_1^2)}{2(\sigma^2+d_1)E_1^4}. \tag{58}$$

This allows us to find the second derivative of $L(\gamma)$. Substituting β_1 and β_2 into Eq. (43) and expanding to the second order in Δ, we obtain that

$$L(\gamma) = -d_1b^2E[xy]^T[b^2(\sigma^2+d_1)A+\gamma I]^{-1}E[xy] + E[y^2] + \gamma d_1b^2 \tag{59}$$

$$= -d_1(\beta_1\Delta+\beta_2\Delta^2)\mathbb{E}[xy]^T[(\beta_1\Delta+\beta_2\Delta^2)(\sigma^2+d_1)A_0+\gamma I]^{-1}\mathbb{E}[xy] + \gamma d_1(\beta_1\Delta+\beta_2\Delta). \tag{60}$$

At the critical point,

$$\frac{d^2}{d\gamma^2}L(\gamma)|_{\gamma=||\mathbb{E}[xy]||_-} = -d_1\beta_2E_0 + d_1\beta_1^2(\sigma^2+d_1)\frac{E_1^2}{E_0^2} + d_1\beta_1 + d_1\beta_1 + d_1\beta_2E_0 \tag{61}$$

$$= 2d_1\beta_1 + d_1\beta_1^2(\sigma^2+d_1)\frac{E_1^2}{E_0^2} \tag{62}$$

$$= d_1\beta_1 \tag{63}$$

$$= -\frac{d_1}{\sigma^2+d_1}\frac{||\mathbb{E}[xy]||^2}{||\mathbb{E}[xy]||^2_{A_0}}. \tag{64}$$

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

$$\frac{d^2}{d\gamma^2}L(\gamma)|_{\gamma=||\mathbb{E}[xy]||_-} = -\frac{d_1}{\sigma^2+d_1}\frac{||\mathbb{E}[xy]||^2}{||\mathbb{E}[xy]||^2_{A_0}} < 0. \tag{65}$$

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

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

B.7 Proof of Theorem 5

Proof. By definition, it suffices to show that $\frac{d}{d\gamma}L(\gamma)$ is not continuous. We prove 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^*

$$\frac{d}{d\gamma}L(\gamma) = \sum_i \frac{d_0^{2D} b_*^{2D} \mathbb{E}[x'y]_i^2}{[d_0^D(\sigma^2+d_0)^D a_i b_*^{2D}+\gamma]^2} + \gamma D d_0^2 b_*^2 \geq 0. \tag{66}$$

Both terms in the last line 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 $\gamma D d_0^2 b_*^2$ is continuous in γ. By Proposition 3 of Ref. [24], there exist constants c_0, c_1 such that $0 < c_0 \leq c_1$, and

$$\begin{cases} b_* = 0 & \text{if } \gamma < c_0; \\ b_* > 0, & \text{if } \gamma > c_1. \end{cases} \tag{67}$$

Additionally, if $b_* > 0$, b_* must be lower-bounded by some nonzero value [24]:

$$b_* \geq \frac{1}{d_0}\left(\frac{\gamma}{||\mathbb{E}[xy]||}\right)^{\frac{1}{D-1}} > \frac{1}{d_0}\left(\frac{c_1}{||\mathbb{E}[xy]||}\right)^{\frac{1}{D-1}} > 0. \tag{68}$$

Therefore, for any $D > 1$, $b_*(\gamma)$ must have a discontinuous jump from 0 to a value larger than $\frac{1}{d_0}\left(\frac{c_0}{||\mathbb{E}[xy]||}\right)^{\frac{1}{D-1}}$, 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. □

B.8 Proof of Theorem 6

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. [24], when $\gamma > 0$, any nonzero global minimum must obey the following two inequalities:

$$\frac{1}{d_0}\left[\frac{\gamma}{||\mathbb{E}[xy]||}\right]^{\frac{1}{D-1}} \leq b^* \leq \left[\frac{||\mathbb{E}[xy]||}{d_0(\sigma^2+d_0)^D a_{\max}}\right]^{\frac{1}{D+1}}, \tag{69}$$

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

$$\frac{1}{d_0}\left[\frac{\gamma}{||\mathbb{E}[xy]||}\right]^{\frac{1}{D-1}} \to \frac{1}{d_0}. \tag{70}$$

The upper bound becomes

$$\left[\frac{||\mathbb{E}[xy]||}{d_0(\sigma^2+d_0)^D a_{\max}}\right]^{\frac{1}{D+1}} \to \frac{1}{\sigma^2+d_0}. \tag{71}$$

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. [19] 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. □