A Theory of High Dimensional Regression with Arbitrary Correlations between Input Features and Target Functions: Sample Complexity, Multiple Descent Curves and a Hierarchy of Phase Transitions

Gabriel C. Mel [1] **Surya Ganguli** [2]

Abstract

The performance of neural networks depends on precise relationships between four distinct ingredients: the architecture, the loss function, the statistical structure of inputs, and the ground truth target function. Much theoretical work has focused on understanding the role of the first two ingredients under highly simplified models of random uncorrelated data and target functions. In contrast, performance likely relies on a conspiracy between the statistical structure of the input distribution and the structure of the function to be learned. To understand this better we revisit ridge regression in high dimensions, which corresponds to an exceedingly simple architecture and loss function, but we analyze its performance under arbitrary correlations between input features and the target function. We find a rich mathematical structure that includes: (1) a dramatic reduction in sample complexity when the target function aligns with data anisotropy; (2) the existence of multiple descent curves; (3) a sequence of phase transitions in the performance, loss landscape, and optimal regularization as a function of the amount of data that explains the first two effects.

1. Introduction and Motivation

The field of machine learning, especially in the form of deep learning, has undergone major revolutions in the last several years, leading to significant advances in multiple domains (LeCun et al., 2015). Many works have attempted to gain a theoretical understanding of the key reasons underlying the empirical success of deep neural networks (see e.g. (Bahri et al., 2020) for a review). A major open puzzle concerns understanding the ability of not only deep networks, but also many other machine learning (ML) methods, to successfully generalize to test examples drawn from the same distribution as the training data. Such theory would ideally guide the choice of various ML hyperparameters to achieve minimal test error.

However, any such theory that is powerful enough to guide hyperparameter choices in real world settings necessarily involves assumptions about the nature of the ground truth mapping to be learned, the statistical structure of the inputs one has access to, the amount of data available, and the machine learning method itself. Many theoretical works studying the generalization properties of ML methods have made highly simplified assumptions about the statistical structure of inputs and the ground truth mapping to be learned.

Much work in computer science for example seeks to find *worst case* bounds on generalization error over the choice of the worst data set and the worst ground truth function (Vapnik, 1998). Such bounds are often quite pessimistic as the types of data and ground truth functions that occur in natural problems are far from worst case. These worst case bounds are then often vacuous (Zhang et al., 2017) and cannot guide choices made in practice, spurring the recent search for nonvacuous data-dependent bounds (Dziugaite & Roy, 2017).

Also, much work in statistical physics has focused instead on computing exact formulas for the test error of various ML algorithms (Engel & den Broeck, 2001; Advani et al., 2013; Advani & Ganguli, 2016a;b) including compressed sensing (Rangan et al., 2009; Donoho et al., 2009; Ganguli & Sompolinsky, 2010b;a), single layer (Seung et al., 1992) and multi-layer (Monasson & Zecchina, 1995; Lampinen & Ganguli, 2018) neural networks, but under highly simplified random uncorrelated assumptions about the inputs. Moreover, the ground truth target function was often chosen randomly in a manner that was uncorrelated with the choice of the random inputs. In most situations where the ground truth function possessed no underlying simple structure, these statistical physics works based on random uncorrelated functions and inputs generally concluded that the ratio

[1]Neurosciences PhD Program, School of Medicine, Stanford University, CA, US [2]Department of Applied Physics, Stanford, CA, US. Correspondence to: G.C. Mel <meldefon@gmail.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

of the number of data points must be proportional to the number of unknown parameters in order to achieve good generalization. In practice, this ratio is much less than 1.

A key, and in our view foundational ingredient underlying the theory of generalization in ML is an adequate model of the statistical structure of the inputs, the nature of the ground truth function to be learned, and most importantly, an adequate understanding of how the alignment between these two can potentially dramatically impact the success of generalization. For example, natural images and sounds all have rich statistical structure, and the functions we wish to learn in these domains (e.g. image and speech recognition) are not just random functions, but rather are functions that are intimately tied to the statistical structure of the inputs. Several works have begun to explore the effect of structured data on learning performance in the context of compressed sensing in a dynamical system (Ganguli & Sompolinsky, 2010a), the Hidden Manifold Model (Goldt et al., 2020; Gerace et al., 2020), linear and kernel regression (Chen et al., 2021; Canatar et al., 2021), and two-layer neural networks (Ghorbani et al., 2020). As pointed out by these authors, it is likely the case that a fundamental and still poorly understood conspiracy between the structure of the inputs we receive and the aligned nature of the functions we seek to compute on them plays a key role in the success of various ML methods in achieving good generalization.

Therefore, with these high level motivations in mind, we sought to develop an asymptotically exact analytic theory of generalization for ridge regression in the high dimensional statistical limit where the number of samples N and number of features P are both large but their ratio is $O(1)$. Ridge regression constitutes a widely exploited method especially for high dimensional data; indeed it was shown to be optimal for small amounts of isotropic but non-Gaussian data (Advani & Ganguli, 2016a;b). Moreover the high dimensional statistical limit is increasingly relevant for many fields in which we can simultaneously measure many variables over many observations. The novel ingredient we add is that we assume the inputs have an arbitrary covariance structure, and the ground truth function has an arbitrary alignment with this covariance structure. We focus in particular on examples involving highly heterogeneous multi-scale data. Our key contributions are: (1) We derive exact analytic formulas for test error in high dimensions; (2) we show that for a random target function, isotropic data yields the lowest error; (3) we demonstrate that for anisotropic data, alignment of the target function with this anisotropy yields lowest test error; (4) we derive an analytic understanding of the optimal regularization parameter as a function of the structure of the data and target function; (5) for multi-scale data we find a sequence of phase transitions in the test error and the optimal regularization parameter that result in multiple descent curves with arbitrary numbers of peaks, thereby generalizing the phenomenon of double descent (Belkin et al., 2019; Mei & Montanari, 2020; d'Ascoli et al., 2020); (6) we analytically compute the spectrum of the Hessian of the loss landscape and show that this spectrum undergoes a sequence of phase transitions as successively finer scales become visible with more data; (7) we connect these spectral phase transitions to the phenomenon of multiple descent and to phase transitions in the optimal regularization.

Finally, we note that the phenomenon of double descent in a generalization curve refers to non-monotonic behavior in this curve, corresponding to a single intermediate peak, as a function of the ratio of the number data points to parameters. This can occur either when the amount of data is held fixed and the number of parameters increases (Belkin et al., 2019; Mei & Montanari, 2020; d'Ascoli et al., 2020; Chen et al., 2021), or when the number of parameters is held fixed but the amount of data increases (Seung et al., 1992; Engel & den Broeck, 2001). Here we exhibit multiple descent curves with arbitrary numbers of peaks in the latter setting. We explain their existence in terms of a hierarchy of phase transitions in the empirical covariance matrix of multi-scale correlated data. And furthermore, we show that such non-monotonic behavior is a consequence of suboptimal regularization. Indeed we show how to recover monotonically *decreasing* generalization error curves with *increasing* amounts of data by using an optimal regularization that depends on the ratio of data points to parameters.

2. Overall Framework

Generative Model. We study a generative model of data consisting of N independent identically distributed (iid) random Gaussian input vectors $\mathbf{x}^\mu \in \mathbb{R}^P$ for $\mu = 1, \ldots, N$, each drawn from a zero mean Gaussian with covariance matrix $\mathbf{\Sigma}$ (i.e. $\mathbf{x}^\mu \sim \mathcal{N}(0, \mathbf{\Sigma})$), N corresponding iid scalar noise realizations $\epsilon^\mu \sim \mathcal{N}\left(0, \sigma^2\right)$, and N outputs y^μ given by

$$y^\mu = \mathbf{x}^\mu \cdot \mathbf{w} + \epsilon^\mu,$$

where $\mathbf{w} \in \mathbb{R}^P$ is an unknown ground truth regression vector. The outputs y^μ decompose into a signal component $\mathbf{x}^\mu \cdot \mathbf{w}$ and noise component ϵ^μ and signal to noise ratio (SNR) given by the relative power

$$SNR := \frac{\text{Var}[\mathbf{x} \cdot \mathbf{w}]}{\text{Var}[\epsilon]} = \frac{\mathbf{w}^T \mathbf{\Sigma} \mathbf{w}}{\sigma^2}, \tag{1}$$

plays a critical role in estimation performance. For convenience below we also define the fractional signal power $f_s := \frac{\mathbf{w}^T \mathbf{\Sigma} \mathbf{w}}{\mathbf{w}^T \mathbf{\Sigma} \mathbf{w} + \sigma^2}$ and fractional noise power $f_n := \frac{\sigma^2}{\mathbf{w}^T \mathbf{\Sigma} \mathbf{w} + \sigma^2}$. Note $f_s + f_n = 1$ and $SNR = \frac{f_s}{f_n}$.

High Dimensional Statistics Limit. We will be working in the nontrivial high dimensional statistics limit where

$N, P \to \infty$ but the measurement density $\alpha = N/P$ remains $O(1)$. We assume that $\mathbf{\Sigma}$ has P eigenvalues that are each $O(1)$ so that both $\sigma_x^2 := \frac{1}{P}\mathbb{E}\,\mathbf{x}^T\mathbf{x} = \frac{1}{P}\text{Tr}\,\mathbf{\Sigma}$ and the individual components of $\mathbf{x}$ remain $O(1)$ as $P \to \infty$. Furthermore we assume that $\mathbf{w}^T\mathbf{w}$ is $O(1)$ (i.e. each component of $\mathbf{w}$ is $O(1/\sqrt{P})$ and the noise variance σ^2 is $O(1)$ so that the SNR remains $O(1)$.

Estimation Procedure. We construct an estimate $\hat{\mathbf{w}}$ of $\mathbf{w}$ from the data $\{\mathbf{x}^\mu, y^\mu\}_{\mu=1}^N$ using ridge regression:

$$\hat{\mathbf{w}} = \arg\min_w \frac{1}{N}\sum_{\mu=1}^{N}(y^\mu - \mathbf{x}^\mu \cdot w)^2 + \lambda||w||^2. \quad (2)$$

The solution to this optimization problem is given by

$$\hat{\mathbf{w}} = \left(\mathbf{X}^T\mathbf{X} + \lambda N \mathbf{I}_P\right)^{-1}\mathbf{X}^T\mathbf{y}, \quad (3)$$

where $\mathbf{X}$ is an $N \times P$ matrix whose μ^{th} row is $\mathbf{x}^{\mu T}$, $\mathbf{y} \in \mathbb{R}^N$ with components y^μ and $\mathbf{I}_P$ is the $P \times P$ identity matrix.

Performance Evaluation. We will be interested in the fraction of unexplained variance $\mathcal{F}$ on a new test example $(\mathbf{x}, \epsilon, y)$, averaged over realizations of the inputs $\mathbf{x}^\mu$ and noise ϵ^μ that determine the training set:

$$\mathcal{F} := \frac{\mathbb{E}_{\mathbf{x}^\mu,\epsilon^\mu,\mathbf{x},\epsilon}\left[(y - \mathbf{x}\cdot\hat{\mathbf{w}})^2\right]}{\mathbb{E}_{\mathbf{x},\epsilon}[y^2]}. \quad (4)$$

In the high dimensional statistics limit, the error $\mathcal{F}$ will concentrate about its mean value, and will depend upon the measurement density α, the noise level σ^2, the input covariance $\mathbf{\Sigma}$, the ground truth vector $\mathbf{w}$ and the regularization λ. In particular we will be interested in families of problems of the same SNR but with varying degrees of alignment between the ground truth $\mathbf{w}$ with eigenspaces of $\mathbf{\Sigma}$.

Models of Multi-Scale Covariance Matrices. The alignment of $\mathbf{w}$ with different eigenspaces of $\mathbf{\Sigma}$ (at fixed SNR) becomes of particular interest when $\mathbf{\Sigma}$ contains a hierarchy of multiple scales. In particular consider a D scale model in which $\mathbf{\Sigma}$ has D distinct eigenvalues S_d^2 for $d = 1, \ldots, D$ where each eigenvalue has multiplicity P_d where $\sum_{d=1}^D P_d = P$. A special case is the isotropic single scale model where $\mathbf{\Sigma} = S_1^2\mathbf{I}_P$. Another important special case is the anisotropic two scale model with two distinct eigenvalues S_1^2 and S_2^2 with multiplicities P_1 and P_2, corresponding to long and short data directions. Throughout we will quantify the input anisotropy via the aspect ratio $\gamma := \frac{S_1}{S_2}$ (with $\gamma = 1$ reducing to isotropy). For simplicity, we will balance the number of long and short directions (i.e. $P_1 = P_2$).

3. Results

3.1. Exact High Dimensional Error Formula.

Our first result is a formula for $\mathcal{F}$ for arbitrary α, σ^2, $\mathbf{\Sigma}$, $\mathbf{w}$ and λ,that is asymptotically exact in the high dimensional statistical limit. To understand this formula, it is useful to first compare to the scalar case where $P = 1$ and N is large but finite. $\mathbf{\Sigma}$ then has a single eigenvalue S^2 and $\mathcal{F}$ in this scalar case is given by (see SM for details)

$$\mathcal{F}_{scalar} \approx f_n + f_s\left(\frac{\lambda}{S^2+\lambda}\right)^2 + f_n\frac{1}{N}\left(\frac{S^2}{S^2+\lambda}\right)^2. \quad (5)$$

The first term comes from unavoidable noise in the test example. The second term originates from the discrepancy between $\hat{\mathbf{w}}$ and $\mathbf{w}$. Since increasing λ shrinks $\hat{\mathbf{w}}$ away from $\mathbf{w}$ leading to underfitting, this second term increases with increasing regularization. The third term originates primarily from the noise in the training data. Since increasing λ reduces the sensitivity of $\hat{\mathbf{w}}$ to this training noise, this third term decreases with increasing regularization. Balancing underfitting versus noise reduction sets an optimal λ. Increasing (decreasing) the SNR increases (decreases) the weight of the second term relative to the third, tilting the balance in favor of a decreased (increased) optimal λ.

Our main result is that in the high dimensional anisotropic setting we obtain a similar formula

$$\mathcal{F} = f_n + \frac{1}{\rho_f}\sum_{i=1}^{P}\left\{f_s\hat{\mathbf{v}}_i^2\left(\frac{\tilde{\lambda}}{S_i^2+\tilde{\lambda}}\right)^2 + f_n\frac{1}{\alpha}\frac{1}{P}\left(\frac{S_i^2}{S_i^2+\tilde{\lambda}}\right)^2\right\}, \quad (6)$$

but with several fundamental modifications compared to the low dimensional setting, as can be seen by comparing (5) and (6). First the original regularization parameter λ is replaced with an effective regularization parameter $\tilde{\lambda}$. Second, the single scalar mode is replaced with an average over the P eigenmodes of Σ. Third, the scalar measurement density N is converted to the high dimensional measurement density α. Fourth, there is an excess multiplicative factor on the last two terms of the error that increases as the fractional participation ratio ρ_f decreases. We now define and discuss each of these important elements in turn.

The Effective Regularization $\tilde{\lambda}$. As we show through extensive calculations in SM (which we sketch in section 3.7) a key quantity that governs the performance of ridge regression in high dimensions is the inverse Hessian of the cost function in (2). This inverse Hessian appears in the estimate of $\hat{\mathbf{w}}$ in (3) and is given by $\mathbf{B} := \left(\frac{1}{N}\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_P\right)^{-1}$. A closely related matrix is $\tilde{\mathbf{B}} := \left(\frac{1}{N}\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I}_N\right)^{-1}$. Indeed the two matrices have identical spectra except for the number of eigenvalues equal to λ, corresponding to the zero eigenvalues of $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}\mathbf{X}^T$. The effective regulariza-

tion $\tilde{\lambda}$ can be expressed in terms of the spectrum of $\tilde{\mathbf{B}}$ via

$$\tilde{\lambda} := 1/\left(\frac{1}{N}\text{Tr}\,\tilde{\mathbf{B}}\right). \tag{7}$$

In the high dimensional limit, $\tilde{\lambda}$ converges to the solution of

$$\lambda = \tilde{\lambda} - \frac{1}{\alpha}\frac{1}{P}\sum_{j=1}^{P}\frac{\tilde{\lambda}S_j^2}{\tilde{\lambda}+S_j^2}, \tag{8}$$

as we prove in SM . We will explore the properties of the effective regularization and its dependence on λ, α, and the spectrum of $\mathbf{\Sigma}$ in more detail in Sec. 3.3.

The Target-Input Alignment $\hat{\mathbf{v}}_i^2$. In the third term, all P modes are equally averaged via the factor $\frac{1}{P}$ while in the second term, each mode is averaged with a different weight $\hat{\mathbf{v}}_i^2$ which captures the alignment between the target function $\mathbf{w}$ and the input distribution. To define this weight, let the eigendecomposition of the true data covariance be given by $\mathbf{\Sigma} = \mathbf{U}\mathbf{S}^2\mathbf{U}^T$. The total signal power can be written $\mathbf{w}^T\mathbf{\Sigma}\mathbf{w} = \sum_i \mathbf{v}_i^2$, where $\mathbf{v} = \mathbf{S}\mathbf{U}^T\mathbf{w}$, so that the components $\mathbf{v}_i^2$ can be interpreted as the signal power in the i^{th} mode of the data. $\hat{\mathbf{v}}$ in (6) is defined to be the unit vector in the same direction as $\mathbf{v}$, so the components $\hat{\mathbf{v}}_i^2$ quantify the *fractional* signal power in the i^{th} mode. Both $\mathbf{v}^T\mathbf{v} = \mathbf{w}^T\mathbf{\Sigma}\mathbf{w}$ and $\hat{\mathbf{v}}^T\hat{\mathbf{v}} = 1$ are $O(1)$, so $\mathbf{v}_i^2$ and $\hat{\mathbf{v}}_i^2$ are $O(1/P)$, and so the second term in (6) has a well defined high dimensional limit. Thus the second term in (6) can be thought of as a weighted average over each mode i, where the weight $\hat{\mathbf{v}}_i^2$ is the fractional signal power in that mode.

The Fractional Participation Ratio ρ_f. We define the participation ratio ρ of the spectrum of $\tilde{\mathbf{B}}$ as

$$\rho := \frac{\left(\text{Tr}\,\tilde{\mathbf{B}}\right)^2}{\text{Tr}\,\tilde{\mathbf{B}}^2}. \tag{9}$$

ρ measures the number of active eigenvalues in $\tilde{\mathbf{B}}$ in a scale-invariant manner, and always lies between 1 (when $\tilde{\mathbf{B}}$ has a single nonzero eigenvalue) and N (when $\tilde{\mathbf{B}}$ has N identical eigenvalues) (see SM). The fractional participation ratio is then $\rho_f := \rho/N$ which satisfies $\frac{1}{N} \le \rho_f \le 1$. For a typical spectrum, the numerator in (9) is $O(N^2)$ and the denominator is $O(N)$, so $\rho = O(N)$ and $\rho_f = O(1)$. We also prove a relation between ρ_f and $\tilde{\lambda}$ (see SM): $\frac{d\tilde{\lambda}}{d\lambda} = \frac{1}{\rho_f}$. Thus a high sensitivity of the effective regularization $\tilde{\lambda}$ to changes in the actual regularization λ coincide with reduced participation ratio ρ_f and higher error $\mathcal{F}$ in (6).

3.2. Proof Sketch of High Dimensional Error Formula.

We first insert (3) into (4) and average over all variables except $\mathbf{X}$, which appear inside a matrix inverse. We then expand this matrix inverse as a power series:

$$\frac{1}{N}\mathbf{B} = \left(\mathbf{X}^T\mathbf{X} + \lambda N\mathbf{I}_P\right)^{-1} = -z\sum_n z^n\left(\frac{1}{N}\mathbf{X}^T\mathbf{X}\right)^n, \tag{10}$$

where $z = -\frac{1}{\lambda}$. This series is a generating function for the matrix sequence $A_n = \left(\frac{1}{N}\mathbf{X}^T\mathbf{X}\right)^n$. We show that computing the average of A_n over the training inputs $\mathbf{X}$ reduces to a combinatorial problem of counting weighted paths (weighted by the eigenvalues of $\mathbf{\Sigma}$) of length $2n$ in a bipartite graph with two groups of nodes corresponding to the N samples and P features respectively. In the limit $N, P \to \infty$ for fixed $\alpha = N/P$ we further show that only paths whose (paired) edges form a tree contribute to A_n. Thus we show the matrix inverse in (10) averaged over the training data $\mathbf{X}$ is a generating function for weighted trees embedded in a bipartite graph. We further exploit the recursive structure of such trees to produce a recurrence relation satisfied by the A_n averaged over $\mathbf{X}$. This recurrence relation yields recursive equations for both the generating function in (10) and the effective regularization $\tilde{\lambda}$ in (8). From these recursive equations we also finally obtain the formula for the error $\mathcal{F}$ in (6) averaged over all the training and test data (see SM).

3.3. Properties of Effective Regularization $\tilde{\lambda}$ and Fractional Participation Ratio ρ_f

We show that the effective regularization $\tilde{\lambda}$ is an increasing, concave function of λ satisfying (see SM)

$$\lambda \le \tilde{\lambda} \le \lambda + \frac{\sigma_x^2}{\alpha}. \tag{11}$$

Furthermore, for large λ, $\tilde{\lambda}$ tends to the upper bound in (11) with error $O(1/\lambda)$. Thus, when λ is large enough, $\tilde{\lambda}$ can be thought of as λ plus a constant correction σ_x^2/α which vanishes in the oversampled low dimensional regime of large α.

The important qualitative features of $\tilde{\lambda}$ can be illustrated in the isotropic case, where

$$\tilde{\lambda} = \frac{\lambda + S^2\left(\frac{1}{\alpha}-1\right) + \sqrt{\left(\lambda + S^2\left(\frac{1}{\alpha}+1\right)\right)^2 - 4\frac{1}{\alpha}S^4}}{2}. \tag{12}$$

For $\alpha > 1$, the graph of $\tilde{\lambda}(\lambda)$ rises from the origin and gradually approaches the line $\tilde{\lambda} = \lambda + S^2/\alpha$, coinciding with the upper bound in (11) (Figure 1A; cyan curves). As α is decreased the slope at the origin becomes steeper until, at the critical value $\alpha = 1$, the slope $\frac{d\tilde{\lambda}}{d\lambda}$ at $\lambda = 0$ is infinite. After this point for $\alpha < 1$, $\tilde{\lambda}$ has nonzero y-intercept (Figure 1A, magenta curves; B shows zoomed in view with critical $\alpha = 1$ curve in black). Figure 1C and D show $\tilde{\lambda}$ and $1/\rho_f$ as a joint function of λ, α. Recalling that $\frac{1}{\rho_f} = \frac{d\tilde{\lambda}}{d\lambda}$, the

bright bar in D representing small ρ_f corresponds exactly to $\tilde{\lambda}$'s infinite derivative at $\alpha = 1$ (Figure 1B; C, dashed line). Thus large sensitivity in $\tilde{\lambda}$ to changes in λ corresponds to a small fractional participation ratio ρ_f, which in turn leads to increased error in (6). For a covariance with multiple distinct scales, $\tilde{\lambda}$ behaves analogously, and can attain near infinite slope at more than one critical value of α (two scale model shown in Figure 1E,F). We observe in 3.6 that this can lead to multiple descent curves, and in 3.7 we show how these effects can be understood in terms of the spectrum of the inverse Hessian $\mathbf{B}$.

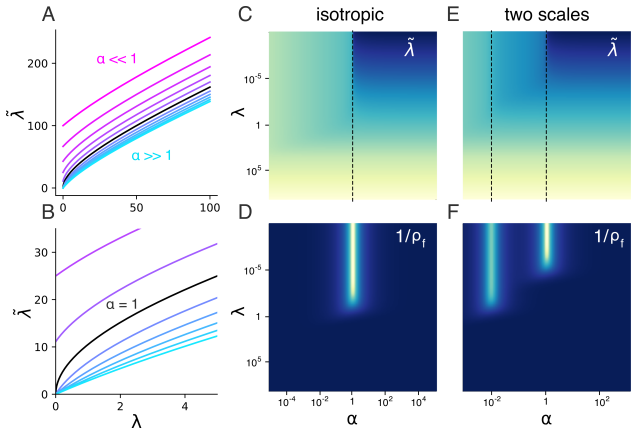

Figure 1. $\tilde{\lambda}$ is the effective regularization parameter satisfying (8). A: $\tilde{\lambda}$ vs λ for different values of α. In all cases $\tilde{\lambda}$ tends to λ plus a constant positive correction factor (see (11)). In the undersampled regime $\alpha \ll 1$, the intercept is nonzero (magenta traces), while in the oversampled regime the intercept is zero (cyan traces). B: Zooming in on the traces in A shows that at $\alpha = 1$ the slope $\frac{d\tilde{\lambda}}{d\lambda}$ becomes infinite, separating the cases with zero and nonzero intercept. Right heatmaps: upper row, $\tilde{\lambda}$ as a function of α, λ; bottom row, $1/\rho_f$. Brighter colors indicate higher values. For the isotropic case with $S^2 = 1$ (C,D), and similarly for slightly anisotropic cases, $\tilde{\lambda}, \rho_f$ behave as the traces shown in A,B. For two widely separated scales ($S_1, S_2 = 1, 10^{-2}$, and $\frac{P_1}{P}, \frac{P_2}{P} = \frac{1}{101}, \frac{100}{101}$; these values were chosen to illustrate the behavior of $\tilde{\lambda}$. For all other figures, $P_1 = P_2$ unless stated otherwise.) (E,F), $\tilde{\lambda}$ has a near infinite derivative at two critical values of α. We show below that this has consequences for the error when fitting with highly anisotropic data.

3.4. Input Anisotropy Increases Error for Generic Target Functions

We first address how anisotropy in $\mathbf{\Sigma}$ alone affects performance when there is no special alignment between $\mathbf{w}$ and $\mathbf{\Sigma}$. To this end we draw $\mathbf{w}$ from a zero mean Gaussian with covariance $\mathbf{C} = \sigma^2 \frac{SNR_0}{P} \mathbf{\Sigma}^{-1}$. This choice essentially fixes the expected $SNR = \frac{\mathbf{w}^T \mathbf{\Sigma} \mathbf{w}}{\sigma^2}$ to the constant value SNR_0 and spreads it evenly across the modes, with the expected value of the fractional SNR per mode $\hat{\mathbf{v}}_i^2 = 1/P$ in (6).

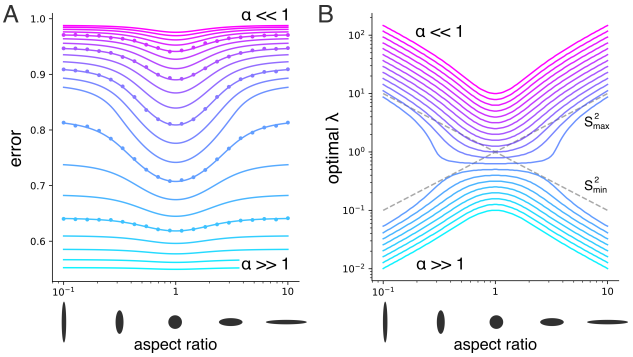

Figure 2. Learning a random function is easiest with isotropic data. A: Error as a function of data anisotropy in a model with 2 distinct variances (diagrams below x-axis show relative length scales of the data distribution). Different traces correspond to different measurement densities α. λ is optimized (either empirically or using the error formula (6)) for each α and aspect ratio. Across different α's, error is minimized on isotropic input data distributions. Dots showing average error from actual ridge regression at 5 α's ($P_1 = P_2 = 50$, $P = 100$ and $N = 25, 50, 100, 200, 400$) closely match theoretical curves. B: Optimal λ as a function of the data distribution's aspect ratio. Traces and colors correspond to A. Dashed lines show the value of the large and small variance. For highly anisotropic data, the optimal λ jumps suddenly from $\propto S^2_{min}$ to $\propto S^2_{max}$ as α increases (note the sudden jump in cyan to magenta curves (decreasing α) at the extremal aspect ratios).

For a two-scale model the main parameter of interest is the aspect ratio $\gamma = S_1/S_2$. Figure 2A shows the error $\mathcal{F}$ in (6), optimized over λ, as a function of γ (with total SNR held constant as γ varies), for different values of the measurement density α. The dots correspond to results from ridge regression, which shows excellent agreement with the theoretical curves. For all α, $\mathcal{F}$ is minimized when $\gamma = 1$, demonstrating that for a generic unaligned target function, error at constant SNR is minimized when the data is isotropic.

Figure 2B shows the optimal λ as a function of aspect ratio. As expected, optimal regularization decreases with α, reflecting the fact that as the number of training examples increases, the need to regularize decreases. More interestingly, crossing the boundary between undersampled and oversampled around $\alpha = O(1)$, the optimal regularization undergoes a sudden jump from $\approx S^2_{max}$ to $\approx S^2_{min}$ (diagonal dashed lines). Consistent with this, we show in SM that when the aspect ratio is very large or small, the optimal regularization λ^* takes the following values in the under- and over-sampled limits:

$$\lambda^* \to \begin{cases} \frac{1}{\alpha} \frac{1}{SNR} S^2_{min} & \alpha \gg 1 \\ \frac{1}{\alpha} \left(\frac{1}{2} + \frac{1}{SNR} \right) S^2_{max} & \alpha \ll 1. \end{cases} \tag{13}$$

This can be roughly understood as follows: in the undersampled regime, where strong regularization is required to control noise, $\tilde{\lambda}$ should be increased until the third term in (6), $\left(\frac{S_i^2}{S_i^2+\tilde{\lambda}}\right)^2$, stops improving, around $\tilde{\lambda} \approx S_{max}^2$. On the other hand, in the oversampled regime, where the weights can be estimated reliably and the biggest source of error is underfitting from the regularization itself, $\tilde{\lambda}$ should be reduced until the second term $\left(\frac{\tilde{\lambda}}{S_i^2+\tilde{\lambda}}\right)^2$ stops improving, around $\tilde{\lambda} \approx S_{min}^2$. This argument provides intuition for why the optimal regularization λ^* transitions from approximately S_{min}^2 to S_{max}^2 as α decreases.

3.5. Weight-Data Alignment Reduces Sample Complexity and Changes the Optimal Regularization

We now explore the effect of alignment on sample complexity and optimal regularization λ. First, if the total SNR is held constant, error decreases as the true weights $\mathbf{w}$ become more aligned with the high variance modes of $\mathbf{\Sigma}$. To see this, note that fixed SNR implies $\mathbf{w}^T\mathbf{\Sigma}\mathbf{w} = \sigma^2 SNR$, so both f_s, f_n are constant. Thus $\mathcal{F}$ can only change through the unit vector $\hat{\mathbf{v}}_i$ in the second term in (6), where the unnormalized alignment vector is $\mathbf{v} = \mathbf{S}\mathbf{U}^T\mathbf{w}$. Thus $\mathcal{F}$ is minimized by aligning $\mathbf{w}$ with the highest variance direction of the data for *any* regularization λ *even if* the magnitude of $\mathbf{w}$ must be reduced to keep the total SNR fixed.

Consequently the sample complexity required to achieve a given performance level can decrease with increasing weight-data alignment even if total SNR is fixed. Figure 3A illustrates this in a two-scale model with aspect ratio $\gamma = S_1/S_2 = 10$. Learning curves of $\mathcal{F}$ as a function of measurement density α for optimal λ are plotted for different weight-data alignments θ, defined as the angle between $\hat{\mathbf{v}}$ and the high variance subspace. The superimposed dots correspond to results from numerical ridge regression and show excellent agreement with the theoretical curves. To achieve any fixed error, models with greater weight-data alignment (blue traces) require fewer samples than those with less weight-data alignment (red traces).

How does weight-data alignment affect the optimal regularization λ? As we show in Figure 3 B,C, this depends on the measurement density α. In the undersampled regime (B), as alignment increases (moving from red to blue traces) the optimal λ achieving minimal error decreases, while in the oversampled regime (C), the optimal λ increases. Figure 3D shows that this trend holds more generally. Each curve shows the optimal λ as a function of alignment. Curves are decreasing in the undersampled regime where $\alpha \ll 1$ (magenta) and increasing in the oversampled regime where $\alpha \gg 1$ (cyan). In SM , we derive the following formula for the optimal lambda in the over- and under-sampled regimes:

$$\lambda^* \to \begin{cases} \frac{1}{\alpha}\sigma^2 \frac{\frac{1}{P}\mathrm{Tr}\left[\mathbf{\Sigma}^{-1}\right]}{\mathbf{w}^T\mathbf{\Sigma}^{-1}\mathbf{w}} & \alpha \ll 1 \\ \frac{1}{\alpha}\sigma^2\left(1+SNR\right)\frac{\frac{1}{P}\mathrm{Tr}\left[\mathbf{\Sigma}^2\right]}{\mathbf{w}^T\mathbf{\Sigma}^2\mathbf{w}} - \frac{\sigma_x^2}{\alpha} & \alpha \gg 1 \end{cases} \tag{14}$$

These approximate formulas are plotted in Figure 3D (dashed lines), and show excellent agreement with the exact optima at extremal α values (solid). The two approximate formulas depend on $\mathbf{w}$ through $\mathbf{w}^T\mathbf{\Sigma}^{-1}\mathbf{w}$ and $\mathbf{w}^T\mathbf{\Sigma}^2\mathbf{w}$, explaining why the first set of curves decreases with alignment while the other increases (see SM for details).

Figure 3E quantifies how helpful regularization is, by comparing the error at optimal λ to that of no regularization with $\lambda = 0$. The effect of alignment again depends on α. In the undersampled regime (magenta), regularization is most helpful for misaligned cases, while in the oversampled regime (cyan) regularization is most helpful for aligned cases. Figure 3F shows the optimal λ as a function of both α, θ. For low α, increasing θ increases the optimal λ, while the opposite is true for high α. The behavior switches sharply at a curved boundary in α, λ space.

3.6. Multiple Descent Curves from Anisotropic Data

It is thought that the performance of a general regression algorithm should improve as the number of training samples increases. However this need not be true if regularization is not tuned properly. For small regularization values, the test error increases - and in fact becomes infinite in the high dimensional limit - around $\alpha = 1$, before eventually monotonically decreasing. This is one example of a phenomenon known as "double descent".

We next show how, when the input data is very anisotropic, learning curves can exhibit one, two, or in fact *any* number of local peaks, a phenomenon we refer to as "multiple descent". We first consider a $D = 2$ scale model with highly disparate scales $S_1 = 1, S_2 = 10^{-2}$ with $\frac{P_1}{P} = \frac{P_2}{P} = \frac{1}{2}$. Figure 4A shows the learning curves for this model, where each black trace corresponds to a fixed value of λ. For very low values of λ (orange trace), the learning curve shows 2 strong peaks before descending monotonically, corresponding to "triple descent". We confirm that this effect can be seen in numerical ridge regression experiments (orange dots). This effect disappears when λ is optimized separately for each value of α (red trace).

In the isotropic regime, singularities in the error typically arise when the number of samples is equal to the number of parameters, that is, $\alpha = 1$. Interestingly, for the 2-scale model, approximate singularities appear at $\alpha = 1$ and $\alpha = \frac{1}{2}$, suggesting that singularities can appear generically whenever the number of samples is equal to the number of parameters *with "large" variance*: when $\alpha = \frac{1}{2}$, there are exactly as many samples as parameters at the larger scale

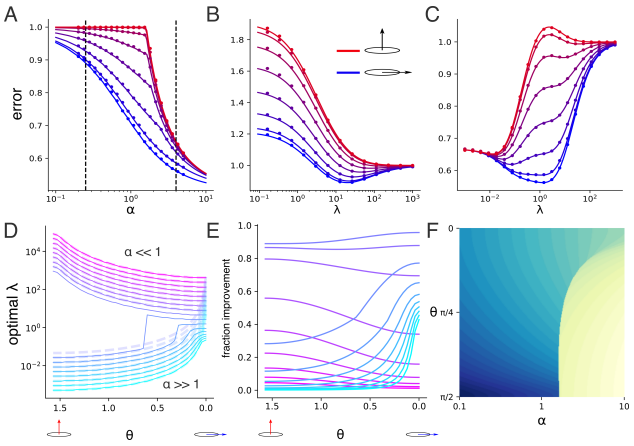

Figure 3. Weight-data alignment improves sample complexity and changes the optimal regularization. A: Learning curves showing error $\mathcal{F}$ vs measurement density α for fixed SNR. Different curves correspond to different weight data alignment θ. Performance is optimized over the regularization λ for each α and alignment. Error decreases faster when weights and data are aligned (blue traces) than when they are misaligned (red traces). B,C: error vs. λ for the same θs as in A, but α fixed to $1/4$ (B), or 4 (C), corresponding to the dashed lines in A. For undersampled cases (B), optimal λ decreases with alignment, while for oversampled cases (C), optimal λ increases with alignment. Dots in panels A-C show average error from numerical ridge regression and closely match theoretical curves. For A, $P = 100$ and N is sampled log-evenly from 20 to 500. For B, $P = 200, N = 50$; for C, $P = 50, N = 200$. D: Trend shown in B,C holds more generally: for $\alpha \ll 1$, optimal λ decreases with alignment, while for $\alpha \gg 1$, optimal λ increases with alignment. Approximate formulas in (14) (dashed lines) match exact curves well. E: Error improvement afforded by optimal regularization as a function of alignment. Just as for the optimal regularization value, oversampled ($\alpha \gg 1$) and undersampled ($\alpha \ll 1$) show opposite dependence on θ. F: Higher resolution view of optimal regularization value as a function of θ, α. Colormap is reversed relative to other figures for visual clarity: dark blue corresponds to high values and light yellow corresponds to low values (cf. panel D). Note the sharp transition between the increasing and decreasing phases.

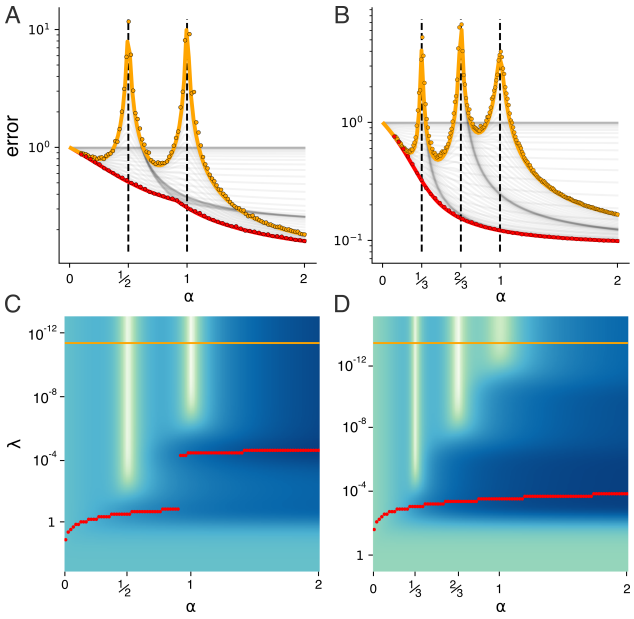

Figure 4. Multiple descent curves emerge from highly anisotropic data. A: learning curve for $\lambda \approx 0$ (orange trace) shows triple descent as a function of measurement density α (i.e. 3 disjoint regions of α where error descends with increasing α). Error peaks occur at $\frac{1}{2}$ and 1. Black traces correspond to different fixed λ. The learning curve with λ optimized for each α instead decreases monotonically (red trace). B: Similar plot showing quadruple descent with peaks at approximate locations $\alpha = \frac{1}{3}, \frac{2}{3}, 1$. Dots in A,B show average error from numerical simulations of ridge regression and closely match theoretical curves ($P = 100$ and N is sampled evenly from 10 to 200). C,D: Global view of the top plots showing error as a function of λ and α. Light bars correspond to error peaks. Horizontal orange (red) slices correspond to orange (red) traces in top row. The kink in the red curve in panel A corresponds to the optimal λ suddenly jumping from one local minimum to another as α is increased (the discontinuity in the red curve in panel C).

$S_1 = 1$, and at $\alpha = 1$, there are as many samples as parameters at the largest two scales S_1, S_2 (which, in this case, is all the parameters).

As a test of this intuition, we show the learning curves for a three scale model (with scales $S_1, S_2, S_3 = 10^{-1}, 10^{-3}, 10^{-5}$) in Figure 4B. As expected, for low λ (orange trace), we see sharp increases in the error when $\alpha = \frac{1}{3}, \frac{2}{3}$, and 1, that is, when the number of samples N equals the number of parameters at the top one (S_1), two (S_1, S_2), or three (S_1, S_2, S_3) scales.

Figure 4C,D show heatmaps of the error $\mathcal{F}$ as a joint function of α and λ. The orange and red slices correspond to the low λ (orange) and optimal λ (red) traces in A,B. The bright vertical bars corresponding to high error at the critical values of α can be completely avoided by choosing λ appropriately, showing how multiple descent can be thought of as an artifact from inadequate regularization. Figure 4C also illustrates how, even when regularizing properly, learning curves can display "kinks" (Figure 3A, 4A) around the critical values of α as the optimal regularization jumps from one local minimum to another. In the next section, we give a more detailed explanation of how the spectrum of the inverse Hessian $\mathbf{B}$ can explain these effects.

3.7. Random Matrix Theory Explains Multiple Descent

We now sketch how multiple descent can be understood in terms of the spectrum of the inverse Hessian $\mathbf{B}$, or as noted

above, of the matrix $\tilde{\mathbf{B}} = \left(\frac{1}{N}\mathbf{X}\mathbf{X}^T + \lambda \mathbf{I}_N\right)^{-1}$, which contains the same information. For detailed proofs see SM . First, a key quantity encoding information about the spectral density of $\frac{1}{N}\mathbf{X}\mathbf{X}^T$, which we'll call $\rho(x)$, is the Stieltjes transform defined as $G(x) = \int \frac{\rho(t)}{x-t} dt$ (Nica & Speicher, 2006). The spectrum $\rho(x)$ can be recovered from G through the inversion formula $\rho(x) = \frac{1}{i\pi} \lim_{\epsilon \to 0^+} G(x + i\epsilon)$. To obtain an equation for G, we use (8) and the fact that $G(\lambda) = -\frac{1}{N}\mathrm{Tr}\,\tilde{\mathbf{B}}(-\lambda) = -\frac{1}{\lambda(-\lambda)}$, giving the following:

$$\frac{1}{G} - \frac{1}{\alpha}\frac{1}{P}\sum_{i=1}^{P} \frac{S_i^2}{S_i^2 G - 1} = \lambda. \tag{15}$$

Although (15) is difficult to solve for general α, λ, S_i, we can obtain approximate formulas for the spectrum when the scales S_i are very different from one another, corresponding to highly anisotropic data. We also show in SM how to obtain exact values for the *boundaries* of the spectrum using the discriminant of (15) even when the scales are not well separated. We state the main results (see SM for detailed proofs):

Density for Widely Separated Scales. Consider a D-scale model where $\mathbf{\Sigma}$ has P_d eigenvalues with value S_d^2 for $d = 1, \ldots, D$. Define $f_d := P_d/P$. Assume the scales are arranged in descending order, and are very different from one another - that is, $\epsilon_d^2 = S_{d+1}^2/S_d^2 \ll 1$. In the limit of small ϵ_d, the spectral density ρ consists of D disjoint components ρ_d, roughly centered on the D distinct scales S_d^2, satisfying

$$\rho_d(x) = \frac{\sqrt{(x_+ - x)(x - x_-)}}{2\pi S_d^2 \lambda}$$
$$x_\pm = S_d^2 \left(1 - \frac{1}{\alpha}\sum_{d'<d} f_{d'}\right)\left(1 \pm \sqrt{\frac{f_d}{\alpha - \sum_{d'<d} f_{d'}}}\right)^2 \tag{16}$$

The d^{th} component is proportional to a Marchenko-Pastur density supported on $[x_-, x_+]$ and exists for values of $\alpha > \sum_{d'<d} f_{d'}$. This result reveals that as α increases one encounters a sequence of phase transitions in the spectrum of $\tilde{\mathbf{B}}$; at each transition a successively finer scale in the data becomes visible and the spectrum associated with the previous already visible scale acquires an extended tail just before the transition. These essential spectral features can be illustrated in the balanced 2-scale model with $S_1 = 1, S_2 = 10^{-2}$, whose learning curves were previously shown to exhibit triple descent in Figure 4A,C. Figure 5A shows the empirical histograms of the nonzero eigenvalues of randomly generated $\frac{1}{N}\mathbf{X}\mathbf{X}^T$ for several values of α (colored histograms; log scale), which show excellent agreement with the densities predicted above (black traces). As α increases we see a phase transition from one large visible scale to two visible scales, both the large and the small one. This transition occurs at the predicted critical value $\alpha = 1/2$ by the formulas above, and a precursor to this transition is a large spread in the spectrum.

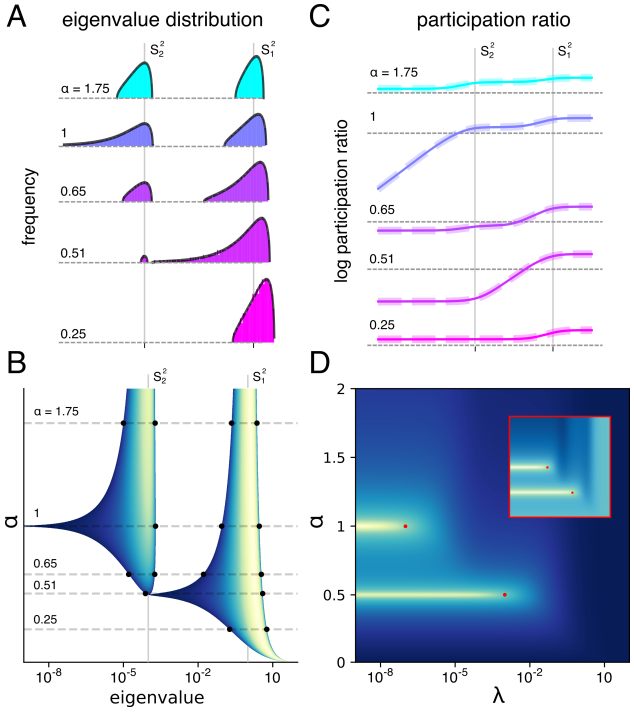

Figure 5. Changes in the spectrum of $\frac{1}{N}\mathbf{X}\mathbf{X}^T$ explain multiple descent. All panels show behavior of the 2-scale model with $S_1 = 1, S_2 = 10^{-2}$ shown in Figure 4A,C. A: empirical histograms for nonzero eigenvalues of $\frac{1}{N}\mathbf{X}\mathbf{X}^T$ for 5 values of α (colored histograms; log scale). Black traces show approximate formulas for density for widely separated scales (see (16)). Distinct components center on the scales S_1^2, S_2^2. B: Heatmap of the eigenvalue density. Each horizontal line is normalized by its maximum value to allow comparison for different values of α. Dashed lines correspond to the values of α sampled in A. Note rapid changes in the spectrum around $\alpha = \frac{1}{2}, 1$. C: Log of the participation ratio ρ_f vs λ for the same values of α as in A. For small values of λ, the participation ρ_f becomes very small around the critical values $\alpha = \frac{1}{2}, 1$. Horizontal dashed lines (corresponding to $\rho_f = \frac{1}{\sqrt{10}}$ for each trace) are drawn to aid comparison. D: Inverse participation ratio vs α, λ. $\frac{1}{\rho_f}$ shows strong increases around the critical α, explaining the local increases in error seen in the multiple descent curves in Figure 4A,C. The error heatmap in Figure 4C is copied in the red inset. Red dots show corresponding points in the $\alpha - \lambda$ plane.

Figure 5B shows the spectral density for all $\alpha \in [0, 2]$. The horizontal dashed lines show the values of α sampled in A, and the black dots show the boundaries of the empirical histograms. Consistent with A, the density undergoes phase transitions at critical values of $\alpha = \frac{1}{2}, 1$. Intriguingly, these critical values correspond exactly to the α values where the triple descent curves in Figure 4A achieve their local

maxima. The reason for this connection lies in the fact that the inverse fractional participation ratio $1/\rho_f$ appears in the error $\mathcal{F}$ in (6) and $1/\rho_f$ can be written as $\frac{1}{\rho_f} = \left(\frac{\sigma_\gamma}{\mu_\gamma}\right)^2 + 1$ where μ_γ and σ_γ are the mean and standard deviation of the nonzero spectrum of $\mathbf{B} = \left(\frac{1}{N}\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_P\right)^{-1}$. Thus when the spectrum of $\mathbf{X}\mathbf{X}^T$ is spread out, ρ_f is small and $\mathcal{F}$ is large. We confirm this intuition in (Figure 5C) which shows a match between theory (with the ρ_f calculated analytically in SM) and experiment (with ρ_f calculated numerically for random matrices). Furthermore, both theory and experiment indicate that ρ_f at small λ drops precipitously as a function of α at precisely the critical values of α (Figure 5C) at which the triple descent curves in Figure 4A peak. Indeed plotting $1/\rho_f$ as a joint function of α and λ matches exceedingly well in terms of the location of peaks, the error $\mathcal{F}$ as a joint function of α and λ (see Figure 5D). Thus the structure of phase transitions in the spectrum of the random matrix $\frac{1}{N}\mathbf{X}\mathbf{X}^T$ drawn from a true covariance $\mathbf{\Sigma}$ with multiple scales can explain the emergence of multiple descent, with a one to one correspondence between the number of widely separated data scales and the number of peaks.

4. Discussion

Thus we obtain a relatively complete analytic theory for a widespread ML algorithm in the important high dimensional statistical limit that takes into account multi-scale anisotropies in inputs that can be aligned in arbitrary ways to the target function to be learned. Our theory shows how and why successful generalization is possible with very little data when such alignment is high. We hope the rich mathematical structure of phase transitions and multiple descent that arises when we model correlations between inputs and target functions and their impact on generalization performance motivates further research along these lines in other settings, in order to better bridge the gap between the theory and practice of successful generalization.

References

Advani, M. and Ganguli, S. Statistical mechanics of optimal convex inference in high dimensions. *Physical Review X*, 2016a.

Advani, M. and Ganguli, S. An equivalence between high dimensional bayes optimal inference and m-estimation. *Adv. Neural Inf. Process. Syst.*, 2016b.

Advani, M., Lahiri, S., and Ganguli, S. Statistical mechanics of complex neural systems and high dimensional data. *J. Stat. Mech: Theory Exp.*, 2013(03):P03014, 2013.

Bahri, Y., Kadmon, J., Pennington, J., Schoenholz, S. S., Sohl-Dickstein, J., and Ganguli, S. Statistical mechanics of deep learning. *Annual Review of Condensed Matter Physics*, March 2020.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1903070116. URL https://www.pnas.org/content/116/32/15849.

Canatar, A., Bordelon, B., and Pehlevan, C. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12(1), May 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-23103-1. URL http://dx.doi.org/10.1038/s41467-021-23103-1.

Chen, L., Min, Y., Belkin, M., and Karbasi, A. Multiple descent: Design your own generalization curve. *arXiv*, 2021. URL https://arxiv.org/abs/2008.01036.

d'Ascoli, S., Sagun, L., and Biroli, G. Triple descent and the two kinds of overfitting: Where why do they appear? *arXiv*, 2020. URL https://arxiv.org/abs/2006.03509.

Donoho, D. L., Maleki, A., and Montanari, A. Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci.*, 106(45):18914, 2009.

Dziugaite, G. K. and Roy, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv*, 2017. URL https://arxiv.org/abs/1703.11008.

Engel, A. and den Broeck, C. V. *Statistical Mechanics of Learning*. Cambridge Univ. Press, 2001.

Ganguli, S. and Sompolinsky, H. Short-term memory in neuronal networks through dynamical compressed sensing. In *Neural Information Processing Systems (NIPS)*, 2010a.

Ganguli, S. and Sompolinsky, H. Statistical mechanics of compressed sensing. *Phys. Rev. Lett.*, 104(18):188701, 2010b.

Gerace, F., Loureiro, B., Krzakala, F., Mézard, M., and Zdeborová, L. Generalisation error in learning with random features and the hidden manifold model. *arXiv*, 2020. URL https://arxiv.org/abs/2002.09339.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. When do neural networks outperform kernel methods? In Larochelle, H., Ranzato, M.,

Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 14820–14830. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/a9df2255ad642b923d95503b9a7958d8-Paper.pdf.

Goldt, S., Mézard, M., Krzakala, F., and Zdeborová, L. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10(4), Dec 2020. ISSN 2160-3308. doi: 10.1103/physrevx.10.041044. URL http://dx.doi.org/10.1103/PhysRevX.10.041044.

Lampinen, A. K. and Ganguli, S. An analytic theory of generalization dynamics and transfer learning in deep linear networks. In *International Conference on Learning Representations (ICLR)*, 2018.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521(7553):436–444, May 2015.

Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv*, 2020. URL https://arxiv.org/abs/1908.05355.

Monasson, R. and Zecchina, R. Weight space structure and internal representations: a direct approach to learning and generalization in multilayer neural networks. *Phys. Rev. Lett.*, 75(12):2432–2435, 1995.

Nica, A. and Speicher, R. *Lectures on the Combinatorics of Free Probability*. London Mathematical Society Lecture Note Series. Cambridge University Press, 2006. doi: 10.1017/CBO9780511735127.

Rangan, S., Fletcher, A. K., and V.k., G. Asymptotic analysis of MAP estimation via the replica method and applications to compressed sensing. *CoRR*, abs/0906.3234, 2009.

Seung, H. S., Sompolinsky, H., and Tishby, N. Statistical mechanics of learning from examples. *Phys. Rev. A*, 45 (8):6056, 1992.

Vapnik, V. N. *Statistical learning theory*. Wiley-Interscience, 1998.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv*, 2017. URL https://arxiv.org/abs/1611.03530.