Exact Solutions of a Deep Linear Network

Liu Ziyin[1], **Botao Li**[2], **Xiangming Meng**[3]
[1]Department of Physics, The University of Tokyo
[2]Laboratoire de Physique de l'Ecole normale supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris Cité, Paris, France
[3]Institute for Physics of Intelligence, Graduate School of Science, The University of Tokyo

Abstract

This work finds the analytical expression of the global minima of a deep linear network with weight decay and stochastic neurons, a fundamental model for understanding the landscape of neural networks. Our result implies that zero is a special point in deep neural network architecture. We show that weight decay strongly interacts with the model architecture and can create bad minima at zero in a network with more than 1 hidden layer, qualitatively different from a network with only 1 hidden layer. Practically, our result implies that common deep learning initialization methods are insufficient to ease the optimization of neural networks in general.

1 Introduction

Applications of neural networks have achieved great success in various fields. One central open question is why neural networks, being nonlinear and containing many saddle points and local minima, can sometimes be optimized easily (Choromanska et al., 2015a) while becoming difficult and requiring many tricks to train in some other scenarios (Glorot and Bengio, 2010; Gotmare et al., 2018). One established approach is to study the landscape of deep linear nets (Choromanska et al., 2015b), which are believed to approximate the landscape of a nonlinear net well. A series of works proved the famous results that for a deep linear net, all local minima are global (Kawaguchi, 2016; Lu and Kawaguchi, 2017; Laurent and Brecht, 2018), which is regarded to have successfully explained why deep neural networks are so easy to train because it implies that initialization in any attractive basin can reach the global minimum without much effort (Kawaguchi, 2016). However, the theoretical problem of when and why neural networks can be hard to train is understudied.

In this work, we theoretically study a deep linear net with weight decay and stochastic neurons, whose loss function takes the following form in general:

$$\underbrace{\mathbb{E}_x \mathbb{E}_{\epsilon^{(1)},\epsilon^{(2)},\ldots,\epsilon^{(D)}} \left(\sum_{i,i_1,i_2,\ldots,i_D}^{d,d_1,d_2,\ldots d_D} U_{i_D} \epsilon^{(D)}_{i_D} \ldots \epsilon^{(2)}_{i_2} W^{(2)}_{i_2 i_1} \epsilon^{(1)}_{i_1} W^{(1)}_{i_1 i} x_i - y \right)^2}_{L_0} + \underbrace{\gamma_u \|U\|_2^2 + \sum_{i=1}^{D} \gamma_i \|W^{(i)}\|_F^2}_{L_2\ reg.}, \tag{1}$$

where $\mathbb{E}_x$ denotes the expectation over the training set, U and $W^{(i)}$ are the model parameters, D is the depth of the network,[1] ϵ is the noise in the hidden layer (e.g., due to dropout), d_i is the width of the i-th layer, and γ is the strength of the weight decay. Previous works have studied special cases of this loss function. For example, Kawaguchi (2016) and Lu and Kawaguchi (2017) study the landscape of L_0 when ϵ is a constant (namely, when there is no noise). Mehta et al. (2021) studies

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

[1]In this work, we use "depth" to refer to the number of hidden layers. For example, a linear regressor has depth 0.

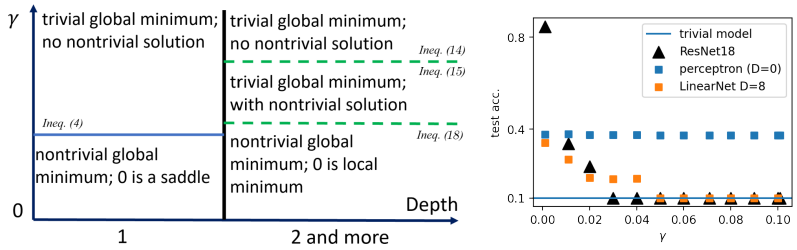

Figure 1: **Left**: A summary of the network landscape that is implied by the main results of this work when one increases the weight decay strength γ while fixing other terms. We show that the landscape of a depth-1 net can be precisely divided into two regimes, while, for $D \geq 2$, there exists at least three regimes. The solid blue line indicates that the division of the regimes is precisely understood. The dashed lines indicate that the conditions we found are not tight and may be improved in the future. **Right**: ResNet18 on CIFAR10. The performance of a linear regressor never drops to that of a trivial model, whereas the performance of ResNet18 drops to the level of a trivial model, like a deep linear net with similar depth.

L_0 with (a more complicated type of) weight decay but without stochasticity and proved that all the stationary points are isolated. Another line of works studies L_0 when the noise is caused by dropout (Mianjy and Arora, 2019; Cavazza et al., 2018). Our setting is more general than the previous works in two respects. First, apart from the mean square error (MSE) loss L_0, an L_2 regularization term (weight decay) with arbitrary strength is included; second, the noise ϵ is arbitrary. Thus, our setting is arguably closer to the actual deep learning practice, where the injection of noises to latent layers is common, and the use of weight decay is virtually ubiquitous (Krogh and Hertz, 1992; Loshchilov and Hutter, 2017). One major limitation of our work is that we assume the label y to be 1-dimensional, and it can be an important future problem to prove whether an exact solution exists or not when y is high-dimensional.

Our foremost contribution is to prove that all the global minimum of an arbitrarily deep and wide linear net takes a simple analytical form. In other words, we identify in closed form the global minima of Eq. (1) up to a single scalar, whose analytical expression does not exist in general. We then show that it has nontrivial properties that can explain many phenomena in deep learning. In particular, the implications of our result include (but are not limited to):

1. Weight decay makes the landscape of neural nets more complicated;
   - we show that bad minima[2] emerge as weight decay is applied, whereas there is no bad minimum when there is no weight decay. This highlights the need to escape bad local minima in deep learning with weight decay.
2. Deeper nets are harder to optimize than shallower ones;
   - we show that a $D \geq 2$ linear net contains a bad minimum at zero, whereas a $D = 1$ net does not. This partially explains why deep networks are much harder to optimize than shallower ones in deep learning practice.
3. Depending on the task, the common initialization methods (such as the Kaiming init.) can initialize a deep model in the basin of attraction of the bad minimum at zero;
   - common initialization methods initialize the models at a radius of roughly $1/\sqrt{width}$ around the origin; however, we show that the width of the bad minimum is task-dependent and can be larger than the initialization radius for tasks with a small margin ($\|\mathbb{E}[xy]\|$);
4. Thus, the use of (effective) weight decay is a major cause of various types of collapses in deep learning (for example, see Figure 1).

Organization: In the next section, we discuss the related works. In Section 3, we derive the exact solution for a two-layer net. Section 4 extends the result to an arbitrary depth. In Section 5, we study and discuss the relevance of our results to many commonly encountered problems in deep learning. The last section concludes the work and discusses unresolved open problems. All proofs are delayed to Section B. Moreover, additional theoretical results on the effect of including a bias term, the zero weight decay limit, and the zero noise limit are considered in Section D, E, and F, respectively.

[2]Unless otherwise specified, we use the word "bad minimum" to mean a local minimum that is not a global minimum.

Notation. For a matrix W, we use $W_{i:}$ to denote the i-th row vector of W. $\|Z\|$ denotes the L_2 norm if Z is a vector and the Frobenius norm if Z is a matrix. The notation $*$ signals an optimized quantity. Additionally, we use the superscript * and subscript $_*$ interchangeably, whichever leads to a simpler expression. For example, b_*^2 and $(b^*)^2$ denote the same quantity, while the former is "simpler."

2 Related Works

In many ways, linear networks have been used to help understand nonlinear networks. For example, even at depth 0, where the linear net is just a linear regressor, linear nets are shown to be relevant for understanding the generalization behavior of modern overparametrized networks (Hastie et al., 2019). Saxe et al. (2013) studies the training dynamics of a depth-1 network and uses it to understand the dynamics of learning of nonlinear networks. These networks are the same as a linear regression model in terms of expressivity. However, the loss landscape is highly complicated due to the existence of more than one layer, and linear nets are widely believed to approximate the loss landscape of a nonlinear net (Kawaguchi, 2016; Hardt and Ma, 2016; Laurent and Brecht, 2018). In particular, the landscape of linear nets has been studied as early as 1989 in Baldi and Hornik (1989), which proposed the well-known conjecture that all local minima of a deep linear net are global. This conjecture is first proved in Kawaguchi (2016), and extended to other loss functions and deeper depths in Lu and Kawaguchi (2017) and Laurent and Brecht (2018). Many relevant contemporary deep learning problems can be understood with deep linear models. For example, two-layer linear VAE models are used to understand the cause of the posterior collapse problem (Lucas et al., 2019; Wang and Ziyin, 2022). Deep linear nets are also used to understand the neural collapse problem in contrastive learning (Tian, 2022). We also provide more empirical evidence in Section 5.

3 Two-layer Linear Net

This section finds the global minima of a two-layer linear net. The data point is a d-dimensional vector $x \in \mathbb{R}^d$ drawn from an arbitrary distribution and the labels are generated through an arbitrary function $y = y(x) \in \mathbb{R}$. For generality, we let different layers have different strengths of weight decay even though they often take the same value in practice. We want to minimize the following objective:

$$L_{d,d_1}(U,W) = \mathbb{E}_x \mathbb{E}_\epsilon \left(\sum_j^{d_1} U_j \epsilon_j \sum_i^d W_{ji} x_i - y \right)^2 + \gamma_w \|W\|^2 + \gamma_u \|U\|^2, \tag{2}$$

where d_1 is the width of the hidden layer and ϵ_i are independent random variables. $\gamma_w > 0$ and $\gamma_u > 0$ are the weight decay parameters. Here, we consider a general type of noise with $\mathbb{E}[\epsilon_i] = 1$ and $\mathbb{E}[\epsilon_i \epsilon_j] = \delta_{ij}\sigma^2 + 1$ where δ_{ij} is the Kronecker's delta, and $\sigma^2 > 0$. While we formally require γ and σ to nonzero, we discuss the cases of zero limit in Appendix E and F. For shorthand, we use the notation $A_0 := \mathbb{E}[xx^T]$, and the largest and the smallest eigenvalues of A_0 are denoted as $a_{\max}$ and $a_{\min}$ respectively. a_i denotes the i-th eigenvalue of A_0 viewed in any order. For now, it is sufficient for us to assume that the global minimum of Eq. (2) always exists. We will prove a more general result in Proposition 1, when we deal with multilayer nets.

3.1 Main Result

We first present two lemmas showing that the global minimum can only lie on a rather restrictive subspace of all possible parameter settings due to invariances in the objective.

Lemma 1. *At the global minimum of Eq. (2), $U_j^2 = \frac{\gamma_w}{\gamma_u} \sum_i W_{ji}^2$ for all j.*

Proof Sketch. We use the fact that the first term of Eq. (2) is invariant to a simultaneous rescaling of rows of the weight matrix to find the optimal rescaling, which implies the lemma statement. □

This lemma implies that for all j, $|U_j|$ must be proportional to the norm of its corresponding row vector in W. This lemma means that using weight decay makes all layers of a deep neural network balanced. This lemma has been referred to as the "weight balancing" condition in recent works (Tanaka et al., 2020), and, in some sense, is a unique and potentially essential feature of neural networks that encourages a sparse solution (Ziyin and Wang, 2022). The following lemma further shows that, at the global minimum, all elements of U must be equal.

Lemma 2. *At the global minimum, for all i and j, we have*

$$\begin{cases} U_i^2 = U_j^2; \\ U_i W_{i:} = U_j W_{j:}. \end{cases} \tag{3}$$

Proof Sketch. We show that if the condition is not satisfied, then an "averaging" transformation will strictly decrease the objective. □

This lemma can be seen as a formalization of the intuition suggested in the original dropout paper (Srivastava et al., 2014). Namely, using dropout encourages the neurons to be independent of one another and results in an averaging effect. The second lemma imposes strong conditions on the solution of the problem, and the essence of this lemma is the reduction of the original problem to a lower dimension. We are now ready to prove our first main result.

Theorem 1. *The global minimum U_* and W_* of Eq. (2) is $U_* = 0$ and $W_* = 0$ if and only if*

$$\|\mathbb{E}[xy]\|^2 \leq \gamma_u \gamma_w. \tag{4}$$

When $\|\mathbb{E}[xy]\|^2 > \gamma_u \gamma_w$, the global minima are

$$\begin{cases} U_* = b\mathbf{r}; \\ W_* = \mathbf{r}\mathbb{E}[xy]^T b \left[b^2\left(\sigma^2 + d_1\right) A_0 + \gamma_w I\right]^{-1}, \end{cases} \tag{5}$$

where $\mathbf{r} = (\pm 1, ..., \pm 1)$ is an arbitrary vertex of a d_1-dimensional hypercube, and b satisfies:

$$\left\| \left[b^2\left(\sigma^2 + d_1\right) A_0 + \gamma_w I\right]^{-1} \mathbb{E}[xy] \right\|^2 = \frac{\gamma_u}{\gamma_w}. \tag{6}$$

Apparently, $b = 0$ is the trivial solution that has not learned any feature due to overregularization. Henceforth, we refer to this solution (and similar solutions for deeper nets) as the "trivial" solution. We now analyze the properties of the nontrivial solution b^* when it exists.

The condition for the solution to become nontrivial is interesting: $\|\mathbb{E}[xy]\|^2 \geq \gamma_u \gamma_w$. The term $\|\mathbb{E}[xy]\|$ can be seen as the effective strength of the signal, and $\gamma_u \gamma_w$ is the strength of regularization. This precise condition means that the learning of a two-layer can be divided into two qualitatively different regimes: an "overregularized regime" where the global minimum is trivial, and a "feature learning regime" where the global minimum involves actual learning. Lastly, note that our main result does not specify the exact value of b^*. This is because b^* must satisfy the condition in Eq. (6), which is equivalent to a high-order polynomial in b with coefficients being general functions of the eigenvalues of A_0, whose solutions are generally not analytical by Galois theory. One special case where an analytical formula exists for b is when $A_0 = \sigma_x^2 I$. See Section C for more discussion.

3.2 Bounding the General Solution

While the solution to b^* does not admit an analytical form for a general A_0, one can find meaningful lower and upper bounds to b^* such that we can perform an asymptotic analysis of b^*. At the global minimum, the following inequality holds:

$$\begin{aligned} \|\left[b^2\left(\sigma^2 + d_1\right) a_{\max} I + \gamma_w I\right]^{-1} \mathbb{E}[xy]\|^2 &\leq \|[b^2\left(\sigma^2 + d_1\right) A_0 + \gamma_w I]^{-1} \mathbb{E}[xy]\|^2 \\ &\leq \|\left[b^2\left(\sigma^2 + d_1\right) a_{\min} I + \gamma_w I\right]^{-1} \mathbb{E}[xy]\|^2, \end{aligned} \tag{7}$$

where $a_{\min}$ and $a_{\max}$ are the smallest and largest eigenvalue of A_0, respectively. The middle term is equal to γ_u/γ_w by the global minimum condition in (33), and so, assuming $a_{\min} > 0$, this inequality is equivalent to the following inequality of b^*:

$$\frac{\sqrt{\frac{\gamma_w}{\gamma_u}}\|\mathbb{E}[xy]\| - \gamma_w}{(\sigma^2 + d_1) a_{\max}} \leq b_*^2 \leq \frac{\sqrt{\frac{\gamma_w}{\gamma_u}}\|\mathbb{E}[xy]\| - \gamma_w}{(\sigma^2 + d_1) a_{\min}}. \tag{8}$$

Namely, the general solution b^* should scale similarly to the homogeneous solution in Eq. (105) if we treat the eigenvalues of A_0 as constants.

4 Exact Solution for An Arbitrary-Depth Linear Net

This section extends our result to multiple layers. We first derive the analytical formula for the global minimum of a general arbitrary-depth model. We then show that the landscape for a deeper network is highly nontrivial.

4.1 General Solution

The loss function is

$$\mathbb{E}_x \mathbb{E}_{\epsilon^{(1)},\epsilon^{(2)},\ldots,\epsilon^{(D)}} \left(\sum_{i,i_1,i_2,\ldots,i_D}^{d,d_1,d_2,\ldots d_D} U_{i_D} \epsilon_{i_D}^{(D)} \ldots \epsilon_{i_2}^{(2)} W_{i_2 i_1}^{(2)} \epsilon_{i_1}^{(1)} W_{i_1 i}^{(1)} x_i - y \right)^2 + \gamma_u \|U\|^2 + \sum_{i=1}^{D} \gamma_i \|W^{(i)}\|^2, \tag{9}$$

where all the noises ϵ are independent, and for all i and j, $\mathbb{E}[\epsilon_j^{(i)}] = 1$ and $\mathbb{E}[(\epsilon_j^{(i)})^2] = \sigma_i^2 + 1 > 1$. We first show that for general D, the global minimum exists for this objective.

Proposition 1. *For $D \geq 1$ and strictly positive $\gamma_u,\ \gamma_1, \ldots, \gamma_D$, the global minimum for Eq.(9) exists.*

Note that the positivity of the regularization strength is crucial. If one of the γ_i is zero, the global minimum may not exist. The following theorem is our second main result.

Theorem 2. *Any global minimum of Eq. (9) is of the form*

$$\begin{cases} U = b_u \mathbf{r}_D; \\ W^{(i)} = b_i \mathbf{r}_i \mathbf{r}_{i-1}^T; \\ W^{(1)} = \mathbf{r}_1 \mathbb{E}[xy]^T (b_u \prod_{i=2}^D b_i) \mu \left[(b_u \prod_{i=2}^D b_i)^2 s^2 \left(\sigma^2 + d_1\right) A_0 + \gamma_w I \right]^{-1}, \end{cases} \tag{10}$$

where $\mu = \prod_{i=2}^D d_i$, $s^2 = \prod_{i=2}^D d_i(\sigma^2 + d_i)$, $b_u \geq 0$ and $b_i \geq 0$, and $\mathbf{r}_i = (\pm 1, \ldots, \pm 1)$ is an arbitrary vertex of a d_i-dimensional hypercube for all i. Furthermore, let $b_1 := \sqrt{\|W_{i:}\|^2/d}$ and $b_{D+1} := b_u$, b_i satisfies

$$\gamma_{k+1} d_{k+1} b_{k+1}^2 = \gamma_k d_{k-1} b_k^2. \tag{11}$$

Proof Sketch. We prove by induction on the depth D. The base case is proved in Theorem 1. We then show that for a general depth, the objective involves optimizing subproblems, one of which is a $D-1$ layer problem that follows by the induction assumption, and the other is a two-layer problem that has been solved in Theorem 1. Putting these two subproblems together, one obtains Eq. (10). □

Remark. *We deal with the technical case of having a bias term for each layer in Appendix D. For example, we will show that if one has preprocessed the data such that $\mathbb{E}[x] = 0$ and $\mathbb{E}[y] = 0$, our main results remain precisely unchanged.*

The condition in Eq. (11) shows that the scaling factor b_i for all i is not independent of one another. This automatic balancing of the norm of all layers is a consequence of the rescaling invariance of the multilayer architecture and the use of weight decay. It is well-known that this rescaling invariance also exists in a neural network with the ReLU activation, and so this balancing condition is also directly relevant for ReLU networks.

Condition (11) implies that all the b_i can be written in terms of one of the b_i:

$$b_u \prod_{i=2}^D b_i = c_0 \mathrm{sgn}\left(b_u \prod_{i=2}^D b_i\right) |b_2^D| := c_0 \mathrm{sgn}\left(b_u \prod_{i=2}^D b_i\right) b^D \tag{12}$$

where $c_0 = \frac{(\gamma_2 d_2 d_1)^{D/2}}{\sqrt{\gamma_u \prod_{i=2}^D \gamma_i} \prod_{i=2}^D d_i \sqrt{d_1}}$ and $b \geq 0$. Consider the first layer ($i = 1$), Eq (11) shows that the global minimum must satisfy the following equation, which is equivalent to a high-order polynomial in b that does not have an analytical solution in general:

$$\|\mathbb{E}[xy]^T c_0 b^D \mu \left[c_0^2 b^{2D} s^2 \left(\sigma^2 + d_1\right) A_0 + \gamma_w I \right]^{-1} \|^2 = d_2 b^2. \tag{13}$$

Thus, this condition is an extension of the condition (6) for two-layer networks.

At this point, it pays to clearly define the word "solution," especially given that it has a special meaning in this work because it now becomes highly nontrivial to differentiate between the two types of solutions.

Definition 1. *We say that a non-negative real b is a solution if it satisfies Eq. (13). A solution is trivial if $b = 0$ and nontrivial otherwise.*

Namely, a global minimum must be a solution, but a solution is not necessarily a global minimum. We have seen that even in the two-layer case, the global minimum can be the trivial one when the strength of the signal is too weak or when the strength of regularization is too strong. It is thus natural to expect 0 to be the global minimum under a similar condition, and one is interested in whether the condition becomes stronger or weaker as the depth of the model is increased. However, it turns out this naive expectation is not true. In fact, when the depth of the model is larger than 2, the condition for the trivial global minimum becomes highly nontrivial.

The following proposition shows why the problem becomes more complicated. In particular, we have seen that in the case of a two-layer net, some elementary argument has helped us show that the trivial solution $b = 0$ is either a saddle or the global minimum. However, the proposition below shows that with $D \geq 2$, the landscape becomes more complicated in the sense that the trivial solution is always a local minimum, and it becomes difficult to compare the loss value of the trivial solution with the nontrivial solution because the value of b^* is unknown in general.

Proposition 2. *Let $D \geq 2$ in Eq. (9). Then, the solution $U = 0$, $W^{(D)} = 0$, ..., $W^{(1)} = 0$ is a local minimum with a diagonal positive-definite Hessian γI.*

Comparing the Hessian of $D \geq 2$ and $D = 1$, one notices a qualitative difference: for $D \geq 2$, the Hessian is always diagonal (at 0); for $D = 1$, in sharp contrast, the off-diagonal terms are nonzero in general, and it is these off-diagonal terms that can break the positive-definiteness of the Hessian. This offers a different perspective on why there is a qualitative difference between $D = 1$ and $D = 2$.

Lastly, note that, unlike the depth-1 case, one can no longer find a precise condition such that a $b \neq 0$ solution exists for a general A_0. The reason is that the condition for the existence of the solution is now a high-order polynomial with quite arbitrary intermediate terms. The following proposition gives a sufficient but stronger-than-necessary condition for the existence of a nontrivial solution, when all the σ_i, intermediate width d_i and regularization strength γ_i are the same.[3]

Proposition 3. *Let $\sigma_i^2 = \sigma^2 > 0$, $d_i = d_0$ and $\gamma_i = \gamma > 0$ for all i. Assuming $a_{\min} > 0$, the only solution is trivial if*

$$\frac{D+1}{2D}||\mathbb{E}[xy]||d_0^{D-1}\left(\frac{(D-1)||\mathbb{E}[xy]||}{2Dd_0(\sigma^2+d_0)^D a_{\min}}\right)^{\frac{D-1}{D+1}} < \gamma. \tag{14}$$

Nontrivial solutions exist if

$$\frac{D+1}{2D}||\mathbb{E}[xy]||d_0^{D-1}\left(\frac{(D-1)||\mathbb{E}[xy]||}{2Dd_0(\sigma^2+d_0)^D a_{\max}}\right)^{\frac{D-1}{D+1}} \geq \gamma. \tag{15}$$

Moreover, the nontrivial solutions are both lower and upper-bounded:[4]

$$\frac{1}{d_0}\left[\frac{\gamma}{||\mathbb{E}[xy]||}\right]^{\frac{1}{D-1}} \leq b^* \leq \left[\frac{||\mathbb{E}[xy]||}{d_0(\sigma^2+d_0)^D a_{\max}}\right]^{\frac{1}{D+1}}. \tag{16}$$

Proof Sketch. The proof follows from the observation that the l.h.s. of Eq. (13) is a continuous function and must cross the r.h.s. under certain sufficient conditions. □

One should compare the general condition here with the special condition for $D = 1$. One sees that for $D \geq 2$, many other factors (such as the width, the depth, and the spectrum of the data covariance A_0) come into play to determine the existence of a solution apart from the signal strength $\mathbb{E}[xy]$ and the regularization strength γ.

[3] This is equivalent to setting $c_0 = \sqrt{d_0}$. The result is qualitatively similar but involves additional factors of c_0 if σ_i, d_i, and γ_i all take different values. We thus only present the case when σ_i, d_i, and γ_i are the same for notational concision and for emphasizing the most relevant terms. Also, note that this proposition gives a *sufficient and necessary* condition if $A_0 = \sigma_x^2 I$ is proportional to the identity.

[4] For $D = 1$, we define the lower-bound as $\lim_{\eta\to 0^+}\lim_{D\to 1^+}\frac{1}{d_0}\left[\frac{\gamma+\eta}{||\mathbb{E}[xy]||}\right]^{\frac{1}{D-1}}$, which equal to zero if $\mathbb{E}[xy] \geq \gamma$, and ∞ if $\mathbb{E}[xy] < \gamma$. With this definition, this proposition applies to a two-layer net as well.

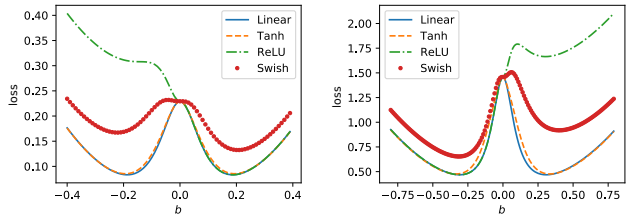

Figure 2: The training loss as a function of b for a $D = 1$ network with different activation functions in the hidden layer. For simplicity, dropout is not implemented. The non-linear activation functions we considered are ReLU, Tanh, and Swish. The left and right panels use different data. **Left:** X are Gaussian random vectors, and $y = v \cdot x$ is a linear function of x. **Right:** x are Gaussian random vectors, and $y = v \cdot tanh(x)$ are nonlinear functions of data; the weight v is obtained as a Gaussian random vector.

4.2 Which Solution is the Global Minimum?

Again, we set $\gamma_i = \gamma > 0$, $\sigma_i^2 = \sigma^2 > 0$ and $d_i = d_0 > 0$ for all i for notational concision. Using this condition and applying Lemma 3 to Theorem 2, the solution now takes the following form, where $b \geq 0$,

$$\begin{cases} U = \sqrt{d_0} b \mathbf{r}_D; \\ W^{(i)} = b \mathbf{r}_i \mathbf{r}_{i-1}^T; \\ W^{(1)} = \mathbf{r}_1 \mathbb{E}[xy]^T d_0^{D-\frac{1}{2}} b^D \left[d_0^D (\sigma^2 + d_0)^D b^{2D} A_0 + \gamma \right]^{-1}. \end{cases} \tag{17}$$

The following theorem gives a sufficient condition for the global minimum to be nontrivial. It also shows that the landscape of the linear net becomes complicated and can contain more than 1 local minimum.

Theorem 3. *Let $\sigma_i^2 = \sigma^2 > 0$, $d_i = d_0$ and $\gamma_i = \gamma > 0$ for all i and assuming $a_{\min} > 0$. Then, if*

$$\|\mathbb{E}[xy]\|^2 \geq \frac{\gamma^{\frac{D+1}{D}} D^2 (\sigma^2 + d_0)^{D-1} a_{max}^{\frac{D-1}{D}}}{d_0^{D-1} (D-1)^{\frac{D-1}{D}}} \tag{18}$$

the global minimum of Eq. (9) is one of the nontrivial solutions.

While there are various ways this bound can be improved, it is general enough for our purpose. In particular, one sees that, for a general depth, the condition for having a nontrivial global minimum depends not only on the $\mathbb{E}[xy]$ and γ but also on the model architecture in general. For a more general architecture with different widths etc., the architectural constant c_0 from Eq. (13) will also enter the equation. In the limit of $D \to 1^+$, relation (18) reduces to

$$\|\mathbb{E}[xy]\|^2 \geq \gamma^2, \tag{19}$$

which is the condition derived for the 2-layer case.

5 Implications

Relevance to nonlinear models. We first caution the readers that the following discussion should be taken with a caveat and is based on the philosophy that deep linear nets can approximate the nonlinear ones. This approximation certainly holds for fully connected models with differentiable activation functions such as tanh or Swish because they are, up to first-order Taylor expansion, a deep linear net around zero, which is the region for which our theory is the most relevant. We empirically demonstrate that close to the origin, the landscape of linear nets can indeed approximate that of nonlinear nets quite well. To compare, we plug in the solution in Theorem 4 to both linear and nonlinear models of the same architecture and compare the loss values at different values of b around $b = 0$. For simplicity, we only consider the case $D = 1$. The activation functions we consider are ReLU, Tanh, and Swish (Ramachandran et al., 2017), a modern and differentiable variant of ReLU. See Fig. 2.

The regressor $x \in \mathbb{R}^d$ is sampled as Gaussian random vectors. We consider two methods of generating y; the first one (left) is $y = v \cdot x$. The second one (right) is $y = v \cdot \tanh(x)$, where the weight $v \in \mathbb{R}^d$ is obtained as a Gaussian random vector. Fig. 2 shows that the landscape consisting of Tanh is always close to the linear landscape. Swish is not as good as Tanh, but the Swish landscape shows a similar tendency to the linear landscape. The ReLU landscape is not so close to the linear landscape

either for $b > 0$ or $b < 0$, but it agrees completely with the linear landscape on the other side, as expected. Besides the quantitative closeness, it is also important to note that all the landscapes agree qualitatively, containing the same number of local minima at similar values of b.

Landscape of multi-layer neural networks. The combination of Theorem 3 and Proposition 2 shows that the landscape of a deep neural network can become highly nontrivial when there is a weight decay and when the depth of the model is larger than 2. This gives an incomplete but meaningful picture of a network's complicated but interesting landscape beyond two layers (see Figure 1 for an incomplete summary of our results). In particular, even when the nontrivial solution is the global minimum, the trivial solution is still a local minimum that needs to be escaped. Our result suggests the previous understanding that all local minima of a deep linear net are global cannot generalize to many practical settings where deep learning is found to work well. For example, a series of works attribute the existence of bad (non-global) minima to the use of nonlinearities nonlinear nets (Kawaguchi, 2016) or the use of a non-regular (non-differentiable) loss function (Laurent and Brecht, 2018). Our result, in contrast, shows that the use of a simple weight decay is sufficient to create a bad minimum.[5] Moreover, the problem with such a minimum is two-fold: (1) (optimization) it is not global and so needs to be "overcome" and (2) (generalization) it is a minimum that has not learned any feature at all because the model constantly outputs zero. To the best of our knowledge, previous to our work, there has not been any proof that a bad minimum can generically exist in a rather arbitrary network without any restriction on the data.[6] Thus, our result offers direct and solid theoretical justification for the widely believed importance of escaping local minima in the field of deep learning (Kleinberg et al., 2018; Liu et al., 2021; Mori et al., 2022). In particular, previous works on escaping local minima often hypothesize landscapes that are of unknown relevance to an actual neural network. With our result, this line of research can now be established with respect to landscapes that are actually deep-learning-relevant.

Previous works also argue that having a deeper depth does not create a bad minimum (Lu and Kawaguchi, 2017). While this remains true, its generality and applicability to practical settings now also seem low. Our result shows that as long as weight decay is used, and as long as $D \geq 2$, there is indeed a bad local minimum at 0. In contrast, there is no bad minimum at 0 for a depth-2 network: the point $b = 0$ is either a saddle or the global minimum.[7] Having a deeper depth thus alters the qualitative nature of the landscape, and our results agree better with the common observation that a deeper network is harder, if not impossible, to optimize.

We note that our result can also be relevant for more modern architectures such as the ResNet. 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. See Figure 1. The experimental details are given in Section A.

Learnability of a neural network. Now we analyze the solution when D tends to infinity. We first note that the existence condition bound in (15) becomes exponentially harder to satisfy as D becomes large:

$$\|\mathbb{E}[xy]\|^2 \geq 4d_0^2 a_{\max}\gamma e^{D\log[(\sigma^2+d_0)/d_0]} + O(1). \tag{20}$$

When this bound is not satisfied, the given neural network cannot learn the data. Recall that for a two-layer net, the existence condition is nothing but $\|\mathbb{E}[xy]\|^2 > \gamma^2$, independent of the depth, width, or stochasticity in the model. For a deeper network, however, every factor comes into play, and the

[5] Some previous works do suggest the existence of bad minima when weight decay is present, but no direct proof exists yet. For example, Taghvaei et al. (2017) shows that when the model is approximated by a linear dynamical system, regularization can cause bad local minima. Mehta et al. (2021) shows the existence of bad local minima in deep linear networks with weight decay through numerical simulations.

[6] In the case of nonlinear networks without regularization, a few works proved the existence of bad minima. However, the previous results strongly depend on the data and are rather independent of architecture. For example, one major assumption is that the data cannot be perfected and fitted by a linear model (Yun et al., 2018; Liu, 2021; He et al., 2020). Some other works explicitly construct data distribution (Safran and Shamir, 2018; Venturi et al., 2019). Our result, in contrast, is independent of the data.

[7] Of course, in practice, the model trained with SGD can still converge to the trivial solution even if it is a saddle point (Ziyin et al., 2021) because minibatch SGD is, in general, not a good estimator of the local minima.

architecture of the model has a strong (and dominant) influence on the condition. In particular, a factor that increases polynomially in the model width and exponentially in the model depth appears.

A practical implication is that the use of weight decay may be too strong for deep networks. If one increases the depth or width of the model, one should also roughly decrease γ according to Eq. (20).

Insufficiency of the existing initialization schemes. We have shown that 0 is often a bad local minimum for deep learning. Our result further implies that escaping this local minimum can be highly practically relevant because standard initialization schemes are trapped in this local minimum for tasks where the signal $\mathbb{E}[xy]$ is weak. See Inequality (16): any nontrivial global minimum is lower-bounded by a factor proportional to $(\gamma/\|\mathbb{E}[xy]\|^{1/(D-1)})/d_0$, which can be seen as an approximation of the radius of the local minimum at the origin. In comparison, standard deep learning initialization schemes such as Kaiming init. initialize at a radius roughly $1/\sqrt{d_0}$. Thus, for tasks $\mathbb{E}[xy] \ll \gamma/\sqrt{d_0}$, these initialization methods are likely to initialize the model in the basin of attraction of the trivial regime, which can cause a serious failure in learning. To demonstrate, we perform a numerical simulation shown in the right panel of Figure 3, where we train $D = 2$ nonlinear networks with width 32 with SGD on tasks with varying $\|\mathbb{E}[xy]\|$. For sufficiently small $\|\mathbb{E}[xy]\|$, the model clearly is stuck at the origin.[8] In contrast, linear regression is never stuck at the origin. Our result thus suggests that it may be desirable to devise initialization methods that are functions of the data distribution.

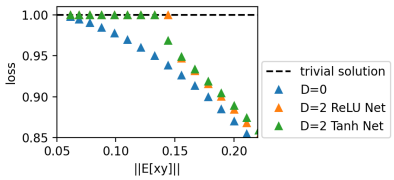

Figure 3: Training loss of $D = 2$ neural networks with ReLU and Tanh activations across synthetic tasks with different $\|\mathbb{E}[xy]\|$. We see that with the Kaiming initialization, both the Tanh net and the ReLU net are stuck at the trivial solution in expectation of our theory. In contrast, an optimized linear regressor ($D = 0$) is better than the trivial solution when $\|\mathbb{E}[xy]\| > 0$. See Section A for experimental details.

Prediction variance of stochastic nets. A major extension of the standard neural networks is to make them stochastic, namely, to make the output a random function of the input. In a broad sense, stochastic neural networks include neural networks trained with dropout (Srivastava et al., 2014; Gal and Ghahramani, 2016), Bayesian networks (Mackay, 1992), variational autoencoders (VAE) (Kingma and Welling, 2013), and generative adversarial networks (Goodfellow et al., 2014). Stochastic networks are thus of both practical and theoretical importance to study. Our result can also be used for studying the theoretical properties of stochastic neural networks. Here, we present a simple application of our general solution to analyze the properties of a stochastic net. The following theorem summarizes our technical results.

Theorem 4. *Let $\sigma_i^2 = \sigma^2 > 0$, $d_i = d_0$ and $\gamma_i = \gamma > 0$ for all i. Let $A_0 = \sigma_x^2 I$. Then, at any global minimum of Eq. (9), holding other parameters fixed,*

1. *in the limit of large d_0, $Var[f(x)] = O\left(d_0^{-1}\right)$;*
2. *in the limit of large σ^2, $Var[f(x)] = O\left(\frac{1}{(\sigma^2)^D}\right)$;*
3. *In the limit of large D, $Var[f(x)] = O\left(e^{-2D\log[(\sigma^2+d_0)/d_0]}\right)$.*

Interestingly, the scaling of prediction variance in asymptotic σ^2 is different for different widths. The third result shows that the prediction variance decreases exponentially fast in D. In particular, this result answers a question recently proposed in Ziyin et al. (2022b): does a stochastic net trained on MSE have a prediction variance that scales towards 0? We improve on their result in the case of a deep linear net by (a) showing that the d_0^{-1} is tight in general, independent of the depth or other factors of the model, and (b) proving a bound showing that the variance also scales towards zero as depth increases, which is a novel result of our work. Our result also offers an important insight into the cause of the vanishing prediction variance. Previous works (Alemi et al., 2018) often attribute the cause to the fact that a wide neural network is too expressive. However, our result implies that this is not always the case because a linear network with limited expressivity can also have a vanishing variance as the model tends to an infinite width.

Collapses in deep learning. Lastly, our result also sheds light on the well-known problem of posterior collapse in Bayesian deep learning, the neural collapse problem in supervised learning,

[8] There are many natural problems where the signal is extremely weak. One well-known example is the problem of future price prediction in finance, where the fundamental theorem of finance forbids a large $\|\mathbb{E}[xy]\|$ (Fama, 1970).

and the dimensional collapse contrastive learning. For neural collapse, our result agrees with the recent works that identify weight decay as a main cause (Rangamani and Banburski-Fahey, 2022; Rangamani et al., 2021). Our result thus suggests that neural collapse for shallow and deeper models can behave quite differently. For Bayesian deep learning, Wang and Ziyin (2022) identified the cause of the posterior collapse in a two-layer VAE structure to be that the regularization of the mean of the latent variable z is too strong, which then causes a change in the stability of Hessian matrix at the origin. More recently, the origin and its stability have also been identified as a crucial feature for the dimensional collapse in self-supervised learning (Ziyin et al., 2022a).

Although appearing in different contexts of deep learning, the three types of collapses share the same phenomenology that the model converges to a "collapsed" regime where the learned representation becomes a constant, which is in agreement with the behavior of the trivial regime we identified. Our result thus shows that these problems may share a common cause, which is a unique feature of deep learning. One important future step is to investigate this connection in greater detail.

6 Conclusion

In this work, we derived the exact solution of a deep linear net with arbitrary depth and width and with stochasticity. Our work sheds light on the highly complicated landscape of a deep neural network. Compared to the previous works that mostly focus on the qualitative understanding of the linear net, our result offers a more precise quantitative understanding of deep linear nets. Quantitative understanding is one major benefit of knowing the exact solution, whose usefulness we have also demonstrated with the various implications. The results, although derived for linear models, are also empirically shown to be relevant for networks with nonlinear activations. Lastly, our results strengthen the line of thought that analytical approaches to deep linear models can be used to understand deep neural networks, and it is the sincere hope of the authors to attract more attention to this promising field.

Acknowledgement

Ziyin is financially supported by the GSS scholarship of the University of Tokyo and the JSPS fellowship. Li is financially supported by CNRS. X. Meng is supported by JST CREST Grant Number JPMJCR1912, Japan.

References

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.

Baldi, P. and Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58.

Cavazza, J., Morerio, P., Haeffele, B., Lane, C., Murino, V., and Vidal, R. (2018). Dropout as a low-rank regularizer for matrix factorization. In *International Conference on Artificial Intelligence and Statistics*, pages 435–444. PMLR.

Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., and LeCun, Y. (2015a). The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, pages 192–204.

Choromanska, A., LeCun, Y., and Arous, G. B. (2015b). Open problem: The landscape of the loss surfaces of multilayer networks. In *Conference on Learning Theory*, pages 1756–1760. PMLR.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, 25(2):383–417.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Gotmare, A., Keskar, N. S., Xiong, C., and Socher, R. (2018). A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. *arXiv preprint arXiv:1810.13243*.

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

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

He, F., Wang, B., and Tao, D. (2020). Piecewise linear activations substantially shape the loss surfaces of neural networks. *arXiv preprint arXiv:2003.12236*.

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

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kleinberg, B., Li, Y., and Yuan, Y. (2018). An alternative view: When does sgd escape local minima? In *International Conference on Machine Learning*, pages 2698–2707. PMLR.

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

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.

Liu, B. (2021). Spurious local minima are common for deep neural networks with piecewise linear activations. *arXiv preprint arXiv:2102.13233*.

Liu, K., Ziyin, L., and Ueda, M. (2021). Noise and fluctuation of finite learning rate stochastic gradient descent.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lu, H. and Kawaguchi, K. (2017). Depth creates no bad local minima. *arXiv preprint arXiv:1702.08580*.

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

Mackay, D. J. C. (1992). *Bayesian methods for adaptive models*. PhD thesis, California Institute of Technology.

Mehta, D., Chen, T., Tang, T., and Hauenstein, J. (2021). The loss surface of deep linear networks viewed through the algebraic geometry lens. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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

Mori, T., Ziyin, L., Liu, K., and Ueda, M. (2022). Power-law escape rate of SGD.

Ramachandran, P., Zoph, B., and Le, Q. V. (2017). Searching for activation functions.

Rangamani, A. and Banburski-Fahey, A. (2022). 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)*, pages 4243–4247. IEEE.

Rangamani, A., Xu, M., Banburski, A., Liao, Q., and Poggio, T. (2021). Dynamics and neural collapse in deep classifiers trained with the square loss. *Center for Brains, Minds and Machines (CBMM) Memo*, (117).

Safran, I. and Shamir, O. (2018). Spurious local minima are common in two-layer relu neural networks. In *International conference on machine learning*, pages 4433–4441. PMLR.

Saxe, A. M., McClelland, J. L., and Ganguli, S. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Taghvaei, A., Kim, J. W., and Mehta, P. (2017). How regularization affects the critical points in linear networks. *Advances in neural information processing systems*, 30.

Tanaka, H., Kunin, D., Yamins, D. L., and Ganguli, S. (2020). Pruning neural networks without any data by iteratively conserving synaptic flow. *Advances in Neural Information Processing Systems*, 33:6377–6389.

Tian, Y. (2022). Deep contrastive learning is provably (almost) principal component analysis. *arXiv preprint arXiv:2201.12680*.

Venturi, L., Bandeira, A. S., and Bruna, J. (2019). Spurious valleys in one-hidden-layer neural network optimization landscapes. *Journal of Machine Learning Research*, 20:133.

Wang, Z. and Ziyin, L. (2022). Posterior collapse of a linear latent variable model. *arXiv preprint arXiv:2205.04009*.

Yun, C., Sra, S., and Jadbabaie, A. (2018). Small nonlinearities in activation functions create bad local minima in neural networks. *arXiv preprint arXiv:1802.03487*.

Ziyin, L., Li, B., Simon, J. B., and Ueda, M. (2021). SGD with a Constant Large Learning Rate Can Converge to Local Maxima.

Ziyin, L., Lubana, E. S., Ueda, M., and Tanaka, H. (2022a). What shapes the loss landscape of self-supervised learning? *arXiv preprint arXiv:2210.00638*.

Ziyin, L. and Wang, Z. (2022). Sparsity by Redundancy: Solving L_1 with a Simple Reparametrization. *arXiv preprint arXiv:2210.01212*.

Ziyin, L., Zhang, H., Meng, X., Lu, Y., Xing, E., and Ueda, M. (2022b). Stochastic neural networks with infinite width are deterministic.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] The experiments are only for demonstration and are straightforward to reproduce following the theory.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] The fluctuations are visually negligible.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] They are done on a single 3080Ti GPU.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [N/A]

   (b) Did you mention the license of the assets? [N/A]

   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Experimental Details

For the experiment in Figure 3, the input data consists of 1000 data points sampled from a multivariate Gaussian distribution: $x \sim \mathcal{N}(0, I_5)$. The target is generated by a linear transformation $y = v \cdot x$, where the norm of v is rescaled to obtain different values of $\|\mathbb{E}[xy]\|$ as the control parameter of the simulation. The models are with $D = 2$ neural networks with bias terms and with hidden width 32 for both hidden layers. The training proceeds with gradient descent with a learning rate of 0.1 for 10^4 iterations when the training loss has stopped decreasing for all the experiments.

For the CIFAR10 experiments, we train a standard ResNet18 with roughly 10^7 parameters under the standard procedure, with a batch size of 256 for 100 epochs.[9] For the linear models, we use a hidden width of 32 without any bias term. The training proceeds with SGD with batch size 256 for 100 epochs with a momentum of 0.9. The learning rate is 0.002, chosen as the best learning rate from a grid search over $[0.001, 0.002, ..., 0.01]$.

B Proofs

B.1 Proof of Lemma 1

Proof. Note that the first term in the loss function is invariant to the following rescaling for any $a > 0$:

$$\begin{cases} U_i \to aU_i; \\ W_{ij} \to W_{ij}/a; \end{cases} \tag{21}$$

meanwhile, the L_2 regularization term changes as a changes. Therefore, the global minimum must have a minimized a with respect to any U and W.

One can easily find the solution:

$$a^* = \arg\min_a \left(\gamma_u a^2 U_i^2 + \gamma_w \sum_j \frac{W_{ij}^2}{a^2} \right) = \left(\frac{\gamma_w \sum_j W_{ij}^2}{\gamma_u U_i^2} \right)^{1/4}. \tag{22}$$

Therefore, at the global minimum, we must have $\gamma_u a^2 U_i^2 = \gamma_w \sum_j \frac{W_{ij}^2}{a^2}$, so that

$$(U_i^*)^2 = (a^* U_i)^2 = \frac{\gamma_w}{\gamma_u} \sum_j (W_{ij}^*)^2, \tag{23}$$

which completes the proof. □

B.2 Proof of Lemma 2

Proof. By Lemma 1, we can write U_i as b_i and $W_{i:}$ as $b_i w_i$ where w_i is a unit vector, and finding the global minimizer of Eq. (2) is equivalent to finding the minimizer of the following objective,

$$\mathbb{E}_{x,\varepsilon} \left[\left(\sum_{i,j} b_i^2 \epsilon_i w_{ij} x_j - y \right)^2 \right] + (\gamma_u + \gamma_w) \|b\|_2^2, \tag{24}$$

$$= \mathbb{E}_x \left[\left(\sum_{i,j} b_i^2 w_{ij} x_j - y \right)^2 \right] + \sigma^2 \sum_{ij} b_i^4 \left(\sum_k w_{ik} x_k \right)^2 + (\gamma_u + \gamma_w) \|b\|_2^2, \tag{25}$$

The lemma statement is equivalent to $b_i = b_j$ for all i and j.

We prove this by contradiction. Suppose there exist i and j such that $b_i \neq b_j$, we can choose i to be the index of b_i with maximum b_i^2, and let j be the index of b_j with minimum b_j^2. Now, we can construct a different solution by the following replacement of $b_i w_{i:}$ and $b_j w_{j:}$:

$$\begin{cases} b_i^2 w_{i:} \to c^2 v; \\ b_j^2 w_{j:} \to c^2 v, \end{cases} \tag{26}$$

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

where c is a positive scalar and v is a unit vector such that $2c^2 v = b_i^2 w_{i:} + b_j^2 w_{j:}$. Note that, by the triangular inequality, $2c^2 \le b_i^2 + b_j^2$. Meanwhile, all the other terms, b_k for $k \neq i$ and $k \neq j$, are left unchanged. This transformation leaves the first term in the loss function (25) unchanged, and we now show that it decreases the other terms.

The change in the second term is

$$\left(b_i^2 \sum_k w_{ik} x_k\right)^2 + \left(b_j^2 \sum_k w_{jk} x_k\right)^2 \to 2\left(c^2 \sum_k v_k x_k\right)^2 = \frac{1}{2}\left(b_i^2 \sum_k w_{ik} x_k + b_j^2 \sum_k w_{jk} x_k\right)^2. \quad (27)$$

By the inequality $a^2 + b^2 \ge (a+b)^2/2$, we see that the left-hand side the larger than the right-hand side.

We now consider the L_2 regularization term. The change is

$$(\gamma_u + \gamma_w)(b_i^2 + b_j^2) \to 2(\gamma_u + \gamma_w)c^2, \quad (28)$$

and the left-hand side is again larger than the right-hand side by the inequality mentioned above: $2c^2 \le b_i^2 + b_j^2$. Therefore, we have constructed a solution whose loss is strictly smaller than that of the global minimum: a contradiction. Thus, the global minimum must satisfy

$$U_i^2 = U_j^2 \quad (29)$$

for all i and j.

Likewise, we can show that $U_i W_{i:} = U_j W_{j:}$ for all i and j. This is because the triangular inequality $2c^2 \le b_i^2 + b_j^2$ is only an equality if $U_i W_{i:} = U_j W_{j:}$. If $U_i W_{i:} \neq U_j W_{j:}$, following the same argument above, we arrive at another contradiction. □

B.3 Proof of Theorem 1

Proof. By Lemma 2, at any global minimum, we can write $U_* = b\mathbf{r}$ for some $b \in \mathbb{R}$. We can also write $W_* = \mathbf{r}v^T$ for a general vector $v \in \mathbb{R}^d$. Without loss of generality, we assume that $b > 0$ (because the sign of b can be absorbed into $\mathbf{r}$).

The original problem in Eq. (2) is now equivalently reduced following problem because $\mathbf{r}^T\mathbf{r} = d_1$:

$$\min_{b,v} \mathbb{E}_x\left[\left(bd_1 \sum_j v_j x_j - y\right)^2 + b^2 d_1 \sigma^2 \left(\sum_k v_k x_k\right)^2\right] + \gamma_u d_1 b^2 + \gamma_w d_1 \|v\|_2^2. \quad (30)$$

For any fixed b, the global minimum of v is well known:[10]

$$v = b\mathbb{E}[xy]^T \left[b^2\left(\sigma^2 + d_1\right)A_0 + \gamma_w I\right]^{-1}. \quad (31)$$

By Lemma 1, at a global minimum, b also satisfies the following condition:

$$b^2 = \frac{\gamma_w}{\gamma_u}\|v\|^2, \quad (32)$$

One solution to this equation is $b = 0$, and we are interested in whether solutions with $b \neq 0$ exist. If there is no other solution, then $b = 0$ must be the unique global minimum; otherwise, we need to identify which of the solutions are actual global minima. When $b \neq 0$,

$$\left\|\left[b^2\left(\sigma^2 + d_1\right)A_0 + \gamma_w I\right]^{-1}\mathbb{E}[xy]\right\|^2 = \frac{\gamma_u}{\gamma_w}. \quad (33)$$

Note that the left-hand side is monotonically decreasing in b^2, and is equal to $\gamma_w^{-2}\|\mathbb{E}[xy]\|^2$ when $b = 0$. When $b \to \infty$, the left-hand side tends to 0. Because the left-hand side is a continuous and monotonic function of b, a unique solution $b_* > 0$ that satisfies Eq. (33) exists if and only if $\gamma_w^{-2}\|\mathbb{E}[xy]\|^2 > \gamma_u/\gamma_w$, or,

$$\|\mathbb{E}[xy]\|^2 > \gamma_u \gamma_w. \quad (34)$$

[10] Namely, it is the solution of a ridgeless linear regression problem.

Therefore, at most, three candidates for global minima of the loss function exist:

$$\begin{cases} b = 0,\ v = 0 & \text{if } \|\mathbb{E}[xy]\|^2 \le \gamma_u\gamma_w; \\ b = \pm b_*,\ v = b\left[b^2\left(\sigma^2 + d_1\right)A_0 + \gamma_w I\right]^{-1}\mathbb{E}[xy], & \text{if } \|\mathbb{E}[xy]\|^2 > \gamma_u\gamma_w, \end{cases} \tag{35}$$

where $b^* > 0$.

In the second case, one needs to discern the saddle points from the global minima. Using the expression of v, one finds the expression of the loss function as a function of b

$$d_1(d_1+\sigma^2)b^4\sum_i \frac{\mathbb{E}[x'y]_i^2 a_i}{[b^2(\sigma^2+d_1)a_i+\gamma_w]^2} - 2b^2 d_1 \sum_i \frac{\mathbb{E}[x'y]_i^2}{b^2(\sigma^2+d_1)a_i+\gamma_w} + \mathbb{E}[y^2] + \gamma_u d_1 b^2 + \gamma_w d_1 \sum_i \frac{\mathbb{E}[x'y]_i^2 b^2}{[b^2(\sigma^2+d_1)a_i+\gamma_w]^2}, \tag{36}$$

where $x' = Rx$ such that RA_0R^{-1} is a diagonal matrix. We now show that condition (34) is sufficient to guarantee that 0 is not the global minimum.

At $b = 0$, the first nonvanishing derivative of b is the second-order derivative. The second order derivative at $b = 0$ is

$$-2d_1\|\mathbb{E}[xy]\|^2/\gamma_w + 2\gamma_u d_1, \tag{37}$$

which is negative if and only if $\|\mathbb{E}[xy]\|^2 > \gamma_u\gamma_w$. If the second derivative at $b = 0$ is negative, $b = 0$ cannot be a minimum. It then follows that for $\|\mathbb{E}[xy]\|^2 > \gamma_u\gamma_w$, $b = \pm b^*$, $v = b\left[b^2\left(\sigma^2+d_1\right)A_0 + \gamma_w I\right]^{-1}\mathbb{E}[xy]$, if $\|\mathbb{E}[xy]\|^2 > \gamma_u\gamma_w$ are the two global minimum (because the loss is invariant to the sign flip of b). For the same reason, when $\|\mathbb{E}[xy]\|^2 < \gamma_u\gamma_w$, $b = 0$ gives the unique global minimum. This finishes the proof. □

B.4 Proof of Proposition 1

Proof. We first show that there exists a constant r such that the global minimum must be confined within a (closed) r-Ball around the origin. The objective (9) can be upper-bounded by

$$Eq.\ (9) \ge \gamma_u\|U\|^2 + \sum_{i=1}^{D}\gamma_i\|W^{(i)}\|^2 \ge \gamma_{\min}\left(\|U\|^2 + \sum_i \|W^{(i)}\|^2\right), \tag{38}$$

where $\gamma_{\min} := \min_{i\in\{u,1,2,\dots,D\}} > 0$. Now, let w denote be the union of all the parameters $(U, W^{(i)})$ and viewed as a vector. We see that the above inequality is equivalent to

$$Eq.\ (9) \ge \gamma_{\min}\|w\|^2. \tag{39}$$

Now, note that the loss value at the origin is $\mathbb{E}[y^2]$, which means that for any w, whose norm $\|w\|^2 \ge \mathbb{E}[y^2]/\gamma_{\min}$, the loss value must be larger than the loss value of the origin. Therefore, let $r = \mathbb{E}[y^2]/\gamma_{\min}$, we have proved that the global minimum must lie in a closed r-Ball around the origin.

As the last step, because the objective is a continuous function of w and the r-Ball is a compact set, the minimum of the objective in this r-Ball is achievable. This completes the proof. □

B.5 Proof of Theorem 2

We divide the proof into the proof of a proposition and a lemma, and combining the following proposition and lemma obtains the theorem statement.

B.5.1 Proposition 4

Proposition 4. *Any global minimum of Eq. (9) is of the form*

$$\begin{cases} U = b_u\mathbf{r}_D; \\ W^{(i)} = b_i\mathbf{r}_i\mathbf{r}_{i-1}^T; \\ W^{(1)} = \mathbf{r}_1\mathbb{E}[xy]^T(b_u\prod_{i=2}^D b_i)\mu\left[(b_u\prod_{i=2}^D b_i)^2 s^2\left(\sigma^2+d_1\right)A_0 + \gamma_w I\right]^{-1}, \end{cases} \tag{40}$$

where $\mu = \prod_{i=2}^D d_i$, $s^2 = \prod_{i=2}^D d_i(\sigma^2+d_i)$, $b_u \ge 0$ and $b_i \ge 0$, and $\mathbf{r}_i = (\pm1,\dots,\pm1)$ is an arbitrary vertex of a d_i-dimensional hypercube for all i.

Proof. Note that the trivial solution is also a special case of this solution with $b = 0$. We thus focus on deriving the form of the nontrivial solution.

We prove by induction on D. The base case with depth 1 is proved in Theorem 1. We now assume that the same holds for depth $D-1$ and prove that it also holds for depth D.

For any fixed $W^{(1)}$, the loss function can be equivalently written as

$$\mathbb{E}_{\tilde{x}}\mathbb{E}_{\epsilon^{(2)},\ldots,\epsilon^{(D)}}\left(\sum_{i_1,i_2,\ldots,i_D}^{d_1,d_2,\ldots d_D} U_{i_D}\epsilon^{(D)}_{i_D}\ldots\epsilon^{(2)}_{i_2}W^{(2)}_{i_2 i_1}\tilde{x}_{i_1} - y\right)^2 + \gamma_u\|U\|^2 + \sum_{i=2}^{D}\gamma_i\|W^{(i)}\|^2 + const., \tag{41}$$

where $\tilde{x} = \epsilon^{(1)}_{i_1}\sum_i W^{(1)}_{i_1 i}x_i$. Namely, we have reduced the problem to a problem involving only a depth $D-1$ linear net with a transformed input $\tilde{x}$.

By the induction assumption, the global minimum of this problem takes the form of Eq. (10), which means that the loss function can be written in the following form:

$$\mathbb{E}_{\tilde{x}}\mathbb{E}_{\epsilon^{(2)},\ldots,\epsilon^{(D)}}\left(b_u b_D\ldots b_3\sum_{i_1,i_2,\ldots,i_D}^{d_1,d_2,\ldots d_D}\epsilon^{(D)}_{i_D}\ldots\epsilon^{(2)}_{i_2}v_{i_1}\tilde{x}_{i_1} - y\right)^2 + L_2\ reg., \tag{42}$$

for an arbitrary optimizable vector v_{i_1}. The term $\sum_{i_2,\ldots,i_D}^{d_2,\ldots d_D}\epsilon^{(D)}_{i_D}\ldots\epsilon^{(2)}_{i_2} := \eta$ can now be regarded as a single random variable such that $\mathbb{E}[\eta] = \prod_{i=2}^{D} d_i := \mu$ and $\mathbb{E}[\eta^2] = \prod_{i=2}^{D} d_i(\sigma_i^2 + d_i) := s^2$. Computing the expectation over all the noises except for $\epsilon^{(1)}$, one finds

$$\mathbb{E}_{\tilde{x}}\left(b_u b_D\ldots b_3 s\sum_{i_1} v_{i_1}\tilde{x}_{i_1} - \frac{\mu y}{s}\right)^2 + L_2\ reg. + const. \tag{43}$$

$$= \mathbb{E}_{x,\epsilon^{(1)}}\left(b_u b_D\ldots b_3 s\sum_{i,i_1} v_{i_1}\epsilon^{(1)}_{i_1}W^{(1)}_{i_1 i}x_i - \frac{\mu y}{s}\right)^2 + L_2\ reg. + const., \tag{44}$$

where we have ignored the constant term because it does not affect the minimizer of the loss. Namely, we have reduced the original problem to a two-layer linear net problem where the label becomes effectively rescaled for a deep network.

For any fixed $b_u, \ldots, b_3$, we can define $\bar{x} := b_u b_D\ldots b_3 s x$, and obtain the following problem, whose global minimum we have already derived:

$$\mathbb{E}_{\bar{x}}\mathbb{E}_{\epsilon^2,\ldots,\epsilon_D}\left(\sum_{i,i_1} v_{i_1}W^{(1)}_{i_1 i}\bar{x}_i - \frac{\mu y}{s}\right)^2. \tag{45}$$

By Theorem 1, the global minimum is identically 0 if $\|\mathbb{E}[\mu\bar{x}y/s]\|^2 < d_2\gamma_2\gamma_1$, or, $\mathbb{E}[xy] \leq \frac{\gamma_2\gamma_1}{b_3^2\ldots b_u^2(\prod_{i=3}^{D} d_i)}$. When $\mathbb{E}[xy] > \frac{\gamma_2\gamma_1}{b_3^2\ldots b_u^2(\prod_{i=3}^{D} d_i)}$, the solution can be non-trivial:

$$\begin{cases} v_* = b_2^*\mathbf{r}_1; \\ W_* = \mathbf{r}_1\mathbb{E}[xy]^T\mu b_2^* b_3\ldots b_u\left[(b_2^*)^2 d_3^2\ldots d_D^2 b_u^2 s^2\left(\sigma^2 + d_1\right)A_0 + \gamma_1 I\right]^{-1}, \end{cases} \tag{46}$$

for some b_2^*. This proves the theorem. □

B.6 Lemma 3

Lemma 3. *At any global minimum of Eq. (9), let* $b_1 := \sqrt{\|W_{i:}\|^2/d}$ *and* $b_{D+1} := b_u$,

$$\gamma_{k+1}d_{k+1}b_{k+1}^2 = \gamma_k d_{k-1}b_k^2. \tag{47}$$

Proof. It is sufficient to show that for all k and i,

$$\gamma_{k+1}\sum_{ij}(W_{ji}^{k+1})^2 = \gamma_k\sum_{ij}(W_{ij}^{k})^2. \tag{48}$$

We prove by contradiction. Let U^*, W^* be the global minimum of the loss function. Assuming that for an arbitrary k,

$$\gamma_{k+1} \sum_{ij} (W_{ji}^{*,k+1})^2 \neq \gamma_k \sum_{ij} (W_{ij}^{*,k})^2. \tag{49}$$

Introduce W^a such that $W_{ji}^{a,k+1} = aW_{ji}^{*,k+1}$ and $W_{ji}^{a,k} = W_{ji}^{*,k}/a$. The loss without regularization is invariant under the transformation of $W^* \to W^a$, namely

$$L_0(W^*) = L_0(W^a). \tag{50}$$

In the regularization, all the terms remain invariant except two terms:

$$\begin{cases} \gamma_{k+1} \sum_{ij} (W_{ji}^{*,k+1})^2 \to \gamma_{k+1} \sum_{ij} (W_{ji}^{a,k+1})^2 = a^2 \gamma_{k+1} \sum_{ij} (W_{ji}^{*,k+1})^2 \\ \gamma_k \sum_{ij} (W_{ij}^{*,k})^2 \to \gamma_k \sum_{ij} (W_{ji}^{a,k})^2 = a^{-2} \gamma_k \sum_{ij} (W_{ji}^{*,k})^2 \end{cases} \tag{51}$$

It could be shown that, the sum of $a^2 \gamma_{k+1} \sum_{ij} (W_{ji}^{*,k+1})^2$ and $a^{-2} \gamma_k \sum_{ij} (W_{ji}^{*,k})^2$ reaches its minimum when $a^2 = \sqrt{\frac{\gamma_k \sum_{ij} (W_{ji}^{*,k})^2}{\gamma_{k+1} \sum_{ij} (W_{ji}^{*,k+1})^2}}$. If $\gamma_{k+1} \sum_{ij} (W_{ji}^{*,k+1})^2 \neq \gamma_k \sum_{ij} (W_{ij}^{*,k})^2$, one can choose a to minimize the regularization terms in the loss function such that $L(W^a) < L(W^*)$, indicating W^* is not the global minimum. Thus, $\gamma_{k+1} \sum_{ij} (W_{ji}^{*,k+1})^2 \neq \gamma_k \sum_{ij} (W_{ij}^{*,k})^2$ cannot be true. □

B.7 Proof of Proposition 2

Proof. Let

$$L_0 = \mathbb{E}_{\tilde{x}} \mathbb{E}_{\epsilon^2, \ldots, \epsilon_D} \left(\sum_{i_1, i_2, \ldots, i_D}^{d_1, d_2, \ldots d_D} U_{i_D} \epsilon_{i_D}^{(D)} \ldots \epsilon_{i_1}^{(1)} W_{i_1 i}^{(1)} x_i - y \right)^2. \tag{52}$$

L_0 is a polynomial containing $2D + 2$th order, $D + 1$th order, and 0th order terms in terms of parameters U and W. The second order derivative of L is thus a polynomial containing $2D$-th order and $(D-1)$-th order terms; however, other orders are not possible. For $D \geq 2$, there are no constant terms in the Hessian of L, and there is at least a parameter in each of the terms.

The Hessian of the full loss function with regularization is

$$\frac{\partial^2 L}{\partial^2 U_i U_j} = \frac{\partial^2 L_0}{\partial^2 U_i U_j} + (1 - \delta_{ij}) 2\gamma_u (U_i + U_j) + \delta_{ij} 2\gamma_u; \tag{53}$$

$$\frac{\partial^2 L}{\partial^2 W_{jk}^i U_l} = \frac{\partial^2 L_0}{\partial^2 W_{jk}^i U_l} + 2(\gamma_w W_{jk}^i + \gamma_u U_l); \tag{54}$$

$$\frac{\partial^2 L}{\partial^2 W_{jk}^i W_{mn}^l} = \frac{\partial^2 L_0}{\partial^2 W_{jk}^i W_{mn}^l} + (1 - \delta_{il}\delta_{jm}\delta_{kn}) 2\gamma_w (W_{jk}^i + W_{mn}^l) + \delta_{il}\delta_{jm}\delta_{kn} 2\gamma_w. \tag{55}$$

For $U = 0$, $W = 0$, the Hessian of L_0 is 0, since each term in L_0 contains at least a U or a W. The Hessian of L becomes

$$\left. \frac{\partial^2 L}{\partial^2 U_i U_j} \right|_{U,W=0} = \delta_{ij} 2\gamma_u; \tag{56}$$

$$\left. \frac{\partial^2 L}{\partial^2 W_{jk}^i U_l} \right|_{U,W=0} = 0; \tag{57}$$

$$\left. \frac{\partial^2 L}{\partial^2 W_{jk}^i W_{mn}^l} \right|_{U,W=0} = \delta_{il}\delta_{jm}\delta_{kn} 2\gamma_w. \tag{58}$$

The Hessian of L is a positive-definite matrix. Thus, $U = 0$, $W = 0$ is always a local minimum of the loss function L. □

B.8 Proof of Proposition 3

We first apply Lemma 3 to determine the condition for the nontrivial solution to exist. In particular, the Lemma must hold for $W^{(2)}$ and $W^{(1)}$, which leads to the following condition:

$$\|b^{D-1}d_0^{D-1}[b^{2D}d_0^D(\sigma^2+d_0)^D A_0+\gamma]^{-1}\mathbb{E}[xy]\|^2 = 1. \tag{59}$$

Note that the left-hand side is a continuous function that tends to 0 as $b \to \infty$. Therefore, it is sufficient to find the condition that guarantees that there exists b such that the l.h.s. is larger than 1. For any b, the l.h.s. is a monotonically decreasing function of any eigenvalue of A_0, and so the following two inequalities hold:

$$\begin{cases} \|b^{D-1}d_0^{D-1}(b^{2D}d_0^D(\sigma^2+d_0)^D\sigma_x^2+\gamma)^{-1}\mathbb{E}[xy]\| \le \|b^{D-1}d_0^{D-1}(b^{2D}d_0^D(\sigma^2+d_0)^D a_{\min}+\gamma)^{-1}\mathbb{E}[xy]\| \\ \|b^{D-1}d_0^{D-1}(b^{2D}d_0^D(\sigma^2+d_0)^D\sigma_x^2+\gamma)^{-1}\mathbb{E}[xy]\| \ge \|b^{D-1}d_0^{D-1}(b^{2D}d_0^D(\sigma^2+d_0)^D a_{\max}+\gamma)^{-1}\mathbb{E}[xy]\|. \end{cases} \tag{60}$$

The second inequality implies that if

$$\|b^{D-1}d_0^{D-1}[b^{2D}d_0^D(\sigma^2+d_0)^D a_{\max}+\gamma]^{-1}\mathbb{E}[xy]\| > 1, \tag{61}$$

a nontrivial solution must exist. This condition is equivalent to the existence of a b such that

$$d_0^D(\sigma^2+d_0)^D a_{\max} b^{2D} - \|\mathbb{E}[xy]\| b^{D-1}d_0^{D-1} < -\gamma, \tag{62}$$

which is a polynomial inequality that does not admit an explicit condition for b for a general D. Since the l.h.s is a continuous function that increases to infinity as $b \to \infty$, one sufficient condition for (62) to hold is that the minimizer of the l.h.s. is smaller than γ.

Note that the left-hand side of Eq. (62) diverges to ∞ as $b \to \pm\infty$ and tends to zero as $b \to 0$. Moreover, Eq. (62) is lower-bounded and must have a nontrivial minimizer for some $b > 0$ because the coefficient of the b^{D-1} term is strictly negative. One can thus find its minimizer by taking derivative. In particular, the left-hand side is minimized when

$$b^{D+1} = \frac{(D-1)\|\mathbb{E}[xy]\|}{2Dd_0(\sigma^2+d_0)^D a_{\max}}, \tag{63}$$

and we can obtain the following sufficient condition for (62) to be satisfiable, which, in turn, implies that (59) is satisfiable:

$$\frac{D+1}{2D}\|\mathbb{E}[xy]\| d_0^{D-1}\left(\frac{(D-1)\|\mathbb{E}[xy]\|}{2Dd_0(\sigma^2+d_0)^D a_{\max}}\right)^{\frac{D-1}{D+1}} > \gamma, \tag{64}$$

which is identical to the proposition statement in (15).

Now, we come back to condition (60) to derive a sufficient condition for the trivial solution to be the only solution. The first inequality in Condition (60) implies that if

$$\|b^{D-1}d_0^{D-1}[b^{2D}d_0^D(\sigma^2+d_0)^D a_{\min}+\gamma]^{-1}\mathbb{E}[xy]\| \le 1, \tag{65}$$

the only possible solution is the trivial one, and the condition for this to hold can be found using the same procedure as above to be

$$\frac{D+1}{2D}\|\mathbb{E}[xy]\| d_0^{D-1}\left(\frac{(D-1)\|\mathbb{E}[xy]\|}{2Dd_0(\sigma^2+d_0)^D a_{\min}}\right)^{\frac{D-1}{D+1}} \le \gamma, \tag{66}$$

which is identical to (14).

We now prove the upper bound for the solution in ((16)). Because for any b, the first condition in 60 gives an upper bound, and so any b that makes the upper bound less than 1 cannot be a solution. This means that any b for which

$$\|b^{D-1}d_0^{D-1}[b^{2D}d_0^D(\sigma^2+d_0)^D a_{\min}+\gamma]^{-1}\mathbb{E}[xy]\| \le 1 \tag{67}$$

cannot be a solution. This condition holds if and only if

$$d_0^D(\sigma^2+d_0)^D a_{\min} b^{2D} - \|\mathbb{E}[xy]\| b^{D-1}d_0^{D-1} > -\gamma. \tag{68}$$

Because $\gamma > 0$, one sufficient condition to ensure this is that there exists b such that

$$d_0(\sigma^2 + d_0)^D a_{\min} b^{2D} - ||\mathbb{E}[xy]|| b^{D-1} > 0, \tag{69}$$

which is equivalent to

$$b > \left[\frac{||\mathbb{E}[xy]||}{d_0(\sigma^2 + d_0)^D a_{\min}} \right]^{\frac{1}{D+1}}. \tag{70}$$

Namely, any solution b^* satisfies

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

We can also find a lower bound for all possible solutions. When $D > 1$, another sufficient condition for Eq. (68) to hold is that there exists b such that

$$||\mathbb{E}[xy]|| d_0^{D-1} b^{D-1} < \gamma. \tag{72}$$

because the b^{2D} term is always positive. This condition then implies that any solution must satisfy:

$$b^* \geq \frac{1}{d_0} \left[\frac{\gamma}{||\mathbb{E}[xy]||} \right]^{\frac{1}{D-1}}. \tag{73}$$

For $D = 1$, we have by Theorem 1 that

$$b^* > 0 \tag{74}$$

if and only if $\mathbb{E}[xy] > \gamma$. This means that

$$b^* \geq \lim_{\eta \to 0^+} \lim_{D \to 1^+} \frac{1}{d_0} \left[\frac{\gamma + \eta}{||\mathbb{E}[xy]||} \right]^{\frac{1}{D-1}} = \begin{cases} \infty & \text{if } \mathbb{E}[xy] \geq \gamma; \\ 0 & \text{if } \mathbb{E}[xy] < \gamma. \end{cases}. \tag{75}$$

This finishes the proof. □

B.9 Proof of Theorem 3

Proof. When nontrivial solutions exist, we are interested in identifying when $b = 0$ is not the global minimum. To achieve this, we compare the loss of $b = 0$ with the other solutions. Plug the trivial solution into the loss function in Eq. (9), the loss is easily identified to be $L_{\text{trivial}} = E[y^2]$.

For the nontrivial minimum, defining f to be the model,

$$f(x) := \sum_{i, i_1, i_2, \ldots, i_D}^{d, d_1, d_2, \ldots d_D} U_{i_D} \epsilon_{i_D}^{(D)} \ldots \epsilon_{i_2}^{(2)} W_{i_2 i_1}^{(2)} \epsilon_{i_1}^{(1)} W_{i_1 i}^{(1)} x \tag{76}$$

$$= \eta d_0^D b^{2D} \mathbb{E}[xy]^T [b^{2D} d_0^D (\sigma^2 + d_0)^D A_0 + \gamma I]^{-1} x, \tag{77}$$

where, similar to the previous proof, we have defined $\sum_{i_1, \ldots, i_D}^{d_1, \ldots d_D} \epsilon_{i_D}^{(D)} \ldots \epsilon_{i_1}^{(1)} := \eta$ such that $\mathbb{E}[\eta] = \prod_i^D d_i = d_0^D$ and $\mathbb{E}[\eta^2] = \prod_i^D d_i(\sigma_i^2 + d_i) := d_0^D(\sigma^2 + d_0)^D$. With this notation, The loss function becomes

$$\mathbb{E}_x \mathbb{E}_\eta (f(x) - y)^2 + L_2 \; reg. \tag{78}$$

$$= \mathbb{E}_{x,\eta}[f(x)^2] - 2\mathbb{E}_{x,\eta}[y f(x)] + \mathbb{E}_x[y^2] + L_2 \; reg. \tag{79}$$

$$= \sum_i \frac{d_0^{3D}(\sigma^2 + d_0)^D b^{4D} a_i \mathbb{E}[x'y]_i^2}{[d_0^D(\sigma^2 + d_0)^D a_i b^{2D} + \gamma]^2} - 2 \sum_i \frac{d_0^{2D} b^{2D} \mathbb{E}[x'y]_i^2}{d_0^D(\sigma^2 + d_0)^D a_i b^{2D} + \gamma} + \mathbb{E}_x[y^2] + L_2 \; reg. \tag{80}$$

The last equation is obtained by rotating x using a orthogonal matrix such that $R^{-1} A_0 R = \text{diag}(a_i)$ and denoting the rotated x as $x' = Rx$. With x', The $L_2 \; reg$ term takes the form of

$$L_2 \; reg. = \gamma D d_0^2 b^2 + \gamma \sum_i \frac{d_0^{2D} b^{2D} ||\mathbb{E}[x'y]_i||^2}{(d_0^D(\sigma^2 + d_0)^D b^{2D} a_i + \gamma)^2}. \tag{81}$$

Combining the expressions of (81) and (80), we obtain that the difference between the loss at the non-trivial solution and the loss at 0 is

$$-\sum_i \frac{d_0^{2D} b^{2D} \mathbb{E}[x'y]_i^2}{[d_0^D(\sigma^2+d_0)^D a_i b^{2D}+\gamma]} + \gamma D d_0^2 b^2. \tag{82}$$

Satisfaction of the following relation thus guarantees that the global minimum is nontrivial:

$$\sum_i \frac{d_0^{2D} b^{2D} \mathbb{E}[x'y]_i^2}{[d_0^D(\sigma^2+d_0)^D a_i b^{2D}+\gamma]} \geq \gamma D d_0^2 b^2. \tag{83}$$

This relation is satisfied if

$$\frac{d_0^{2D} b^{2D} ||\mathbb{E}[xy]||^2}{[d_0^D(\sigma^2+d_0)^D a_{max} b^{2D}+\gamma]} \geq \gamma D d_0^2 b^2 \tag{84}$$

$$\frac{b^{2D-2}}{[d_0^D(\sigma^2+d_0)^D a_{max} b^{2D}+\gamma]} \geq \frac{\gamma D}{d_0^{2D-2}||\mathbb{E}[xy]||^2}. \tag{85}$$

$$\tag{86}$$

The derivative or l.h.s. with respect to b is

$$\frac{b^{2D-3}[(2D-2)\gamma - 2d_0^D(\sigma^2+d_0)^D a_{max} d^{2D}]}{[d_0^D(\sigma^2+d_0)^D a_{max} b^{2D}+\gamma]^2}. \tag{87}$$

For $b, \gamma \in (0, \infty)$, the derivative dives below 0, indicating the l.h.s. of (86) has a global maximum at a strictly positive b. The value of b is found when setting the derivative to 0, namely

$$\frac{b^{2D-3}[(2D-2)\gamma - 2d_0^D(\sigma^2+d_0)^D a_{max} d^{2D}]}{[d_0^D(\sigma^2+d_0)^D a_{max} b^{2D}+\gamma]^2} = 0 \tag{88}$$

$$(2D-2)\gamma - 2d_0^D(\sigma^2+d_0)^D a_{max} d^{2D} = 0 \tag{89}$$

$$b^{2D} = \frac{(D-1)\gamma}{d_0^D(\sigma^2+d_0)^D a_{max}}. \tag{90}$$

The maximum value then takes the form

$$\frac{(D-1)^{\frac{D-1}{D}}}{D\gamma^{\frac{1}{D}} d_0^{D-1}(\sigma^2+d_0)^{D-1} a_{max}^{\frac{D-1}{D}}}. \tag{91}$$

The following condition thus guarantees that the global minimum is non-trivial

$$\frac{(D-1)^{\frac{D-1}{D}}}{D\gamma^{\frac{1}{D}} d_0^{D-1}(\sigma^2+d_0)^{D-1} a_{max}^{\frac{D-1}{D}}} \geq \frac{\gamma D}{d_0^{2D-2}||\mathbb{E}[xy]||^2} \tag{92}$$

$$||\mathbb{E}[xy]||^2 \geq \frac{\gamma^{\frac{D+1}{D}} D^2(\sigma^2+d_0)^{D-1} a_{max}^{\frac{D-1}{D}}}{d_0^{D-1}(D-1)^{\frac{D-1}{D}}}. \tag{93}$$

This finishes the proof. □

B.10 Proof of Theorem 4

Proof. The model prediction is:

$$f(x) := \sum_{i,i_1,i_2,\ldots,i_D}^{d,d_1,d_2,\ldots d_D} U_{i_D} \epsilon_{i_D}^{(D)} \ldots \epsilon_{i_2}^{(2)} W_{i_2 i_1}^{(2)} \epsilon_{i_1}^{(1)} W_{i_1 i}^{(1)} x \tag{94}$$

$$= \eta d_0^D b^{2D} \mathbb{E}[xy]^T [b^{2D} d_0^D (\sigma^2+d_0)^D \sigma_x^2 I + \gamma I]^{-1} x. \tag{95}$$

One can find the expectation value and variance of a model prediction:

$$\mathbb{E}_\eta[f(x)] = \frac{d_0^{2D} b^{2D} \mathbb{E}[xy]^T x}{b^{2D} d_0^D (\sigma^2+d_0)^D \sigma_x^2 + \gamma} \tag{96}$$

For the trivial solution, the theorem is trivially true. We thus focus on the case when the global minimum is nontrivial.

The variance of the model is

$$
\begin{aligned}
Var[f(x)] &= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2 && (97)\\
&= \frac{(\sigma^2+d_0)^D d_0^{3D} b^{4D} (\mathbb{E}[xy]^T x)^2}{[b^{2D} d_0^D (\sigma^2+d_0)^D \sigma_x^2 + \gamma]^2} - \frac{d_0^{4D} b^{4D} (\mathbb{E}[xy]^T x)^2}{[b^{2D} d_0^D (\sigma^2+d_0)^D]^2 \sigma_x^2 + \gamma]^2} && (98)\\
&= \frac{d_0^{3D}[(\sigma^2+d_0)^D - d_0^D] b^{4D} (\mathbb{E}[xy]^T x)^2}{[b^{2D} d_0^D (\sigma^2+d_0)^D \sigma_x^2 + \gamma]^2} && (99)\\
&= \frac{d_0^{3D}[(\sigma^2+d_0)^D - d_0^D] b^{2D+2} (\mathbb{E}[xy]^T x)^2}{||\mathbb{E}[xy]||^2}, && (100)
\end{aligned}
$$

where the last equation follows from Eq. (13). The variance can be upper-bounded by applying (16),

$$
Var[f(x)] \leq \frac{d_0^D[(\sigma^2+d_0)^D - d_0^D](\mathbb{E}[xy]^T x)^2}{(\sigma^2+d_0)^{2D}\sigma_x^2} \propto \frac{d_0^D[(\sigma^2+d_0)^D - d_0^D]}{(\sigma^2+d_0)^{2D}}. \tag{101}
$$

We first consider the limit $d_0 \to \infty$ with fixed σ^2:

$$
Var[f(x)] \propto \frac{D d_0^{2D-1}\sigma^2}{(d_0+\sigma^2)^{2D}} = O\left(\frac{1}{d_0}\right). \tag{102}
$$

For the limit $\sigma^2 \to \infty$ with d_0 fixed, we have

$$
Var[f(x)] = O\left(\frac{1}{(\sigma^2)^D}\right). \tag{103}
$$

Additionally, we can consider the limit when $D \to \infty$ as we fix both σ^2 and d_0:

$$
Var[f(x)] = O\left(e^{-D2\log[(\sigma^2+d_0)/d_0]}\right), \tag{104}
$$

which is an exponential decay. □

C Exact Form of b^* for $D = 1$

Note that our main result does not specify the exact value of b^*. This is because b^* must satisfy the condition in Eq. (6), which is equivalent to a high-order polynomial in b with coefficients being general functions of the eigenvalues of A_0, whose solutions are generally not analytical by Galois theory. One special case where an analytical formula exists for b is when $A_0 = \sigma_x^2 I$. Practically, this can be achieved for any (full-rank) dataset if we disentangle and rescale the data by the whitening transformation: $x \to \sigma_x \sqrt{A_0^{-1}} x$. In this case, we have

$$
b_*^2 = \frac{\sqrt{\frac{\gamma_w}{\gamma_u}}||\mathbb{E}[xy]|| - \gamma_w}{(\sigma^2+d_1)\sigma_x^2}, \tag{105}
$$

and

$$
v = \pm\sqrt{\frac{\sqrt{\frac{\gamma_u}{\gamma_w}}||\mathbb{E}[xy]|| - \gamma_u}{\sigma_x^2(\sigma^2+d_1)}}\frac{\mathbb{E}[xy]}{||\mathbb{E}[xy]||}, \tag{106}
$$

where $v = W_{i:}$.

D Effect of Bias

This section studies a deep linear network with biases for every layer and compares it with the no-bias networks. We first study a general case when the data does not receive any preprocessing. We then show that the problem reduces to the setting we considered in the main text under the common data preprocessing schemes that centers the input and output data: $\mathbb{E}[x] = 0$, and $\mathbb{E}[y] = 0$.

D.1 Two-layer network

The two-layer linear network with bias is defined as

$$f_b(x; U, W, \beta^U, \beta^W) = \sum_i \epsilon_i U_i (W_{i:} \cdot x + \beta_i^W) + \beta^U, \tag{107}$$

where $\beta^W \in \mathbb{R}^{d_1}$ is the bias in the hidden layer, and $\beta^U \in \mathbb{R}$ is the bias at the output layer. The loss function is

$$L_b(U, W, \beta^U, \beta^W) = \mathbb{E}_{\epsilon,x,y}[(\sum_i \epsilon_i U_i (W_{i:} \cdot x + \beta_i^W) + \beta^U - y)^2] + L_2 \tag{108}$$

$$= \mathbb{E}_{x,y}\left[(UWx + U\beta^W + \beta^U - y)^2 + \sigma^2 \sum_i U_i^2 (W_{i:} \cdot x + \beta_i^W)^2\right] \tag{109}$$

$$+ \gamma_u(\|U\|^2 + (\beta^U)^2) + \gamma_w(\|W\|^2 + \|\beta^W\|^2). \tag{110}$$

It is helpful to concatenate x and 1 into a single vector $x' := (x, 1)^{\mathrm{T}}$ and concatenate W and β^W into a single matrix W' such that W, β^W, x, and W', x' are related via the following equation

$$Wx + \beta^W = W'x'. \tag{111}$$

Using W' and x', the model can be written as

$$f_b(x', U, W', \beta^U) = \sum_i \epsilon_i U_i W'_{i:} \cdot x' + \beta^U. \tag{112}$$

The loss function simplifies to

$$L_b(U, W', \beta) = \mathbb{E}_{\epsilon,x,y}[(\sum_i \epsilon_i U_i W'_{i:} \cdot x' + \beta^U - y)^2] + \gamma_u(\|U\|^2 + (\beta^U)^2) + \gamma_w \|W'\|^2. \tag{113}$$

Note that (113) contains similar rescaling invariance between U and W' and the invariance of aligning $W'_{i:}$ and $W'_{j:}$. One can thus obtain the following two propositions that mirror Lemma 1 and 2.

Proposition 5. *At the global minimum of* (108)*,* $U_j^2 = \frac{\gamma_w}{\gamma_u}\left(\sum_i W_{ji}^2 + (\beta_j^W)^2\right)$.

Proposition 6. *At the global minimum, for all i and j, we have*

$$\begin{cases} U_i^2 = U_j^2; \\ U_i W_{i:} = U_j W_{j:}; \\ U_i \beta_i^W = U_j \beta_j^W. \end{cases} \tag{114}$$

The proofs are omitted because they are the same as those of Lemma 1 and 2, substituting W by W'. By following a procedure similar to finding the solution for a no-bias network, one finds that

Theorem 5. *The global minimum of Eq.* (108) *is of the form*

$$\begin{cases} U = b\mathbf{r}; \\ \beta^U = \frac{d_1 b\left[(d_1+\sigma^2)\frac{\gamma_u}{\gamma_w}b - 1\right] v\mathbb{E}[x] - \left(d_1 \frac{\gamma_u}{\gamma_w} b^2 - 1\right)\mathbb{E}[y]}{(d_1+\sigma^2) d_1 \frac{\gamma_u^2}{\gamma_w^2} b^4 + (\gamma_u - 2) d_1 \frac{\gamma_u}{\gamma_w} b^2 + \gamma_u + 1}; \\ W = \mathbf{r}b\left\{\mathbb{E}[x]\left[b\frac{\gamma_u}{\gamma_w}(d_1 + b\sigma^2) - 1\right]\beta^U + \mathbb{E}[xy]\right\}^T [b^2(d_1+\sigma^2)A_0 + \gamma_w I]^{-1}; \\ \beta^W = -\mathbf{r}\frac{\gamma_u}{\gamma_w} b\beta^U, \end{cases} \tag{115}$$

where b satisfies

$$\gamma_u b^2 = b^2 \frac{\gamma_w(\frac{\mathbb{E}[y]S_1S_3}{S_4}\mathbb{E}[x] - \mathbb{E}[xy])(M^{-1})^2(\frac{\mathbb{E}[y]S_1S_3}{S_4}\mathbb{E}[x] - \mathbb{E}[xy])^T + \frac{\gamma_u^2}{\gamma_w}\left(\frac{S_3}{S_4}\mathbb{E}[y] - b\frac{S_2}{S_4}\mathbb{E}[x]M^{-1}\mathbb{E}[xy]\right)^2}{\left(b\frac{S_2S_1}{S_4}\mathbb{E}[x]M^{-1}\mathbb{E}[x]^T - 1\right)^2}, \tag{116}$$

where M, S_1, S_2, S_3, S_4 are functions of the model parameters and b, defined in Eq. (122)*.*

Proof. First of all, we derive a handy relation satisfied by β^U and β^W at all the stationary points. The zero-gradient condition of the stationary points gives

$$\begin{cases}\mathbb{E}_{x,y}[2(UWx + U\beta^W + \beta^U - y)]U + 2\gamma_w\beta^W = 0;\\ \mathbb{E}_{x,y}[2(UWx + U\beta^W + \beta^U - y)] + 2\gamma_u\beta^U = 0,\end{cases} \tag{117}$$

leading to

$$U\gamma_u\beta^U + \gamma_w\beta^W = 0 \tag{118}$$

$$\beta_i^W = -\frac{\gamma_u}{\gamma_w}U_i\beta^U. \tag{119}$$

Proposition 5 and proposition 6 implies that we can define $b := |U_i|$ and $bv := U_iW_{i:}$. Consequently, $U_i\beta_i^W = -\frac{\gamma_u}{\gamma_w}b^2\beta^U$, and the loss function can be written as

$$\mathbb{E}_x\left[\left(bd_1\sum_j v_jx_j - (d_1\frac{\gamma_u}{\gamma_w}b^2 - 1)\beta^U - y\right)^2 + b^2d_1\sigma^2\left(\sum_k v_kx_k - \frac{\gamma_u}{\gamma_w}b\beta^U\right)^2\right] + \gamma_ud_1b^2$$
$$+\gamma_wd_1\|v\|_2^2 + \gamma_u\left(\frac{b^2d_1\gamma_u}{\gamma_w} + 1\right)(\beta^U)^2. \tag{120}$$

The respective zero-gradient condition for v and β^U implies that for all stationary points,

$$\begin{cases}v = [b^2(d_1+\sigma^2)A_0 + \gamma_wI]^{-1}b\left\{\mathbb{E}[x]\left[b\frac{\gamma_u}{\gamma_w}(d_1 + b\sigma^2) - 1\right]\beta^U + \mathbb{E}[xy]\right\};\\ \beta^U = \frac{d_1b\left[(d_1+\sigma^2)\frac{\gamma_u}{\gamma_w}b-1\right]v\mathbb{E}[x]-\left(d_1\frac{\gamma_u}{\gamma_w}b^2-1\right)\mathbb{E}[y]}{(d_1+\sigma^2)d_1\frac{\gamma_u^2}{\gamma_w^2}b^4+(\gamma_u-2)d_1\frac{\gamma_u}{\gamma_w}b^2+\gamma_u+1}.\end{cases} \tag{121}$$

To shorten the expressions, we introduce

$$\begin{cases}M = b^2(d_1+\sigma^2)A_0 + \gamma_wI;\\ S_1 = b\frac{\gamma_u}{\gamma_w}(d_1 + b\sigma^2) - 1;\\ S_2 = d_1b\left[(d_1+\sigma^2)\frac{\gamma_u}{\gamma_w}b - 1\right];\\ S_3 = d_1\frac{\gamma_u}{\gamma_w}b^2 - 1;\\ S_4 = (d_1+\sigma^2)d_1\frac{\gamma_u^2}{\gamma_w^2}b^4 + (\gamma_u-2)d_1\frac{\gamma_u}{\gamma_w}b^2 + \gamma_u + 1.\end{cases} \tag{122}$$

With M, S_1, S_2, S_3, S_4, we have

$$\begin{cases}v = M^{-1}b(\mathbb{E}[x]S_1\beta^U + \mathbb{E}[xy]);\\ \beta^U = \frac{S_2v\mathbb{E}[x]-S_3\mathbb{E}[y]}{S_4}.\end{cases} \tag{123}$$

The inner product of v and $\mathbb{E}[x]$ can be solved as

$$v\mathbb{E}[x] = b\frac{\frac{S_3}{S_4}\mathbb{E}[x]M^{-1}\mathbb{E}[x]S_1\mathbb{E}[y] - \mathbb{E}[x]M^{-1}\mathbb{E}[xy]}{b\frac{S_2}{S_4}\mathbb{E}[x]M^{-1}\mathbb{E}[x]S_1 - 1}. \tag{124}$$

Inserting the expression of $v\mathbb{E}[x]$ into the expression of β^U one obtains

$$\beta^U = \frac{S_3\mathbb{E}[y] - bS_2\mathbb{E}[x]M^{-1}\mathbb{E}[xy]}{b\mathbb{E}[x]M^{-1}\mathbb{E}[x]S_1S_2 - S_4} \tag{125}$$

The global minimum must thus satisfy

$$\gamma_ub^2 = \gamma_w\|v\|^2 + \gamma_u\frac{b^2d_1\gamma_u}{\gamma_w}(\beta^U)^2 \tag{126}$$

$$= b^2\frac{\gamma_w(\frac{\mathbb{E}[y]S_1S_3}{S_4}\mathbb{E}[x] - \mathbb{E}[xy])(M^{-1})^2(\frac{\mathbb{E}[y]S_1S_3}{S_4}\mathbb{E}[x] - \mathbb{E}[xy])^T + \frac{\gamma_u^2}{\gamma_w}\left(\frac{S_3}{S_4}\mathbb{E}[y] - b\frac{S_2}{S_4}\mathbb{E}[x]M^{-1}\mathbb{E}[xy]\right)^2}{\left(b\frac{S_2S_1}{S_4}\mathbb{E}[x]M^{-1}\mathbb{E}[x]^T - 1\right)^2}. \tag{127}$$

This completes the proof. □

Remark. *As in the no-bias case, we have reduced the original problem to a one-dimensional problem. However, the condition for b becomes so complicated that it is almost impossible to understand. That being said, the numerical simulations we have done all carry the bias terms, suggesting that even with the bias term, the mechanisms are qualitatively similar, and so the approach in the main text is justified.*

When $\mathbb{E}[x] = 0$, the solution can be simplified a little:

$$\begin{cases} U = \mathbf{r}b; \\ \beta^U = -\frac{d_1 \frac{\gamma_u}{\gamma_w} b^2 - 1}{(d_1+\sigma^2) d_1 \frac{\gamma_u^2}{\gamma_w^2} b^4 + (\gamma_u - 2) d_1 \frac{\gamma_u}{\gamma_w} b^2 + \gamma_u + 1} \mathbb{E}[y]; \\ W = \mathbf{r}b\mathbb{E}[xy]^T [b^2(d_1+\sigma^2)A_0 + \gamma_w I]^{-1}; \\ \beta^W = \mathbf{r}\frac{\gamma_u}{\gamma_w} b \frac{d_1 \frac{\gamma_u}{\gamma_w} b^2 - 1}{(d_1+\sigma^2) d_1 \frac{\gamma_u^2}{\gamma_w^2} b^4 + (\gamma_u - 2) d_1 \frac{\gamma_u}{\gamma_w} b^2 + \gamma_u + 1} \mathbb{E}[y], \end{cases} \tag{128}$$

where the value of b is either 0 or determined by

$$\gamma_u = \gamma_w |\mathbb{E}[xy]^T [b^2(d_1+\sigma^2)A_0 + \gamma_w I]^{-1}|^2 + \frac{\gamma_u^2}{\gamma_w} \mathbb{E}[y]^2 \left(\frac{d_1 \frac{\gamma_u}{\gamma_w} b^2 - 1}{(d_1+\sigma^2) d_1 \frac{\gamma_u^2}{\gamma_w^2} b^4 + (\gamma_u - 2) d_1 \frac{\gamma_u}{\gamma_w} b^2 + \gamma_u + 1} \right)^2. \tag{129}$$

In this case, the expression of W is identical to the no-bias model. The bias of both layers is proportional to $\mathbb{E}[y]$. The equation determining the value of b is also similar to the no-bias case. The only difference is the term proportional to $\mathbb{E}[y]^2$.

We also note that the solution becomes significantly simplified when $\mathbb{E}[x] = 0$ and $\mathbb{E}[y] = 0$. This could be seen by finding the partial derivative of L with respect to β^W and β^U and then setting them to 0. When $\mathbb{E}[x] = 0$, $\mathbb{E}[y] = 0$, one obtains:

$$\begin{cases} \frac{\partial L}{\partial \beta_i^W} = U_i U \beta^W + U_i \beta^U + \gamma_w \beta_i^W = 0; \\ \frac{\partial L}{\partial \beta^U} = U \beta^W + \beta^U + \gamma_u \beta^U = 0. \end{cases} \tag{130}$$

These equations lead to

$$U \gamma_u \beta^U + \gamma_w \beta^W = 0, \tag{131}$$

implying

$$\begin{cases} \beta^U = 0; \\ \beta^W = 0. \end{cases} \tag{132}$$

In practice, it is common and usually recommended practice to subtract the average of x and y from the data and achieve precisely $\mathbb{E}[x] = 0$ and $\mathbb{E}[y] = 0$. We generalize this result to deeper networks in the next section.

D.2 Deep linear network

Let β be a $\left(\sum_i^D d_i + 1\right)$-dimensional vector concatenating all $\beta^1, \beta^2, ..., \beta^D, \beta^U$, and denoting the collection of all the weights U, W^D, ..., W^1 by w, the model of a deep linear network with bias is defined as

$$f_b(x, W^D, ..., W^1, U, \beta^D, ..., \beta^1, \beta^U) \tag{133}$$

$$= (\epsilon^U \circ U)((\epsilon^D \circ W^D)(...((\epsilon^2 \circ W^2)((W^1 x + \beta^1) + \beta^2)...) + \beta^D) + \beta^U \tag{134}$$

$$= (\epsilon^U \circ U)(\epsilon^D \circ W^D)...(\epsilon^2 \circ W^2) W^1 x + (\epsilon^U \circ U)(\epsilon^D \circ W^D)...(\epsilon^2 \circ W^2)\beta^1 \tag{135}$$

$$+ (\epsilon^U \circ U)(\epsilon^D \circ W^D)...(\epsilon^3 \circ W^3)\beta^2 + ... + (\epsilon^U \circ U)\beta^D + \beta^U \tag{136}$$

$$= (\epsilon^U \circ U)(\epsilon^D \circ W^D)...(\epsilon^2 \circ W^2) W^1 x + bias(w, \beta), \tag{137}$$

where

$$bias(w, \beta) = (\epsilon^U \circ U)(\epsilon^D \circ W^D)...(\epsilon^2 \circ W^2)\beta^1 + (\epsilon^U \circ U)(\epsilon^D \circ W^D)...(\epsilon^3 \circ W^3)\beta^2 + ... + (\epsilon^U \circ U)\beta^D + \beta^U, \tag{138}$$

and $\circ$ denotes Hadamard product. The loss function is

$$L_b(x, y, w, \beta) = \mathbb{E}_{\epsilon,x,y}[(f_b(x, w, \beta) - y)^2] + L_2(w, \beta). \tag{139}$$

Proposition 5 and Proposition 6 can be generated to deep linear network. Similar to the no-bias case, we can reduce the landscape to a 1-dimensional problem by performing induction on D and using the 2-dimensional case as the base step. However, we do not solve this case explicitly here because the involved expressions now become too long and complicated even to write down, nor can they directly offer too much insight. We thus only focus on the case when the data has been properly preprocessed. Namely, $\mathbb{E}[x] = 0$ and $\mathbb{E}[y] = 0$.

For simplicity, we assume that the regularization strength for all the layers employs the value γ. The following theorem shows that When $\mathbb{E}[x] = 0$ and $\mathbb{E}[y] = 0$, the biases vanish for an arbitrarily deep linear network:

Theorem 6. *Let $\mathbb{E}[x] = 0$ and $\mathbb{E}[y] = 0$. The global minima of Eq. (139) have $\beta^1 = 0, \beta^2 = 0, ..., \beta^D = 0, \beta^U = 0$.*

Proof. At the global minimum, the gradient of the loss function vanishes. In particular, the derivatives with respect to β vanish:

$$\frac{\partial L_b(x, y, w, \beta)}{\partial \beta_i} = 0; \tag{140}$$

$$\mathbb{E}_{\epsilon,x,y}\left[\frac{\partial f_b(x, w, \beta)}{\partial \beta_i}(f_b(x, w, \beta) - y)\right] + \gamma\beta_i = 0; \tag{141}$$

$$\mathbb{E}_{\epsilon,x,y}\left[\frac{\partial bias(w, \beta)}{\partial \beta_i}(f_b(x, w, \beta) - y)\right] + \gamma\beta_i = 0; \tag{142}$$

$$\mathbb{E}_{\epsilon}\left[\frac{\partial bias(w, \beta)}{\partial \beta_i}(f_b(\mathbb{E}[x], w, \beta) - \mathbb{E}[y])\right] + \gamma\beta_i = 0, \tag{143}$$

where β_i is the ith element of β. The last equation is obtained since $f_b(x, w, \beta)$ is a linear function of x. Using the condition $\mathbb{E}[x] = 0$ and $\mathbb{E}[y] = 0$, Equation (143) becomes

$$\mathbb{E}_{\epsilon}\left[\frac{\partial bias(w, \beta)}{\partial \beta_i} bias(w, \beta)\right] + \gamma\beta_i = 0. \tag{144}$$

$bias(w, \beta)$ is a linear combination of β_i. Consequently, $\partial bias(w, \beta)/\partial\beta_i$ does not depend on β, and $bias(w, \beta)\partial bias(w, \beta)/\partial\beta_i$ is a linear combination of β_i. The $\sum_i^D d_i + 1$ equation derived from vanishing gradient yield a set of $\sum_i^D d_i + 1$ linear equations of the form $M(w)\beta = 0$, where $M(w)$ is a $\left(\sum_i^D d_i + 1\right) \times \left(\sum_i^D d_i + 1\right)$ matrix with dependence on w. These linear equations are linearly independent, since the term $\partial L_2(w, \beta)/\partial\beta_i = 2\gamma\beta_i$ and is different in each of the equations. Thus, the linear system $M(w)\beta = 0$ has $\left(\sum_i^D d_i + 1\right)$ independent equations and $\left(\sum_i^D d_i + 1\right)$ variables. The only possible solution to this linear system is

$$\beta = 0. \tag{145}$$

This finishes the proof. □

Thus, for a deep linear network, a model without bias is good enough to describe data satisfying $\mathbb{E}[x] = 0$ and $\mathbb{E}[y] = 0$, which could be achieved by subtracting the mean of the data.

E Case of $\gamma = 0$

When there is no weight decay, we show that the global minimum we found is still a global minimum. In fact, additional symmetries (invariances) emerge in the solution because the weight balance condition in Lemma 3 is no longer required. That being said, one can obtain a manifold of global minimum from our solution by applying the rescaling transformations of our proposed solutions in Theorem 2.

The following theorem allows us to do this by showing that the globally minimal value of the loss is right-continuous at $\gamma = 0$ (and so the solutions we found remain a solution when $\gamma = 0$). For

notational simplicity, we assume that all regularization takes the same value γ and use w to denote all the weights concatenated together. $L(\gamma)$ is the value of the loss function at global minimum with a fixed γ and $l(w, \gamma)$ is the value of the loss function for regularization strength γ at w. We first prove a useful lemma.

Lemma 4. *For any finite $d_i > 0$, $D > 0$ and $\gamma \in [0, \infty)$, $L(\gamma)$ is left continuous.*

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 this 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

$$L(\gamma^* - \epsilon) = \min_w \ell(w, \gamma^* - \epsilon) := \ell(w', \gamma^* - \epsilon), \tag{146}$$

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 $\epsilon > 0$,

$$L(\gamma^* - \epsilon) < L(\gamma^*) - \delta, \tag{147}$$

which implies that

$$\ell(w', \gamma^* - \epsilon) = L(\gamma^* - \epsilon) < L(\gamma^*) - \delta \leq \ell(w', \gamma^*) - \delta. \tag{148}$$

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

$$\ell(w', \gamma^* - \epsilon) \leq \ell(w', \gamma^*) - \delta. \tag{149}$$

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

$$\ell(w, \gamma) - \ell(w, \gamma - \epsilon) = \epsilon \|w\|^2. \tag{150}$$

Thus, by choosing $\epsilon < \delta / \|w\|^2$, the relation in (148) is violated. Thus, $L(\gamma)$ must be left-continuous. □

With Lemma 4, one can prove the following theorem.

Theorem 7. *When $\gamma = 0$, the weight $w_*(\gamma)$ given by Theorem 2 is one of the global minima of the loss function.*

Proof. As a consequence of Lemma 4,

$$\lim_{\epsilon \to 0} L(\epsilon) = \lim_{\epsilon \to 0} l(w_*(\epsilon), \epsilon) = L(0). \tag{151}$$

Namely, as $\gamma \to 0_+$, the value of the loss function obtained by our solution is $L(0)$. By definition,

$$L(0) = l(w_*(0), 0) < l(w, 0), \forall w. \tag{152}$$

Thus, when $\gamma = 0$, w_* is one of global minima. □

Therefore, the solution given by 2 remains as a global minimum if there is no regularization. The regularization imposes constraints on the absolute value of weights in the way described in Lemma 3. These constraints are now removed when $\gamma = 0$, and by multiplying a positive constant and dividing the same constant into two connecting layers in the solution given by 2, one can obtain other solutions. This means that each of the solutions we obtained can be extended to obtain a manifold of globally optimal solutions. However, whether these solutions exhaust all the global minima of the loss function remains an open question.

That being said, we comment that among all global minima at $\gamma = 0$, the limiting solution of our solution is necessarily the one that has the smallest norm and is often expected to have better generalization performance than the average of the rest of the global minima. In fact, many of the previous works study the generalization performance of this special limit of weight decay of linear models (Krogh and Hertz, 1992; Hastie et al., 2019).

F Case of $\sigma^2 = 0$

Similar to the case of $\gamma = 0$, L is also a monotonic increasing function with respect to σ^2. Thus, one finds a lemma similar to Lemma 4:

Lemma 5. *For any finite $d_i > 0$, $D > 0$ and $\sigma^2 \in [0, \infty)$, $L(\sigma^2)$ is left continuous.*

Proof. The proof of this lemma is identical to the proof of Lemma 4, after switching γ into σ^2 and w^2 into $(UW^{(D)}...W^{(1)}x)^2$. □

As a consequence, one finds a theorem similar to Theorem 7.

Theorem 8. *When $\sigma^2 = 0$, the weight $w_*(\sigma^2)$ given by Theorem 2 is one of the global minima of the loss function.*

Proof. The proof of this theorem is identical to the proof of Theorem 7, after switching γ into σ^2. □

Again, when setting σ^2 to 0, the constraint imposed by Lemma 2 is removed, resulting in the degeneracy of the global minimum. For a deep linear network, the dropout eliminates the rotational symmetry between two connecting layers. Leaving out dropout recovers the rotational symmetry between connecting layers, and a manifold of solution at $\sigma^2 = 0$ can be obtained by rotating two consecutive layers simultaneously. However, whether these solutions exhaust all the global minima of the loss function remains an open question. In fact, to the best of our knowledge, there exists no work that completely characterizes all the global minima of an L_2 regularized deep linear net.

Moreover, as in the case of zero-weight-decay, the zero-noise limit of our solution is also worth studying per se because, among all the global minima of $\sigma^2 = 0$, this solution is the one that the most captured the regularization and averaging effect of dropout, and is thus expected have better generalization performance than an arbitrary solution.