Generalized Aliasing Explains Double Descent and Informs Model Design

Mark K. Transtrum,[1, *] Gus L. W. Hart,[1] Tyler J. Jarvis,[2] and Jared P. Whitehead[2]

[1]*Department of Physics and Astronomy, Brigham Young University, Provo, Utah 84602, USA*
[2]*Department of Mathematics, Brigham Young University, Provo, Utah 84602, USA*

A central problem in data science is to use potentially noisy samples of an unknown function to predict function values for unseen inputs. In classical statistics, the predictive error is understood as a trade-off between the bias and the variance that balances model simplicity with its ability to fit complex functions. However, over-parameterized models exhibit counterintuitive behaviors, such as "double descent" in which models of increasing complexity exhibit *decreasing* generalization error. Other models may exhibit more complicated patterns of predictive error with multiple peaks and valleys. Neither double descent nor multiple descent phenomena are well explained by the bias–variance decomposition.

We introduce a novel decomposition that we call the *generalized aliasing decomposition* (GAD) to explain the relationship between predictive performance and model complexity. The GAD decomposes the predictive error into three parts: 1.) *model insufficiency*, which dominates when the number of parameters is much smaller than the number of data points, 2.) *data insufficiency*, which dominates when the number of parameters is much greater than the number of data points, and 3.) *generalized aliasing*, which dominates between these two extremes.

We demonstrate the applicability of the GAD to diverse applications, including random feature models from machine learning, Fourier transforms from signal processing, solution methods for differential equations, and predictive formation enthalpy in materials discovery. Because key components of the generalized aliasing decomposition can be explicitly calculated from the relationship between model class and samples without seeing any data labels, it can answer questions related to experimental design and model selection *before* collecting data or performing experiments. We further demonstrate this approach on several examples and discuss implications for predictive modeling and data science.

I. INTRODUCTION

Predictive models allow scientists and engineers to extend data and anticipate outcomes for unseen cases. A key issue for these models is the problem of how to understand and minimize the generalization error. Traditionally, scientists think about generalization error in terms of a trade-off between bias and variance, but that trade-off does not readily predict the error curves for many models, especially models with more parameters than data points and models involving highly structured scientific and engineering data. In this work, we introduce a new decomposition, the *generalized aliasing decomposition (GAD)*, that explains a wide variety of error curves in predictive models for both small (classical) models and for large, over-parametrized models. This decomposition explains complex generalization curves, including double and multiple descent, and can be used to inform the choice of model and experimental design (training points) to control, reduce, or even minimize generalization error.

Some of the fundamental choices when model building are (1) the sample data and (2) the complexity of the model class. Simple models are generally preferred for many reasons, including interpretability and computational expense [1–7], but one of the more pragmatic justifications for parsimony is a desire to balance over- and under-fitting as understood through the bias–variance decomposition. Models with few parameters avoid making wild predictions but under fit the observed data without much fidelity (high bias), while over-parameterized models fit the sampled data well with wild swings in between data points (high variance). The unquestioned goal has been to find the "sweet spot" of model complexity that balances bias and variance, i.e., a faithful model of moderate complexity (see Figure 1, left panel) that minimizes the so-called "risk", that is, errors made by model predictions on unseen data.

While the foregoing story has long been the standard way to approach these problems, we now know this view of the fitting problem is not the whole story. For extremely over-parameterized models (i.e., more parameters than samples), prediction errors may actually *decrease* with additional parameters, a phenomenon often called "double descent" [8], summarized by the left panel in Figure 1. The boundary between the two regimes, where there are as many parameters as data points, is known as the *interpolation threshold*, because it is (generically) the boundary of where the model can perfectly interpolate the training data, but below that threshold interpolation cannot occur. Non-convex risk curves (such as with double and multiple descent) are most famously recognized in neural networks [9, 10], though this behavior has been observed in other settings as well. (See [11] for the bias–variance decomposition for neural networks, [12] for ordinary least-squares regression, and [13] for a thorough review.) Furthermore, models and data sets can be designed to exhibit complex, multiple descent [14–21].
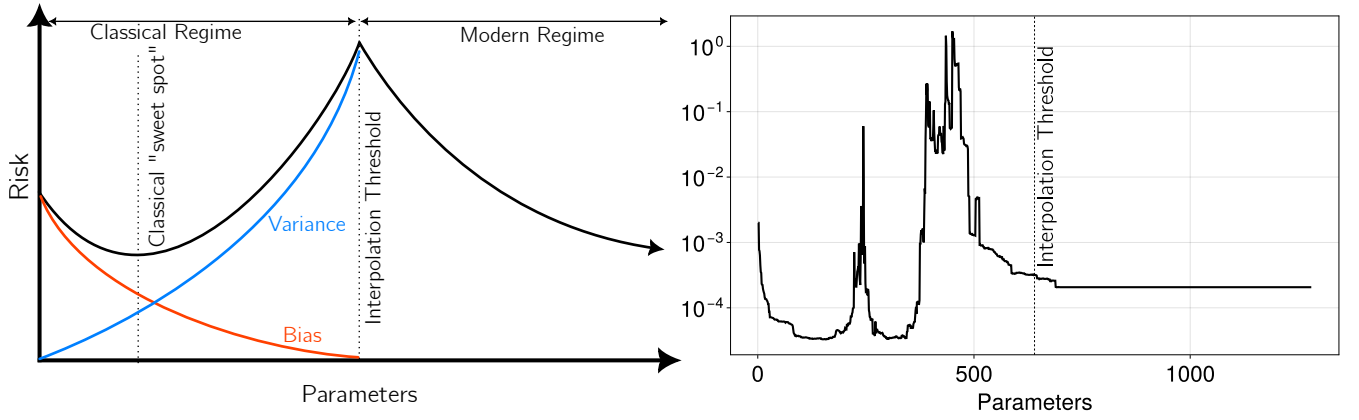
* mktranstrum@byu.edu

FIG. 1. **Limitations of the Bias–Variance Trade-off.** Left: Bias and variance are traditionally understood as monotonically decreasing/increasing contributions to risk to be balanced by tuning model complexity. Double descent illustrates a breakdown of this intuition beyond the interpolation threshold where variance and bias can exhibit counter-intuitive dependence on model class. Right: Highly structured data such as a cluster expansion of alloy formation enthalpy exhibit even more complicated and counter-intuitive dependence on model complexity, not easily explained using the bias–variance trade-off.

Models of highly structured data from scientific and engineering applications often exhibit similar multiple descent behavior. For example, the risk curve in the right panel of Figure 1 comes from a cluster expansion model of the formation enthalpy of alloy structures (see section III D for more detail) in materials science. Not only do the peaks and troughs appear *to the left* of the interpolation threshold where classical bias–variance arguments ought to apply, the naïve interpolation threshold apparently plays no role.

Traditional data analysis techniques are also at odds with the intuition of the bias–variance decomposition. The discrete Fourier transform, for example, is formally equivalent to a regression problem (see section III B) with as many parameters (Fourier coefficients) as data, so bias–variance arguments suggest that the inferred Fourier coefficients should exhibit unreasonable sensitivity to noisy data. In spite of this, the fast Fourier transform which efficiently computes the discrete Fourier transform precisely at the interpolation threshold, is one of the most influential and widely used algorithms in all of science and engineering (even for noisy signals that are not band-limited). Furthermore, techniques such as pseudospectral and collocation methods for solving differential equations are similarly equivalent to regression problems (see Section III C) but are known to exhibit optimal performance at or beyond the interpolation threshold [22].

While the bias–variance decomposition holds as a formal mathematical result, these examples expose the limited insight it provides. Its utility derives from the incorrect expectation that model selection balances the trade-off between monotonically decreasing (bias) and monotonically increasing (variance) error contributions. In reality, the contributions of bias and variance for each of the preceding examples are non-monotonic, complex, and intimately connected with the algorithmic solution to the optimization problem.

Recent work has begun to explain these behaviors, often focusing on regression and the simplest case of double descent, although risk curves may be far more complicated [14]. In [17, 20], the bias–variance decomposition is expanded to explain this non-convex behavior, relying on the interplay between the model design and the actual data. Several other efforts have been made to clarify the relationship between the model class, inherent algorithmic bias, the split between testing and training data, and the appearance of non-monotonic loss and generalization curves. We do not present a comprehensive summary of these efforts, but direct the interested reader to [15, 16, 21, 23], which clarify the nature of double descent and its apparent reliance on the structure of the testing and training data sets.

In contrast to these approaches, we build on insights from signal processing [10] and introduce a new decomposition (Eq. (17)), which we refer to as the *generalized aliasing decomposition* (GAD), summarized for the generic case of double descent in the left panel of Figure 2. The aliasing decomposition explains generic risk curves in both the classical and modern regimes as the contribution of three terms: 1.) model insufficiency, 2.) data insufficiency, and 3.) generalized aliasing.

Model insufficiency quantifies the inability of the model to fit the data (red curve in the left panel of Figure 2). It is usually the dominant error contribution for models with few parameters, and it decreases monotonically with the number of parameters. Though the mapping is not exact, it roughly corresponds to "bias" in the bias–variance paradigm. Adding more parameters to a model does not limit the ability of the model to fit data, so it decreases monotonically as we prove in Section II C 2.

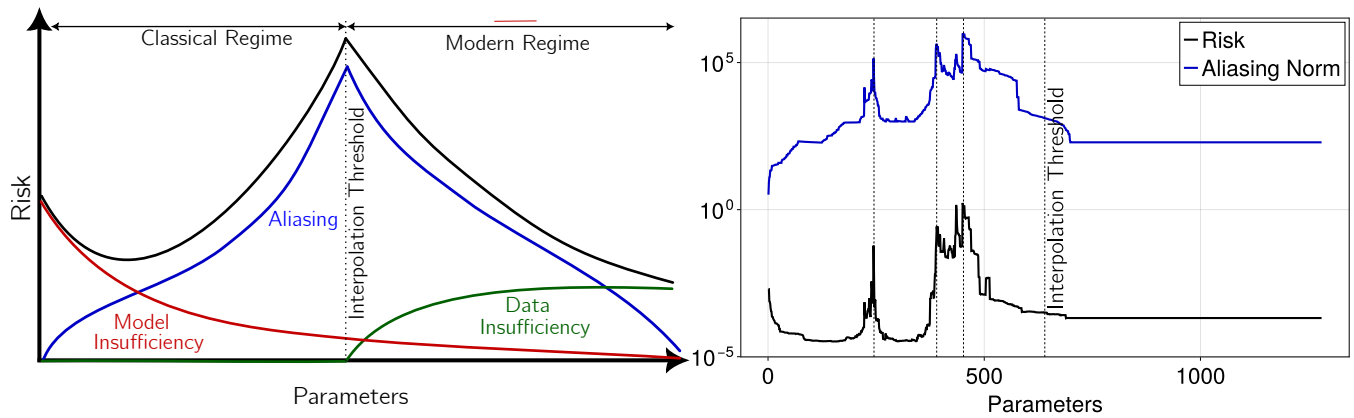*Data insufficiency* quantifies how much model param-

FIG. 2. **Generalized Aliasing Decomposition.** Left: The generalized aliasing decomposition (GAD) expresses risk as the contribution of three terms: model insufficiency, data insufficiency, and generalized aliasing. Model insufficiency dominates for small models and decreases monotonically with the number of parameters. Data insufficiency dominates for large models, increasing monotonically with the number of parameters. Generalized Aliasing accounts for non-convex intermediate behaviors but has a single peak at the interpolation threshold for the generic case, accounting for the phenomenon of Double Descent. Right: For highly structured problems (see Figure 1, right), aliasing explains all of the non-convex behavior of generic risk curves at intermediate model sizes.

eters are unconstrained by available data (green curve in the left panel of Figure 2). It is the dominant error contribution for models with many excess parameters, increasing monotonically as the model grows. Intuitively, adding more parameters does not introduce any additional parametric constraints, and we show this contribution does not increase with additional parameters. However, adding parameters generically also imposes additional data requirements, so data insufficiency generally increases with the number of parameters (see Section II C 1).

Finally, *generalized aliasing* explains all non-monotonic behavior in the intermediate regime (blue curve in the left panel of Figure 2). The name derives from the special case of Fourier aliasing. When high-frequency (noisy) components of a signal cannot be distinguished from low-frequency components at finite sampling rates, high-frequency (unmodeled) contributions are said to *alias* with the low-frequency (modeled) components, corrupting the representation.

In the generic case, aliasing errors are maximized at the interpolation threshold and cause double descent (Figure 2 left). In structured cases such as the cluster expansion example of Section III D, aliasing also fully accounts for the complicated, non-monotonic behavior throughout both the classical and modern regimes (Figure 2, right).

As we demonstrate below, the GAD provides the intuition behind best practices for other analysis techniques, such as pseudospectral approaches to solving differential equations. Although the contribution of generalized aliasing is non-monotonic in the number of parameters, we show its behavior is easily intuited. In Section III D, we use this fact to explain the complicated risk curve in the right panel of Figure 1.

Taken collectively, *the three components of the GAD*

explain all the qualitative features of generic risk curves. The GAD further clarifies the roles of model structure, data sampling, data labels, and the learning algorithm. Indeed, a useful feature of the decomposition is that much of the analysis can be done independently of data labels. Consequently, the GAD facilitates key modeling decisions such as the choice of model class, experimental design, regularization, and learning algorithm.

II. THE GENERALIZED ALIASING DECOMPOSITION (GAD)

In this section we give the details of the GAD and its mathematical justification. We begin with establishing notation to describe the regression problem, define the decomposition, and then describe how the decomposition influences the error or risk of the fitting problem. Several explicit examples and applications are given in Section III.

Readers wishing to focus on the examples and discussion should first read Sections II A and II B where the GAD is defined, but can safely skip over Sections II C 3–II C 5 and all but the first paragraph of Section II C 1.

A. Mathematical Preliminaries

In regression, data $\boldsymbol{y}$ are given at samples of an independent variable t and usually decomposed as the sum of an unknown signal $f_{\boldsymbol{\theta}}(t)$ parameterized by $\boldsymbol{\theta}$ and noise ξ:

$$y_i = f_{\boldsymbol{\theta}}(t_i) + \xi_i, \tag{1}$$

where the subscript i refers to a particular data sample. For linear regression the signal is a linear combination

of basis functions $f(t) = \sum_j \Phi_j(t)\theta_j$, so that the fundamental regression equation (1) becomes

$$\boldsymbol{y} = \mathbf{M}\boldsymbol{\theta} + \boldsymbol{\xi}, \qquad (2)$$

where the design matrix $\mathbf{M}$ is composed of samples of basis functions, $\mathbf{M}_{ij} = \Phi_j(t_i)$.

As an illustration consider a polynomial fit on an interval $[a, b]$, and take the basis functions to be the usual monomial basis $\{1, t, t^2, t^3, \ldots, t^d\}$ for some $d > 0$ [24]; so $\Phi_j = t^{j-1}$, and the design matrix $\mathbf{M}$ is

$$\mathbf{M} = \begin{pmatrix} 1 & t_1 & t_1^2 & \ldots & t_1^d \\ 1 & t_2 & t_2^2 & \ldots & t_2^d \\ \vdots & & & & \vdots \\ 1 & t_n & t_n^2 & \ldots & t_n^d \end{pmatrix}. \qquad (3)$$

Inferred parameter values $\hat{\boldsymbol{\theta}}$ are found by inverting the design matrix. Since $\mathbf{M}$ is generally not square, an appropriate pseudoinverse $\mathbf{M}^+$ is used: $\hat{\boldsymbol{\theta}} = \mathbf{M}^+\boldsymbol{y}$. The Moore–Penrose pseudoinverse is the standard choice for linear regression[25], including in this motivating example. Other cases may require an algorithmic solution, but common algorithmic choices, such as stochastic gradient descent, are known to produce similar norm-minimizing solutions (see [20, 26, 27], for example).

Finally, predictions at unobserved values of the independent variable are constructed

$$\hat{y}(t) = \sum_j \Phi_j(t)\hat{\boldsymbol{\theta}}_j. \qquad (4)$$

An important quantity of interest for validation is the squared error, averaged over a (typically theoretical) distribution of all of the data, not just the training samples. The expectation of the squared error is called *generalization error* or *population risk*

$$R_{\boldsymbol{\theta}}(\hat{\boldsymbol{y}}) = \mathbb{E}[(\boldsymbol{y} - \hat{\boldsymbol{y}})^2], \qquad (5)$$

where the dependence of $R_{\boldsymbol{\theta}}$ on the model class and training data is implicit. A primary goal in data science is to identify the model class and degree of complexity that minimizes this risk (5).

B. Generalized Aliasing

With a common vocabulary established, we now derive the aliasing operator that underpins the generalized aliasing decomposition (GAD). We no longer require that the data points t lie in $\mathbb{R}$; they could belong to any set Ω. But we assume that the model functions $\Phi_j : \Omega \to \mathbb{R}$ may be extended to form a complete set, meaning that the true function $y(t)$ (both signal f and noise ξ) can be uniquely expressed as a convergent series

$$y(t) \sum_j \Phi_j(t)\theta_j. \qquad (6)$$

on the entire domain, not just on the training points. For the example of polynomial regression, if $y(t)$ is a real analytic function, then the infinite monomial basis $\{1, t, t^2, \ldots\}$ is complete and an appropriate extension for this example Said more formally, the noisy signal $y(t)$ is an abstract vector $\boldsymbol{y}$ in a (potentially infinite-dimensional) vector space $\mathcal{D}$ expressed in some Schauder basis $\{\Phi_j\}_{j \in \mathbb{N}}$ as

$$\boldsymbol{y} = \mathbf{M}\boldsymbol{\theta}, \qquad (7)$$

where $\mathbf{M}$ is a bounded linear transformation mapping the vector $\boldsymbol{\theta}$ in the parameter space Θ to $\boldsymbol{y}$ in the data space $\mathcal{D}$. In the case of fitting a polynomial on an interval $[a, b]$, the operator $\mathbf{M}$ could be thought of as a generalized Vandermonde matrix with countably (infinite) many rows corresponding to rational points of $[a, b]$ and countably (infinite) many columns corresponding to t^j for each nonnegative integer j.[28]

Performing linear regression on samples of $y(t)$ and making predictions at unobserved values of t corresponds to partitioning data space $\mathcal{D}$ into a direct sum $\mathcal{T} \oplus \mathcal{P}$ of training $\mathcal{T}$ and prediction $\mathcal{P}$ subspaces. We write $\boldsymbol{y}$ in this decomposition as $\boldsymbol{y} = (\boldsymbol{y}_{\mathcal{T}}, \boldsymbol{y}_{\mathcal{P}})$. We assume that $\mathcal{T}$ has finite dimension n, but $\mathcal{P}$ need not be finite dimensional. The learning problem is this: Given observations in $\boldsymbol{y}_{\mathcal{T}}$, predict the components of $\boldsymbol{y}_{\mathcal{P}}$.

In practice, this is done by similarly partitioning the representation space $\Theta = \mathcal{M} \oplus \mathcal{U}$ into a modeled $\mathcal{M}$ and an unmodeled $\mathcal{U}$ subspace so that $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\mathcal{M}}, \boldsymbol{\theta}_{\mathcal{U}})$, implicitly assuming that $\boldsymbol{\theta}_{\mathcal{U}}$ are negligible. We usually assume that $\mathcal{M}$ has finite dimension m (we have $m = d + 1$ for polynomials of degree at most d), but $\mathcal{U}$ need not be finite dimensional. With these partitions, the relationship of Eq. (7) between data and coordinates takes the block representation described in the definition below.

Definition. *Denote the decomposition of the labeled data as*

$$\begin{pmatrix} \boldsymbol{y}_{\mathcal{T}} \\ \boldsymbol{y}_{\mathcal{P}} \end{pmatrix} = \begin{pmatrix} \mathbf{M}_{\mathcal{TM}} & \mathbf{M}_{\mathcal{TU}} \\ \mathbf{M}_{\mathcal{PM}} & \mathbf{M}_{\mathcal{PU}} \end{pmatrix} \begin{pmatrix} \boldsymbol{\theta}_{\mathcal{M}} \\ \boldsymbol{\theta}_{\mathcal{U}} \end{pmatrix}, \qquad (8)$$

where the linear transformation $\mathbf{M}_{\mathcal{TM}} : \mathbb{R}^m \to \mathbb{R}^n$ *is the usual* design matrix.

By explicitly recognizing the unmodeled blocks in the definition, we emphasize that some contributions to the signal $y(t)$ will remain unknown to us (noise, for example). Essentially, we acknowledge that our model is a subspace of a universal function space which will in turn allow us to reason about the relationship between the modeled and the unmodeled spaces. The significance of this decomposition, as opposed to simply treating unmodeled signal as noise, is discussed further in section IV C. We emphasize that with this definition $\mathbf{M}$ does not denote the classical design matrix. Rather the block $\mathbf{M}_{\mathcal{TM}}$ is the classical design matrix, and $\mathbf{M}$ represents a full basis transformation (with both modeled and unmodeled basis functions) on the complete signal (including both seen

and unseen data). Because $\mathbf{M}$ is bounded and linear, the subblocks $\mathbf{M}_{\mathcal{TU}}$, $\mathbf{M}_{\mathcal{PM}}$, and $\mathbf{M}_{\mathcal{PU}}$ are also bounded linear transformations.

In the case of fitting a polynomial of degree at most d on n training points $t_1, \ldots t_n$, the design matrix (upper left block) $\mathbf{M}_{\mathcal{TM}}$ is the Vandermonde matrix in (3) and the unmodeled training (upper right) block is the semi-infinite matrix

$$\mathbf{M}_{\mathcal{TU}} = \begin{pmatrix} t_1^{d+1} & t_1^{d+2} & \cdots \\ t_2^{d+1} & t_2^{d+2} & \cdots \\ \vdots & \vdots & \\ t_n^{d+1} & t_n^{d+2} & \cdots \end{pmatrix}.$$

The rows of the lower blocks $\mathbf{M}_{\mathcal{PM}}$ and $\mathbf{M}_{\mathcal{PU}}$ correspond to the prediction points, $t_{\mathcal{P}} = [a, b] \setminus \{t_1, \ldots, t_n\}$ (again, a countable dense subset of points in $[a, b] \setminus \{t_1, \ldots, t_n\}$ suffices). The columns of the lower left block $\mathbf{M}_{\mathcal{PM}}$ correspond to the monomials $1, t, t^2, \ldots, t^d$ (spanning the space $\mathcal{M}$) evaluated at the points $t_{\mathcal{P}}$. The columns of the lower right block $\mathbf{M}_{\mathcal{PU}}$ correspond to the unmodeled monomials $t^{d+1}, t^{d+2}, \ldots$ (spanning $\mathcal{U}$), evaluated at points $t_{\mathcal{P}}$.

We learn the modeled parameters $\hat{\boldsymbol{\theta}}_{\mathcal{M}}$ using some pseudoinverse $\mathbf{M}_{\mathcal{TM}}^+$ of the design matrix $\mathbf{M}_{\mathcal{TM}}$:

$$\hat{\boldsymbol{\theta}}_{\mathcal{M}} = \mathbf{M}_{\mathcal{TM}}^+ \, \boldsymbol{y}_{\mathcal{T}}. \qquad (9)$$

Inferring only $\hat{\boldsymbol{\theta}}_{\mathcal{M}}$ is equivalent to assuming that the unmodeled parameters vanish, so $\hat{\boldsymbol{\theta}}_{\mathcal{U}} = \mathbf{0}$ and $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_{\mathcal{M}}, \mathbf{0})$. However, the true representation of the training data $\boldsymbol{y}_{\mathcal{T}}$ includes contributions from both the modeled and unmodeled components of $\boldsymbol{\theta}$:

$$\boldsymbol{y}_{\mathcal{T}} = \mathbf{M}_{\mathcal{TM}} \boldsymbol{\theta}_{\mathcal{M}} + \mathbf{M}_{\mathcal{TU}} \boldsymbol{\theta}_{\mathcal{U}}. \qquad (10)$$

The unmodeled term $\mathbf{M}_{\mathcal{TU}} \boldsymbol{\theta}_{\mathcal{U}}$ corresponds to the noise in Eq. (1). The inferred parameters $\hat{\boldsymbol{\theta}}_{\mathcal{M}}$ are distorted by the unmodeled term, which, in our extended representation, takes the form:

$$\hat{\boldsymbol{\theta}}_{\mathcal{M}} = \left(\mathbf{M}_{\mathcal{TM}}^+ \mathbf{M}_{\mathcal{TM}} \right) \boldsymbol{\theta}_{\mathcal{M}} + \left(\mathbf{M}_{\mathcal{TM}}^+ \mathbf{M}_{\mathcal{TU}} \right) \boldsymbol{\theta}_{\mathcal{U}}. \qquad (11)$$

For conceptual clarity, we write this as

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix} \hat{\boldsymbol{\theta}}_{\mathcal{M}} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{M}_{\mathcal{TM}}^+ \mathbf{M}_{\mathcal{TM}} & \mathbf{M}_{\mathcal{TM}}^+ \mathbf{M}_{\mathcal{TU}} \\ 0 & 0 \end{pmatrix} \boldsymbol{\theta}$$
$$= \begin{pmatrix} \mathbf{B} & \mathbf{A} \\ 0 & 0 \end{pmatrix} \boldsymbol{\theta}, \qquad (12)$$

where we have defined

$$\mathbf{A} = \mathbf{M}_{\mathcal{TM}}^+ \mathbf{M}_{\mathcal{TU}} \quad \text{and} \quad \mathbf{B} = \mathbf{M}_{\mathcal{TM}}^+ \mathbf{M}_{\mathcal{TM}}, \qquad (13)$$

and $\boldsymbol{\theta}$ is the vector of parameters that represents the complete signal precisely.

We call $\mathbf{A}$ the *generalized aliasing operator*. It quantifies how the effects of the unmodeled parameters $\boldsymbol{\theta}_{\mathcal{U}}$ are redirected (aliased) into the modeled parameters. Note

that $\mathbf{A}$ depends not only on the partition between modeled parameters and unmodeled modes, but also on the partition between training points and prediction points and the choice of pseudoinverse or the choice of learning algorithm, more generally.

In the familiar example of Fourier series, the concept of *aliasing* refers to the distortion of a low-frequency signal by high-frequency modes. Expressed in the form we have described, Fourier aliasing is found from $\mathbf{A} = \mathbf{M}_{\mathcal{TM}}^+ \mathbf{M}_{\mathcal{TU}}$, expressed in the Fourier basis for uniform samples (see Section III B for an example of aliasing in Fourier series), where it can be expressed in closed-form. Generalizing beyond the specific concept of frequency, $\mathbf{A}$ quantifies how unmodeled components affect the signal at the sampled points, leading to a misrepresentation of the inferred modeled parameters that we call *generalized aliasing*.

Using $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_{\mathcal{M}}, \mathbf{0})$, we reconstruct the inferred signal over both training and prediction points

$$\hat{\boldsymbol{y}} = \mathbf{M}\hat{\boldsymbol{\theta}} = \mathbf{M}\mathbf{M}_{\mathcal{TM}}^+ \boldsymbol{y}_{\mathcal{T}} = \mathbf{M}(\mathbf{B}\boldsymbol{\theta}_{\mathcal{M}} + \mathbf{A}\boldsymbol{\theta}_{\mathcal{U}}). \qquad (14)$$

Comparing $\hat{\boldsymbol{y}}$ with the true $\boldsymbol{y}$, the GAD decomposes the population risk of Eq. (5) into an intuitive partition. Assuming that the points t are drawn from a uniform distribution on Ω, the risk is

$$R_{\boldsymbol{\theta}}(\hat{\boldsymbol{y}}) = \mathbb{E}_t[(\boldsymbol{y}(t) - \hat{\boldsymbol{y}}(t))^2] = \sum_t (\boldsymbol{y}(t) - \hat{\boldsymbol{y}}(t))^2$$

$$= \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2 \qquad (15)$$

$$= \left\| \mathbf{M} \begin{pmatrix} \mathbf{I}_{\mathcal{M}} - \mathbf{B} & -\mathbf{A} \\ 0 & \mathbf{I}_{\mathcal{U}} \end{pmatrix} \boldsymbol{\theta} \right\|^2, \qquad (16)$$

where the norm $\| \cdot \|$ is the 2-norm $\| \cdot \|_2$. This motivates the definition of the *parameter error operator*

$$\mathbf{E}_{\boldsymbol{\theta}} = \begin{pmatrix} \mathbf{P}_{\mathcal{N}} & -\mathbf{A} \\ 0 & \mathbf{I}_{\mathcal{U}} \end{pmatrix}, \qquad (17)$$

where we define

$$\mathbf{P}_{\mathcal{N}} = \mathbf{I}_{\mathcal{M}} - \mathbf{B}.$$

We have used the subscript $\boldsymbol{\theta}$ on $\mathbf{E}_{\boldsymbol{\theta}}$ to indicate that $\mathbf{E}_{\boldsymbol{\theta}}\boldsymbol{\theta}$ represents errors in the inferred parameters; whereas the errors in the signal are $\boldsymbol{y} - \hat{\boldsymbol{y}} = \mathbf{M}\mathbf{E}_{\boldsymbol{\theta}}\boldsymbol{\theta}$. The notation $\mathbf{P}_{\mathcal{N}} = \mathbf{I}_{\mathcal{M}} - \mathbf{B}$ is motivated by the fact that it is the orthogonal projection onto the kernel $\mathcal{N}$ of $\mathbf{M}_{\mathcal{TM}}$ (see Proposition II.1, below).

We call the block decomposition in Eq. (17) the *generalized aliasing decomposition*, or *GAD*, for short. We call the three nonzero blocks of $\mathbf{E}_{\boldsymbol{\theta}}$ *data insufficiency* $\mathbf{P}_{\mathcal{N}}$, *model insufficiency* $\mathbf{I}_{\mathcal{U}}$, and *generalized aliasing* $-\mathbf{A}$. The effect of these operators acting on the parameters $\boldsymbol{\theta}$ are depicted in the left panel of Figure 2. We now analyze each of these contributions to the error in turn.

C.   Error Analysis

For a given partition of the parameter space between modeled and unmodeled subspaces $\Theta = \mathcal{M} \oplus \mathcal{U}$, the predictions $\hat{\boldsymbol{y}}$ and the risk $R_{\boldsymbol{\theta}}(\hat{\boldsymbol{y}})) = \|\hat{\boldsymbol{y}} - \boldsymbol{y}\|^2$ depend on the choice $T = \{t_1, \ldots, t_n\}$ of the training points. The expected value of the risk (taken over the distribution of T) is often decomposed into a sum of a *bias* term and a *variance* term; see, for example, [29, §20.1]. In many settings, however, it is more natural to analyze the risk $\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2$ and its expected value $\mathbb{E}_T\big[\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2\big]$ directly through the GAD.

In this Section we are primarily interested in estimating the risk for a particular decomposition. Observe that the risk is bounded by the (square of the) product of three norms

$$R_{\boldsymbol{\theta}}(\hat{\boldsymbol{y}}) = \|\mathbf{M}\mathbf{E}_{\boldsymbol{\theta}}\boldsymbol{\theta}\|^2 \leq \|\mathbf{M}\|^2 \|\mathbf{E}_{\boldsymbol{\theta}}\|^2 \|\boldsymbol{\theta}\|^2.$$

The norm used on the linear transformations $\mathbf{M}$ and $\mathbf{E}_{\boldsymbol{\theta}}$ is the *induced* norm, defined as

$$\|\mathbf{M}\| = \max_{\|\boldsymbol{\nu}\|=1} \|\mathbf{M}\boldsymbol{\nu}\| \qquad \text{and} \qquad \|\mathbf{E}\| = \max_{\|\boldsymbol{\nu}\|=1} \|\mathbf{E}\boldsymbol{\nu}\|.$$

In the finite-dimensional case (where the transformations are represented by matrices) it is well known that when the norm on both the domain and range of a matrix is the usual (two-) norm, then the induced norm of a matrix is its largest singular value. In this case the induced norm is often called the *spectral norm*.

We assume the transformation $\mathbf{M}$ has a bounded norm and note that its norm is independent of the choice of model $\mathcal{M}$ and of the choice T of training points. The risk certainly depends on $\boldsymbol{\theta}$ and its norm, particularly in the two extremes of low and high model complexity (left and right end, respectively, of the plots shown here). In the intermediate regime however the risk is generally dominated by the aliasing component of the GAD which is mostly independent of the $\boldsymbol{\theta}$ themselves.

Since we are particularly interested in how these contributions depend on the number of model parameters, consider the situation where the columns of $\mathbf{M}$ are fixed, and a given decomposition $\Theta = \mathcal{M} \oplus \mathcal{U}$ of parameter space is changed by moving one basis element Φ out of the space $\mathcal{U}$ and into the space $\mathcal{M}$. This corresponds to moving the corresponding column $\boldsymbol{\varphi}$ out of the matrix $\mathbf{M}_{\mathcal{T}\mathcal{U}}$ and into the design matrix $\mathbf{M}_{\mathcal{T}\mathcal{M}}$. Fixing the order of the columns of $\mathbf{M}$, and letting m denote the dimension of the modeled space $\mathcal{M}$, we introduce the notation $\mathbf{M}_{\mathcal{T}\mathcal{M}}(m)$ and $\mathbf{M}_{\mathcal{T}\mathcal{U}}(m)$ to make explicit the dependence of the block operators on the dimension m. That is, $\mathbf{M}_{\mathcal{T}\mathcal{M}}(m)$ denotes the design matrix in the case that the first m columns of $\mathbf{M}$ are assigned to $\mathbf{M}_{\mathcal{T}\mathcal{M}}$ and the remaining columns are assigned to $\mathbf{M}_{\mathcal{T}\mathcal{U}}$. With this convention, we now analyze each of the elements of the GAD in turn.

1.   Data Insufficiency $\mathbf{P}_{\mathcal{N}}$

Data insufficiency refers to the upper left block $\mathbf{P}_{\mathcal{N}}$ of Eq. (17) and its effect on the parameters $\boldsymbol{\theta}$. In the case that $\mathbf{M}_{\mathcal{T}\mathcal{M}}^+$ is the Moore–Penrose pseudo inverse of $\mathbf{M}_{\mathcal{T}\mathcal{M}}$, the operator $\mathbf{P}_{\mathcal{N}}$ is the orthogonal projection of $\mathcal{M}$ onto the null space $\mathcal{N} \subseteq \mathcal{M}$ of $\mathbf{M}_{\mathcal{T}\mathcal{M}}$, as the following proposition shows.

Proposition II.1. *If* $\mathbf{M}_{\mathcal{T}\mathcal{M}}^+$ *is the Moore–Penrose pseudoinverse of* $\mathbf{M}_{\mathcal{T}\mathcal{M}}$, *then the data insufficiency operator* $\mathbf{P}_{\mathcal{N}} = \mathbf{I}_{\mathcal{M}} - \mathbf{B}$ *is the orthogonal projection of* $\mathcal{M}$ *to the kernel* $\mathcal{N}$ *of* $\mathbf{M}_{\mathcal{T}\mathcal{M}}$.

Proof. Let

$$\mathbf{M}_{\mathcal{T}\mathcal{M}} = [U_1|U_2] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^{\mathsf{T}} \\ V_2^{\mathsf{T}} \end{bmatrix}$$

be the full SVD of $\mathbf{M}_{\mathcal{T}\mathcal{M}}$, where $\mathbf{M}_{\mathcal{T}\mathcal{M}}$ has rank r, the matrix Σ_1 is invertible of shape $r \times r$, and $V = [V_1|V_2]$ is an orthogonal matrix

$$V_1 V_1^{\mathsf{T}} + V_2 V_2^{\mathsf{T}} = V V^{\mathsf{T}} = \mathbf{I}_{\mathcal{M}},$$

with V_1 having r columns and the columns of V_2 spanning the kernel $\mathcal{N}$ of $\mathbf{M}_{\mathcal{T}\mathcal{M}}$. The Moore–Penrose pseudoinverse of $\mathbf{M}_{\mathcal{T}\mathcal{M}}$ can be written as $\mathbf{M}_{\mathcal{T}\mathcal{M}}^+ = V_1 \Sigma_1^{-1} U_1^{\mathsf{T}}$, which gives

$$\begin{aligned} \mathbf{I}_{\mathcal{M}} - \mathbf{B} &= V V^{\mathsf{T}} - \mathbf{M}_{\mathcal{T}\mathcal{M}}^+ \mathbf{M}_{\mathcal{T}\mathcal{M}} \\ &= (V_1 V_1^{\mathsf{T}} + V_2 V_2^{\mathsf{T}}) - V_1 \Sigma_1^{-1} \Sigma_1 V_1^{\mathsf{T}} \\ &= V_2 V_2^{\mathsf{T}}. \end{aligned}$$

But since the columns of V_2 span the kernel $\mathcal{N}$ of $\mathbf{M}_{\mathcal{T}\mathcal{M}}$, the matrix $V_2 V_2^{\mathsf{T}}$ is exactly the orthogonal projection of $\mathcal{M}$ onto $\mathcal{N}$. $\qquad \square$

The induced norm of any projection operator is always either 0, if it's the zero operator, or 1 otherwise, which yields the following corollary.

Corollary II.2. *The induced norm* $\|\mathbf{P}_{\mathcal{N}}\|$ *of the operator* $\mathbf{P}_{\mathcal{N}}$ *is bounded above by 1, and is always equal to 1 except when* $\mathbf{M}_{\mathcal{T}\mathcal{M}}$ *is injective (full column rank), in which case* $\mathbf{P}_{\mathcal{N}} = 0$ *is the zero operator.*

The corollary shows that when there are enough training data so that $\mathbf{M}_{\mathcal{T}\mathcal{M}}$ is of full column rank, then the norm of the data insufficiency operator $\mathbf{P}_{\mathcal{N}}$ is zero. Generically this happens when there are more data points (rows) than basis functions (columns) in $\mathbf{M}_{\mathcal{T}\mathcal{M}}$, as depicted in the left panel of Figure 2.

Let $\mathcal{N}(m)$ denote the null space of $\mathbf{M}_{\mathcal{T}\mathcal{M}}(m)$. Since $\mathcal{N}(m) \subseteq \mathcal{N}(m+1)$, it follows that the error contribution $\|\mathbf{P}_{\mathcal{N}}\boldsymbol{\theta}_{\mathcal{M}}\|$ from data insufficiency is a nondecreasing function of m. This is depicted on the bottom right of the left panel of Figure 2.

2. Model Insufficiency $\mathbf{I}_{\mathcal{U}}$

Model insufficiency refers to the lower right block $\mathbf{I}_{\mathcal{U}}$ of Eq. (17). It is the most straightforward of the three parts of the GAD to analyze, as it is simply the identity operator on the unmodeled parameters $\mathcal{U}$. Except in the trivial and uninteresting case that $\dim(\mathcal{U}) = 0$, its operator norm is always 1. The contribution to the parameter error from model insufficiency is simply the square $\|\boldsymbol{\theta}_{\mathcal{U}}\|^2$ of the norm of the nescient parameters. It follows that model insufficiency is a non-increasing function of m since it decreases by exactly $|\theta_{m+1}|^2$ as the coordinate θ_{m+1} is removed from the unmodeled space $\mathcal{U}$ and adjoined to the modeled space $\mathcal{M}$.

Model insufficiency dominates when the dimension m of the model $\mathcal{M}$ is small, reflecting the fact that most of the signal is unknown and the model lacks the capacity to capture the signal faithfully (see the bottom left part of the left panel of Fig. 2).

3. Generalized Aliasing $\mathbf{A}$: Overview

Finally, we consider contributions from the *generalized aliasing operator* $\mathbf{A}$. This is the most complicated contribution, and is the source of non-trivial generalization curves such as double or multiple descent, or multiple risk peaks from structured data. This term tends to dominate the risk in the intermediate regime of most models. Importantly, in many cases the effects of $\mathbf{A}$ can be analyzed without knowing $\boldsymbol{\theta}$ or the labels $\boldsymbol{y}$.

Recall from Eq. (13) that the aliasing operator is the product of the pseudoinverse design matrix $\mathbf{M}_{\mathcal{TM}}^{+}$ and the transformation $\mathbf{M}_{\mathcal{TU}}$. Increasing the number of model parameters by moving one column $\boldsymbol{\varphi}$ out of $\mathbf{M}_{\mathcal{TU}}$ and into the design matrix never increases the norm $\|\mathbf{M}_{\mathcal{TU}}\|$, but its effect on $\|\mathbf{M}_{\mathcal{TM}}^{+}\|$ is determined primarily by whether $\boldsymbol{\varphi}$ is linearly independent of the other columns of $\mathbf{M}_{\mathcal{TM}}$ or not, as described in the following theorem (proved in Section II C 4).

Theorem II.3. *When changing the model by moving one column $\boldsymbol{\varphi}$ out of $\mathbf{M}_{\mathcal{TU}}$ and into the design matrix, the norm $\|\mathbf{M}_{\mathcal{TU}}\|$ never increases and*

- *$\|\mathbf{M}_{\mathcal{TM}}^{+}\|$ cannot decrease if $\boldsymbol{\varphi}$ is linearly **independent** of the other columns of $\mathbf{M}_{\mathcal{TM}}$,*

- *$\|\mathbf{M}_{\mathcal{TM}}^{+}\|$ cannot increase if $\boldsymbol{\varphi}$ is linearly **dependent** upon the other columns of $\mathbf{M}_{\mathcal{TM}}$.*

Moreover, as the model dimension m increases to ∞, the norm $\|\mathbf{M}_{\mathcal{TM}}^{+}\|$ shrinks to 0, almost surely.

Although it can be arranged so that $\|\mathbf{M}_{\mathcal{TM}}^{+}\|$ remains constant when moving one column $\boldsymbol{\varphi}$ fom $\mathbf{M}_{\mathcal{TU}}$ to $\mathbf{M}_{\mathcal{TM}}$, in most cases we see a significant increase in $\|\mathbf{M}_{\mathcal{TM}}^{+}\|$ whenever $\boldsymbol{\varphi}$ is independent from the previous columns of $\mathbf{M}_{\mathcal{TM}}$ and a significant decrease in $\|\mathbf{M}_{\mathcal{TM}}^{+}\|$ whenever $\boldsymbol{\varphi}$ is dependent upon the previous columns of $\mathbf{M}_{\mathcal{TM}}$.

Theorem II.3 fully explains the sharp peaks in generalization curves described as double and multiple descent, and it is relevant to other nonmonotonic features in both the under- and over-parameterized regimes, as we now describe.

For a generic $\mathbf{M}$ the columns are typically arranged so that for $m < n$ each column $\boldsymbol{\varphi}_{m+1}$ is independent of the previous columns of $\mathbf{M}_{\mathcal{TM}}(m)$ and each column of $\mathbf{M}_{\mathcal{TU}}$ does not have a large impact on the norm of $\mathbf{M}_{\mathcal{TU}}$. Hence, as m increases the norm $\|\mathbf{M}_{\mathcal{TM}}^{+}(m)\|$ is expected to grow nearly monotonically until the interpolation threshold $m = n$. Once $m \geq n$ the columns of $\mathbf{M}_{\mathcal{TM}}(m)$ are expected to span the column space of the entire training set $(\mathbf{M}_{\mathcal{TM}}|\mathbf{M}_{\mathcal{TU}})$ of the operator $\mathbf{M}$, so each new column added to $\mathbf{M}_{\mathcal{TM}}(m)$ will be linearly dependent on the existing columns, and hence the norms $\|\mathbf{M}_{\mathcal{TM}}^{+}(m)\|$ and $\|\mathbf{A}\|$ cannot increase and typically decrease. In this generic case, $\|\mathbf{M}_{\mathcal{TM}}^{+}(m)\|$ is a nondecreasing function of m until $m = n$, after which it is nonincreasing. The common peak in the generalization error at the interpolation threshold is thus understood as the peak in $\|\mathbf{M}_{\mathcal{TM}}^{+}(m)\|$ at $m = n$.

More complicated generalization curves can be understood by considering whether the next basis vector $\boldsymbol{\varphi}_{m+1}$ is either linearly dependent (the norm $\|\mathbf{M}_{\mathcal{TM}}^{+}(m+1)\| \leq \|\mathbf{M}_{\mathcal{TM}}^{+}(m)\|$) or linearly independent (the norm $\|\mathbf{M}_{\mathcal{TM}}^{+}(m+1)\| \geq \|\mathbf{M}_{\mathcal{TM}}^{+}(m)\|$) on the previously modeled terms, i.e., all those columns already contained in $\mathbf{M}_{\mathcal{TM}}(m)$. Regardless of the ordering of the columns of $\mathbf{M}$, the upper bound

$$\|\mathbf{A}\| \leq \|\mathbf{M}_{\mathcal{TM}}^{+}\|\|\mathbf{M}_{\mathcal{TU}}\|$$

cannot increase when stepping from m to $m + 1$ unless the next column $\boldsymbol{\varphi}_{m+1}$ is independent of the previous columns. Moreover, this upper bound will almost surely shrink to 0 as $m \to \infty$ as both $\|\mathbf{M}_{\mathcal{TU}}\| \to 0$ and $\|\mathbf{M}_{\mathcal{TM}}^{+}\| \to 0$.

Of course, one can arrange to add columns to $\mathbf{M}_{\mathcal{TM}}(m)$ in a way that the rank of $\mathbf{M}_{\mathcal{TM}}(m)$ grows slower than expected, permitting the construction of descent curves for $\|\mathbf{A}\|$ of various shapes. But when the columns are sufficiently general (as, for example, with the random ReLU features (RRF) model and the random Fourier features (RFF) model), the result for $\|\mathbf{A}\|$ is similar in shape to the standard double descent curve for mean-squared error, described in [8] with a single large peak at the interpolation threshold and decreasing monotonically thereafter (see Figure 3).

4. Generalized Aliasing $\mathbf{A}$: Mathematical Treatment

In this section we give more mathematical details of the norm $\|\mathbf{A}\|$ of the aliasing operator and a proof of Theorem II.3.

a. Tools for Analyzing Norms The main tool we use is the following theorem, whose earliest statement seems to be [30, Theorem 17] (see also [31–33]).

Theorem II.4. *Let Φ be an $n \times n$ Hermitian matrix with eigenvalues $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n$ and let C be a positive semidefinite matrix of rank 1. The eigenvalues $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_n$ of the matrix $\Xi = \Phi + C$ satisfy*

$$\beta_1 \geq \alpha_1 \geq \beta_2 \geq \alpha_2 \geq \cdots \geq \beta_n \geq \alpha_n.$$

This theorem immediately gives the corollary that, under the same assumptions on Φ and C, the eigenvalues $\delta_1 \geq \cdots \geq \delta_n$ of $D = \Phi - C$ are below the corresponding eigenvalues of Φ and are interleaved according to:

$$\alpha_1 \geq \delta_1 \geq \alpha_2 \geq \cdots \geq \alpha_n \geq \delta_n.$$

Theorem II.4 also leads to the following fundamental result for analyzing the operator norm of the aliasing operator $\mathbf{A}$, stated here in a more general form.

Theorem II.5. *Let $\mathbf{X}$ be an $n \times m$ matrix of rank r with smallest singluar value $\sigma_r > 0$. Let $\widetilde{\mathbf{X}} = [\mathbf{X}|\boldsymbol{\varphi}]$ be the $n \times (m+1)$ matrix obtained by adjoining an n-dimensional column vector $\boldsymbol{\varphi}$ to $\mathbf{X}$. The smallest singular value $\widetilde{\sigma}_{min} > 0$ of $\widetilde{\mathbf{X}}$ satisfies the following relations:*

$$0 < \widetilde{\sigma}_{min} \leq \sigma_r \qquad \text{if } \operatorname{rank}(\mathbf{X}) < \operatorname{rank}(\widetilde{\mathbf{X}}),$$
$$0 < \sigma_r \leq \widetilde{\sigma}_{min} \qquad \text{if } \operatorname{rank}(\mathbf{X}) = \operatorname{rank}(\widetilde{\mathbf{X}}).$$

Proof. Both $\mathbf{X}\mathbf{X}^\mathsf{T}$ and $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\mathsf{T}$ are $n \times n$ positive definite Hermitian matrices. The singular value decomposition of $\mathbf{X}$ shows that the singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$ and the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_m$ of $\mathbf{X}\mathbf{X}^\mathsf{T}$ satisfy

$$\lambda_1 = \sigma_1^2 \geq \lambda_2 = \sigma_1^2 \geq \cdots \geq \lambda_r = \sigma_r^2 > 0 = \lambda_{r+1}.$$

Similarly, the singular values $\widetilde{\sigma}_1 \geq \widetilde{\sigma}_2 \geq \cdots$ and eigenvalues $\widetilde{\lambda}_1 \geq \cdots \geq \widetilde{\lambda}_m$ of $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\mathsf{T}$ satisfy

$$\widetilde{\lambda}_1 = \widetilde{\sigma}_1^2 \geq \widetilde{\lambda}_2 = \widetilde{\sigma}_2^2 \cdots \geq \widetilde{\lambda}_r = \widetilde{\sigma}_r^2 \geq \widetilde{\lambda}_{r+1} \geq \cdots,$$

where $\widetilde{\lambda}_{r+1} = 0$ if $\operatorname{rank}(\widetilde{\mathbf{X}}) = r$, but $\widetilde{\lambda}_{r+1} > 0$ if $\operatorname{rank}(\widetilde{\mathbf{X}}) = r + 1$. Expanding $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\mathsf{T}$ gives $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\mathsf{T} = \mathbf{X}\mathbf{X}^\mathsf{T} + \boldsymbol{\varphi}\boldsymbol{\varphi}^\mathsf{T}$, where $\boldsymbol{\varphi}\boldsymbol{\varphi}^\mathsf{T}$ is positive semidefinite, so Theorem II.4 implies that

$$\widetilde{\lambda}_1 \geq \lambda_1 \cdots \geq \lambda_{r-1} \geq \widetilde{\lambda}_r \geq \lambda_r \geq \widetilde{\lambda}_{r+1} \geq 0.$$

If $\operatorname{rank}(\widetilde{\mathbf{X}}) = r + 1$ (that is, $\boldsymbol{\varphi}$ is not in the column space of $\mathbf{X}$), then $\lambda_r = \sigma_r^2 \geq \widetilde{\lambda}_{r+1} = \widetilde{\sigma}_{r+1}^2 > 0$. Taking square roots gives $\sigma_r > \widetilde{\sigma}_{r+1} = \widetilde{\sigma}_{min} > 0$.

If $\operatorname{rank}(\widetilde{\mathbf{X}}) = r$ (that is, $\boldsymbol{\varphi}$ is in the column space of $\mathbf{X}$), then the smallest nonzero eigenvalue of $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\mathsf{T}$ is $\widetilde{\lambda}_r$, which satisfies $\lambda_{r-1} \geq \widetilde{\lambda}_r \geq \lambda_r > 0$. Taking square roots gives $\widetilde{\sigma}_{min} = \widetilde{\sigma}_r \geq \sigma_r > 0$, as required. $\square$

b. Decomposing Generalized Aliasing We are interested in how the (induced) operator norm

$$\|\mathbf{A}\| = \|\mathbf{M}_{\mathcal{TM}}^+ \mathbf{M}_{\mathcal{TU}}\| \leq \|\mathbf{M}_{\mathcal{TM}}^+\| \|\mathbf{M}_{\mathcal{TU}}\|$$

changes as the model grows, that is, as a new column is removed from $\mathbf{M}_{\mathcal{TU}}$ and added to $\mathbf{M}_{\mathcal{TM}}$, but the training set (which rows are included) remains unchanged.

As before, assume $\mathbf{M}$ is fixed, and $\mathbf{M}_{\mathcal{TM}}(m)$ corresponds to the design matrix block when the model consists of the first m columns of $\mathbf{M}$, and $\mathbf{M}_{\mathcal{TU}}(m)$ is the corresponding unmodeled block. The matrix $\mathbf{M}_{\mathcal{TM}}(m+1)$ is constructed by moving one column $\boldsymbol{\varphi}_{m+1}$ from the nescience block into the design block. The training-set (top) part of the operator $\mathbf{M}$ decomposes as $[\mathbf{M}_{\mathcal{TM}}(m) \;\; \boldsymbol{\varphi}_{m+1} \;\; \mathbf{M}_{\mathcal{TU}}(m+1)]$.

c. Norm of $\mathbf{M}_{\mathcal{TU}}$ First consider what happens to the matrix $\mathbf{M}_{\mathcal{TU}}$ when a column $\boldsymbol{\varphi}_{m+1}$ is removed from $\mathbf{M}_{\mathcal{TU}}(m) = [\boldsymbol{\varphi}_m \;\; \mathbf{M}_{\mathcal{TU}}(m+1)]$. Expanding the product $\mathbf{M}_{\mathcal{TU}}(m)\mathbf{M}_{\mathcal{TU}}(m)^\mathsf{T}$ gives $\mathbf{M}_{\mathcal{TU}}(m)\mathbf{M}_{\mathcal{TU}}(m)^\mathsf{T} = \boldsymbol{\varphi}_{m+1}\boldsymbol{\varphi}_{m+1}^\mathsf{T} + \mathbf{M}_{\mathcal{TU}}(m+1)\mathbf{M}_{\mathcal{TU}}(m+1)^\mathsf{T}$. Since $\boldsymbol{\varphi}_{m+1}\boldsymbol{\varphi}_{m+1}^\mathsf{T}$ is positive semidefinite, Theorem II.4 applies and guarantees that the norms satisfy $\|\mathbf{M}_{\mathcal{TU}}(m)\| \geq \|\mathbf{M}_{\mathcal{TU}}(m+1)\|$, and thus the norm $\|\mathbf{M}_{\mathcal{TU}}(m)\|$ is a non-increasing function of m.

d. Pseudoinverse of Design: Consider now the pseudoinverse term $\|\mathbf{M}_{\mathcal{TM}}^+\|$ when $\boldsymbol{\varphi}_{m+1}$ is adjoined to $\mathbf{M}_{\mathcal{TM}}(m)$ to create $\mathbf{M}_{\mathcal{TM}}(m+1)$. Theorem II.5 guarantees that whenever $\boldsymbol{\varphi}_{m+1}$ is linearly independent of the old model (does not lie in the column space of $\mathbf{M}_{\mathcal{TM}}(m)$), then the induced norm of the new pseudoinverse is bounded below by the induced norm of the old pseudoinverse:

$$\|\mathbf{M}_{\mathcal{TM}}(m+1)^+\| \geq \|\mathbf{M}_{\mathcal{TM}}(m)^+\|.$$

Similarly, when $\boldsymbol{\varphi}_{m+1}$ is linearly dependent on the old model, then the induced norm of the new pseudoinverse is bounded above by the norm of the previous pseudoinverse

$$\|\mathbf{M}_{\mathcal{TM}}(m+1)^+\| \leq \|\mathbf{M}_{\mathcal{TM}}(m)^+\|.$$

This proves Theorem II.3.

e. Limiting behavior of $\mathbf{A}$ As the number m of model parameters gets large, the norm $\|\mathbf{A}\|$ is dominated by the norm of the pseudoinverse $\|\mathbf{M}_{\mathcal{TM}}^+\|$. For purposes of this analysis, assume that the columns of the training-set (top) part of $\mathbf{M}$ are independent identically distributed (i.i.d.) random vectors $\boldsymbol{\varphi}_i \in \mathbb{R}^n$ with finite second moment $\mathbb{E}[\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^\mathsf{T}] = \Sigma$, where Σ is of full rank (rank t).

The Strong Law of Large Numbers guarantees that

$$\frac{1}{m}\mathbf{M}_{\mathcal{TM}}(m)\mathbf{M}_{\mathcal{TM}}(m)^\mathsf{T}$$
$$= \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^\mathsf{T} \xrightarrow{a.s.} \mathbb{E}[\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^\mathsf{T}]$$
$$= \Sigma$$

as $m \to \infty$. This implies that the smallest eigenvalue of $\frac{1}{m}\mathbf{M}_{\mathcal{TM}}(m)\mathbf{M}_{\mathcal{TM}}(m)^\mathsf{T}$ converges almost surely to the smallest eigenvalue $\lambda_{min} > 0$ of Σ. Thus the smallest eigenvalue of $\mathbf{M}_{\mathcal{TM}}(m)\mathbf{M}_{\mathcal{TM}}(m)^\mathsf{T}$ approches

$m\lambda_{min}$ and goes to infinity almost surely as $m \to \infty$. Thus the smallest singular value $\sigma_{min}(m) = \sqrt{\lambda_{min}(m)}$ of $\mathbf{M}_{\mathcal{TM}}(m)$ also goes to infinity, and this implies $\|\mathbf{M}_{\mathcal{TM}}(m)^+\| = \frac{1}{\sigma_{min}} \xrightarrow{a.s.} 0$.

Because $\|\mathbf{M}_{\mathcal{TU}}(m)\|$ is bounded above and decreasing in m, we have

$$\|\mathbf{A}(m)\| = \|\mathbf{M}_{\mathcal{TM}}(m)^+ \mathbf{M}_{\mathcal{TU}}(m)\|$$
$$\leq \|\mathbf{M}_{\mathcal{TM}}(m)^+\| \|\mathbf{M}_{\mathcal{TU}}(m)\| \xrightarrow{a.s.} 0.$$

In the special case that the columns $\boldsymbol{\varphi}_i$ are i.i.d. standard normal and $m > n$, it is known [34, Thm 2.6] that $\mathbb{E}[\sigma_{min}(m)] \geq \sqrt{m} - \sqrt{n}$, so $\|\mathbf{M}_{\mathcal{TM}}(m)^+\|$ and $\|\mathbf{A}(m)\|$ are $O(m^{-1/2})$ or smaller.

5. *Model–Data Trade-off*

Because the data and model insufficiency terms are non-decreasing and non-increasing respectively, there is an inherent trade-off between them. To study this trade-off, we introduce the combined model and data insufficiency error:

$$\|\mathbf{E}_I \boldsymbol{\theta}\|^2 = \|\mathbf{P}_{\mathcal{N}(m)} \boldsymbol{\theta}_{\mathcal{M}(m)}\|^2 + \|\mathbf{I}_{\mathcal{U}(m)} \boldsymbol{\theta}_{\mathcal{U}(m)}\|^2. \quad (18)$$

where we have made the m-dependence explicit. To make statements about the dependence of the combined insufficiency errors, we consider two prior distributions for the distribution of the components of $\boldsymbol{\theta}$.

We first consider the *random feature model* in which components of $\boldsymbol{\theta}$ are i.i.d. random variables with mean zero and variance σ^2. (Note that the random feature model is sensible only when the parameter space has finite dimension, otherwise the norm $\|\boldsymbol{\theta}\|$ would be infinite.) In this setting the expected total insufficiency error is

$$\mathbb{E}_{\theta} \left[\|\mathbf{E}_I \boldsymbol{\theta}\|^2 \right] = \sigma^2 \left(\text{Tr} \, \mathbf{P}_{\mathcal{N}} + \text{Tr} \, \mathbf{I}_{\mathcal{U}} \right) \quad (19)$$
$$= \sigma^2 (\dim \mathcal{N} + \dim \mathcal{U}). \quad (20)$$

At each step $\dim(\mathcal{U})$ decreases by one, while $\dim(\mathcal{N})$ increases by either zero or one; so, for the random feature model, the expected total insufficiency error is a strictly nonincreasing function of m.

In many ways the random feature model is unrealistic for scientific and engineering applications. Modelers often have prior information about which parameters are most important and preferentially order the parameter vector to reflect this. In such cases and for very small m, as m increases there is often an initial descent of model insufficiency due to the model's rapidly increasing ability to capture the signal faithfully. This is conceptually analogous to reducing bias in the classical bias–variance paradigm. However, for very large models, the data insufficiency grows faster than the decrease in model insufficiency. This growing error for large models is not analogous to variance and cannot be termed over-fitting.

Rather, it reflects the lack of invertibility for large models, specifically, larger parameter bias as more of the mass of $\boldsymbol{\theta}$ is projected into the kernel $\mathcal{N}$ of the design matrix $\mathbf{M}_{\mathcal{TM}}$.

This phenomenon of growing data insufficiency could be thought of as a form of *over-modeling*. It occurs when parameters that are expected to contribute minimally to the signal are included in the model. To be accurately inferred, such parameters place stringent informativity requirements for the data, amplifying the effects of data insufficiency. This growing data insufficiency is less of a problem in random feature models, because all parameters contribute essentially equally, which is why, as shown in the discussion above about random feature models and in the examples in Section III D below illustrate that random feature models tend to have optimal performance in the asymptotic limit as $m \to \infty$. But models that exploit prior information, so that the ordering of the basis functions and/or the choice of the training points are physically motivated, are more likely to suffer from increasing data insufficiency as the number of parameters grows. Thus these models generally have their ideal risk occur in the classical regime.

III. DEMONSTRATIONS AND APPLICATIONS

A. Random Feature Models

Although the motivating example in Section II A for the generalized aliasing decomposition (GAD) was focused on one-dimensional polynomials, the GAD applies much more generally to the problem of fitting a function $f : \Omega \to \mathbb{R}$ or $f : \Omega \to \mathbb{C}$ for a general set Ω. We illustrate this with examples of two different choices of models applied to three different data sets. The two bases are random Fourier features (RFF) and random ReLU features (RRF) (described in [8] and [35]).

All of these basis functions are of the form $\phi_k(\boldsymbol{t}) = \sigma(\langle \boldsymbol{t}, \boldsymbol{v}_k \rangle)$, where the $\boldsymbol{v}_k \in \mathbb{R}^d$ are i.i.d. normal, and σ is some activation function. In the case of the RRF model, the activation function is the usual ReLU, and in the case of RFF the activation function is $\sigma(x) = \exp(i\pi x)$. The models that result from using these two choices (either RRF or RFF) can both be thought of as 2-layer neural networks of the form

$$y(\boldsymbol{t}) = \sum_{k=1}^{m} \theta_k \phi_k(\boldsymbol{t}) = \sum_{k=1}^{m} \theta_k \sigma(\langle \boldsymbol{t}, \boldsymbol{v}_k \rangle).$$

The data sets we use here are images from MNIST and CIFAR-10 and points from the sphere $\mathbb{S}^{d-1}(\sqrt{d})$, as in Mei–Montanari [35]; we have arbitrarily fixed $d = 1024$ for this sphere. In each case 1,000 training points $\boldsymbol{t}_i$ were drawn uniformly and evaluated at 6,000 basis functions (either RRF or RFF). The columns of the resulting design matrix $\mathbf{M}_{\mathcal{TM}}$ and unmodeled block $\mathbf{M}_{\mathcal{TU}}$ are all of the form $\boldsymbol{\varphi}_k = (\phi_k(\boldsymbol{t}_1), \dots, \phi_k(\boldsymbol{t}_n))$.
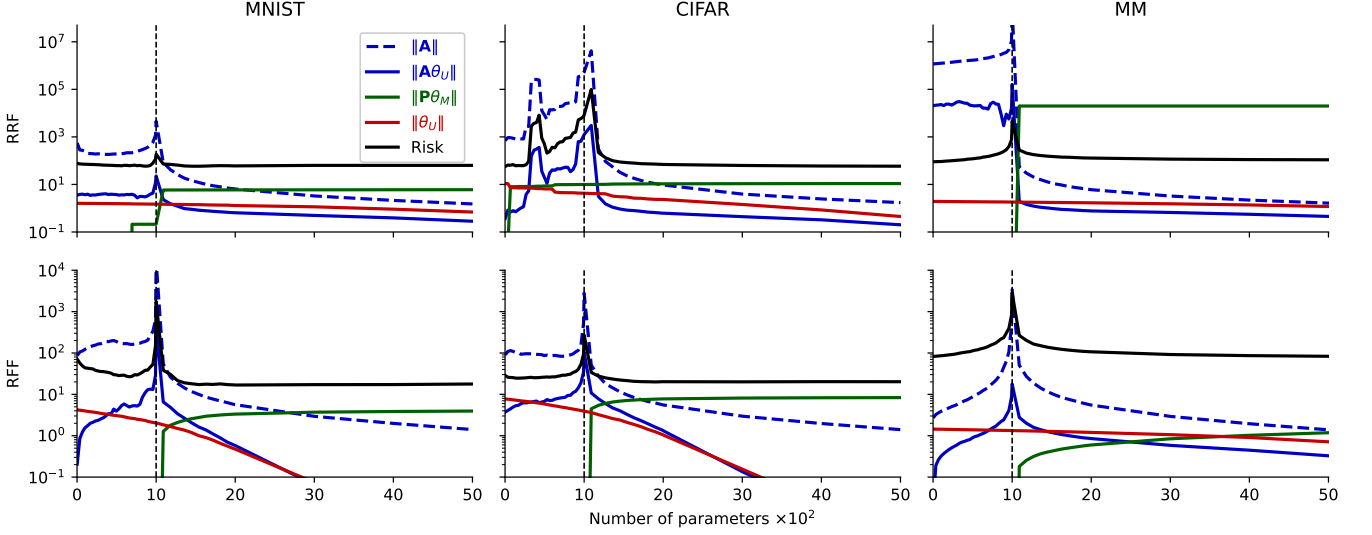
FIG. 3. The induced (spectral) norm of the aliasing operator $\mathbf{A}$ (dashed blue), the aliased parameter error $\|\mathbf{A}\boldsymbol{\theta}_{\mathcal{U}}\|$ (solid blue), the data insufficiency parameter error $\|\mathbf{P}_{\mathcal{N}}\boldsymbol{\theta}_{\mathcal{M}}\|$, the model insufficiency parameter error $\|\mathbf{I}_{\mathcal{U}}\boldsymbol{\theta}_{\mathcal{U}}\|$, and the risk (black) for the random ReLU features (RRF) model (top row) and the random Fourier features (RFF) model (bottom row) on the MNIST and CIFAR-10 datasets, as in [8], as well as on the Mei-Montanari (MM) sphere [35]. In each case the models were trained on 1,000 randomly chosen training points (vertical dashed black line) with the number of modeled parameters ranging from 1 up to 5,000. Although the aliasing operator $\|\mathbf{A}\|$ and aliasing parameter error $\|\mathbf{A}\boldsymbol{\theta}_{\mathcal{U}}\|$ both go to zero almost surely as the number of parameters goes to ∞, the full parameter error has contributions from $\mathbf{I}_{\mathcal{U}}\boldsymbol{\theta}_{\mathcal{U}}$ (also decreasing, but slowly) and the data insufficiency parameter error $\|\mathbf{P}_{\mathcal{N}}\boldsymbol{\theta}_{\mathcal{M}}\|$, which, though bounded above by $\|\boldsymbol{\theta}\|$, is nondecreasing. Data insufficiency $\|\mathbf{P}_{\mathcal{N}}\boldsymbol{\theta}_{\mathcal{U}}\|$ is generally 0 until the interpolation threshold, but in the top center panel (and to a lesser degree in the top left panel) it is nonzero before the interpolation threshold, indicating that the design matrix $\mathbf{M}_{\mathcal{TM}}$ fails to be full rank fairly early. Presumably this happens because ReLU is vanishes for many inputs. The early large decrease in $\|\mathbf{A}\|$ in that top center panel could be partly due to adding linearly dependent columns to $\mathbf{M}_{\mathcal{TM}}$ or to the (usually much less significant) decrease in $\|\mathbf{M}_{\mathcal{TU}}\|$ as columns are moved from $\mathbf{M}_{\mathcal{TU}}$ into $\mathbf{M}_{\mathcal{TM}}$.

In Figure 3 we plot the norm of the aliasing matrix $\mathbf{A}$ and the parameter error contribution from each of: data insufficiency, model insufficiency, and generalized aliasing. The risk is also plotted and each of these terms are displayed as functions of the number m of parameters for these models on the three different datasets. Recall that operator norms $\|\mathbf{P}_{\mathcal{N}}\|$ and $\|\mathbf{I}_{\mathcal{U}}\|$ are always 1 or 0, so we instead plot the norms of the products $\|\mathbf{P}_{\mathcal{N}}\boldsymbol{\theta}_{\mathcal{M}}\|$ and $\|\mathbf{I}_{\mathcal{U}}\boldsymbol{\theta}_{\mathcal{U}}\|$, which are the contributions to parameter error due to data insufficiency and model insufficiency, respectively. We also plot $\|\mathbf{A}\boldsymbol{\theta}_{\mathcal{M}}\|$, to show the effect of each part of the GAD on the parameter error.

The GAD decomposition in these examples closely matches the canonical picture presented in Figure 2, illustrating the dominant effect that the aliasing operator has on the non-monotonic behavior of the full risk. This generic behavior is because the random selection of additional features (columns in $\mathbf{M}_{\mathcal{TM}}$) almost surely guarantees that such new features are linearly independent (on the sample points) from the existing features up to the interpolation threshold so that the risk will increase with model complexity. After the interpolation threshold, additionally added features will be linearly dependent on the existing modeled features, and the aliasing

(and hence risk) will decrease with model complexity.

B. Why call it 'aliasing'? Discrete Fourier series

To clarify the name "generalized aliasing," we turn to an example familiar in the signals-processing community, the Fourier decomposition, which we describe here.

For a square-integrable function on the interval $[0, T]$ we will assume that our training data comes from equally spaced points $0 = t_0 < \cdots < t_n = T$. We let $\omega_n = \exp(2\pi\imath/n)$, be a primitive n-th root of unity and introduce the Fourier basis vectors $\boldsymbol{w}_n^{(k)} = (\omega_n^0, \omega_n^k, \ldots, \omega_n^{(n-1)k})$. The discrete Fourier transform is the vector of coefficients $\hat{\boldsymbol{f}} = (\hat{f}_0, \hat{f}_1, \ldots, \hat{f}_{n-1})$ such that

$$\boldsymbol{f} = \sum_{k=0}^{n-1} \hat{f}_k \boldsymbol{w}_n^{(k)}, \qquad (21)$$

where the vector $\boldsymbol{f}$ is the vector of the sampled values of the function $f(t)$ sampled at the specified sample points. Orthonormality of the Fourier basis in the standard ℓ^2 inner-product space allows us to identify the Fourier co-

efficients

$$\hat{f}_k = \left\langle \boldsymbol{w}_n^{(k)}, \boldsymbol{f} \right\rangle = \frac{1}{n} \sum_{\ell=0}^{n-1} \omega_n^{-k\ell} f_\ell. \qquad (22)$$

In this example (to illustrate the signals-processing version of aliasing) we select the same number of basis functions n as training points, that is $m = n$. The testing points are all other points in the interval. The design matrix is a variant of the Vandermonde matrix

$$\mathbf{M}_{\mathcal{TM}} = \frac{1}{n} \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega_n^{-1} & \omega_n^{-2} & \cdots & \omega_n^{-(n-1)} \\ 1 & \omega_n^{-2} & \omega_n^{-4} & \cdots & \omega_n^{-2(n-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega_n^{-(n-1)} & \omega_n^{-2(n-1)} & \cdots & \omega_n^{-(n-1)^2} \end{pmatrix},$$
$$(23)$$

and the unmodeled block $\mathbf{M}_{\mathcal{TU}}$ is bi-infinite with n rows and columns. Since $\omega_n^{\ell n} = 1$ for any integer ℓ, the unmodeled block $\mathbf{M}_{\mathcal{TU}}$ is equal to an infinite number of copies of the design matrix

$$\mathbf{M}_{\mathcal{TU}} = \begin{pmatrix} \cdots & \mathbf{M}_{\mathcal{TM}} & \mathbf{M}_{\mathcal{TM}} & \cdots \end{pmatrix}.$$

Because we have selected $m = n$, the design matrix is full rank, and $\mathbf{M}_{\mathcal{TM}}^+ = \mathbf{M}_{\mathcal{TM}}^{-1}$. Thus $\mathbf{A} = \mathbf{M}_{\mathcal{TM}}^{-1}\mathbf{M}_{\mathcal{TU}}$ and $\mathbf{B} = \mathbf{I}_n$. This gives

$$\mathbf{A} = \mathbf{M}_{\mathcal{TM}}^{-1} \begin{pmatrix} \cdots & \mathbf{M}_{\mathcal{TM}} & \mathbf{M}_{\mathcal{TM}} & \cdots \end{pmatrix} \qquad (24)$$

$$= \begin{pmatrix} \cdots & \mathbf{I}_n & \mathbf{I}_n & \mathbf{I}_n & \cdots \end{pmatrix}; \qquad (25)$$

that is, $\mathbf{A}$ is a bi-infinite matrix with n rows and infinitely many columns in both directions, and it corresponds to infinitely many copies of the $n \times n$ identity matrix $\mathbf{I}_n$.

This derivation aligns exactly with the traditional concept of aliasing in the signals-processing literature [36], where the first column of the ℓ-th copy of $\mathbf{I}_n$ in $\mathbf{A}$ corresponds to the ℓnth mode of the system, which is exactly aliased to the 0-th mode; the second column of each copy of I_n corresponds to the $(\ell n + 1)$-th mode which is exactly aliased to the first mode of the actual signal, and so forth. Unless the signal is band-limited, an infinite number of modes are aliased to each of the modeled modes. Traditionally, the aliasing effect is not significant because signals are assumed to have most of their strength in the lower frequencies, that is the magnitude of the higher modes θ_k is assumed to decay to 0 rapidly as $k \to \pm\infty$, which means that although $\mathbf{A}$ is bi-infinite, its effect is minimal on the actual representation of the signal.

This mathematical derivation is represented visually in the left panel of Fig. 4: although basis functions are independent over the entire prediction domain, they may make identical predictions over the sampled subset (red dots). If the true signal contains contributions from all basis functions, but only a subset is explicitly modeled, the contribution from the unmodeled modes is aliased into the truncated representation.

The right panel shows three fits for an artificial data set using the Fourier basis. The true signal (black) includes contributions from all Fourier modes (although the low-frequency modes dominate). The classical sweet spot (blue) only models the dominant modes and produces a reasonable interpolation. At the interpolation threshold (red), however, the aliasing operator magnifies the unmodeled modes, producing large swings in the model predictions between the training samples. Beyond the interpolation threshold (green), the additional, high-frequency basis elements temper the aliasing effects by redistributing the signal among multiple basis functions. The result is a rapidly oscillating signal that does not exhibit the wild swings of overfitting. Although the oscillations in this inferred signal do not match those of the true signal, they are statistically similar, leading to reasonable model predictions.

C. Differential Equations

Despite their superficial dissimilarity, it has been recognized for decades that solution methods for differential equations are formally equivalent to data fitting problems, as we see here. A linear ordinary differential equation can be written as $\mathcal{L}[u](x) = f(x)$ where $\mathcal{L}$ is a linear differential operator, $u(x)$ is the unknown function, and $f(x)$ is a given function, often referred to as the "data." As an example in this section we use the simple case where $\mathcal{L}[u] = u''(x)$, which describes the transverse displacement of a string under tension with transverse loading force given by $f(x)$. The fundamental concepts, however, are much more general than this simple example.

To solve such equations numerically, many approximation schemes exist in which $u(x)$ is approximated in some finite-dimensional subspace, such as with finite-differences, Galerkin truncation, or a collocation method. In nearly all cases, the schemes lead to computational problems formally equivalent to the regression problem described in section II A, which we now demonstrate explicitly for collocation. Expand $u(x) = \sum_\ell \theta_\ell \psi_\ell(x)$ in some basis $\{\psi_\ell\}$; the differential equation then becomes $\sum_\ell \theta_\ell \mathcal{L}[\psi_\ell](x) = f(x)$. In the collocation approach, we enforce that this equation is satisfied exactly at several sample points (training points) x_i, so that $\mathcal{L}[u](x_i) = f(x_i)$ for each training point. If we denote $\phi_\ell(x) = \mathcal{L}[\psi_\ell](x)$, these conditions take the form:

$$\sum_\ell \theta_\ell \phi_\ell(x_i) = f(x_i), \quad i = 1, \ldots, n. \qquad (26)$$

The collocation problem is then to choose basis functions $\psi_\ell(x)$ (and by extension $\phi_\ell(x)$) and collocation points x_i such that solving Eq. (26) leads to as small error as possible throughout the entire domain. This problem formulation is equivalent to a regression problem and the generalized aliasing decomposition gives insights into the structure of the errors.
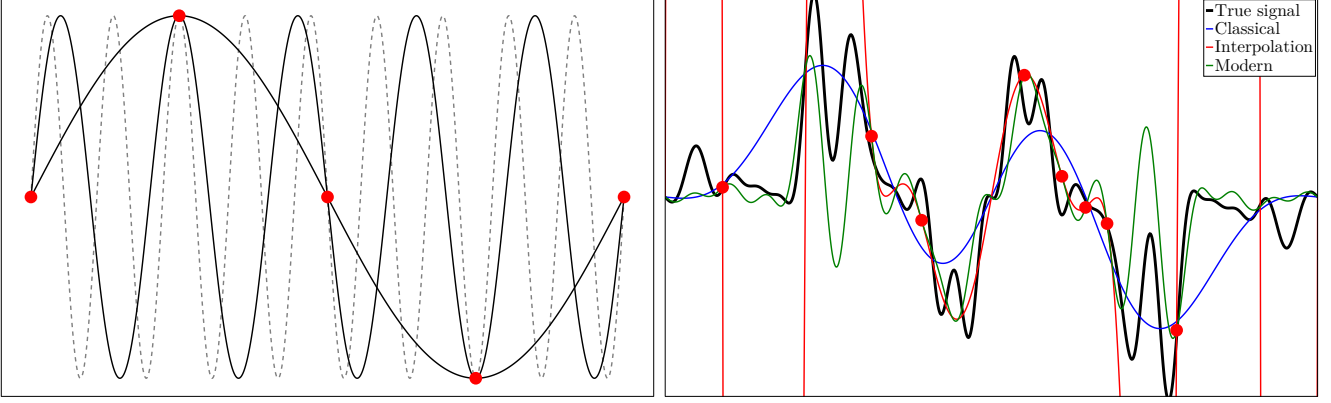
FIG. 4. Aliasing occurs when basis functions that are independent over the entire domain are linearly dependent at the sampled points (left). When fitting a noisy signal (right), the classical sweet spot includes the dominant modes in the signal (blue). Over-fitting occurs when the combined contribution from the unmodeled modes is aliased into the model parameters, producing wild swings in the model predictions (red). Including additional terms allows the learning algorithm to distribute that signal over several basis terms. The result is a model whose predictions oscillate rapidly on a scale that is statistically similar to the true signal (green).

To make these ideas concrete, consider the specific case of

$$u''(x) = x \tag{27}$$
$$u(0) = u(\pi) = 0,$$

which has the solution $u(x) = (x^3 - \pi^2 x)/6$.

We first solve this problem using a sine basis $\psi_\ell(x) = \sin(\ell x)$ with 32 uniformly spaced collocation points and 32 validation points uniformly spaced between them. The resulting GAD for this problem is shown in the upper left of Figure 5. Because the Fourier basis is orthonormal with respect to the uniformly spaced points, the aliasing operator has unit norm. This scenario is equivalent to the traditional understanding of aliasing as presented in section III B, and indeed, aliasing has functionally no effect on the risk curve. The elimination of $\mathbf{A}$ as a factor in the risk arises because this selected basis is composed of exactly the orthonormal eigenfunction basis of the Sturm-Liouville problem defined by Eq. (27).

For this specific, well-adapted basis the risk curve has a weak "U" shape that reflects the trade-off between the data and model insufficiency. The minimum occurs precisely at the interpolation threshold where these two contributions to the error are balanced, demonstrating why the Fourier transform (and the sine series employed here) are most optimal at the interpolation threshold. The striking absence of any aliasing effects is because the basis is optimally adapted to the sampling points. However, this result is sensitive to many aspects of the problem formulation, as we now explore through the lens of the generalized aliasing decomposition.

Moving horizontally across the top row in Figure 5 we explore the sensitivity of the solution to the choice of collocation and validation points. If the sampling points (training and validation) are weakly perturbed from uniform spacing (top row, middle), aliasing emerges around the interpolation threshold along with the characteristic double descent peak. Because the exact solution is continuous, the Fourier series converges rapidly; corresponding to a rapidly decreasing model insufficiency $\|\mathbf{I}_\mathcal{U}\boldsymbol{\theta}\|$. But the aliasing contribution $\|\mathbf{A}\theta_\mathcal{U}\|$ to the parameter error (and hence to risk) is large. Consequently, the optimal solution is no longer at the interpolation threshold; it occurs at the classical sweet spot. In this case, the minimum is to the left of the interpolation threshold rather than to the right, because the basis is ordered with dominant terms first. Consistent with our analysis in section II C 5, the asymptotic limit is suboptimal due to data insufficiency.

The second row of Figure 5 shows the results for a cubic b-spline basis with 65 uniformly spaced knots throughout the domain. We omit the two basis functions that do not satisfy the boundary conditions for a basis of 64 functions. Because there is no natural ordering, we randomly shuffle the basis functions, making the model equivalent to a random feature model. Moving horizontally across the second row of the figure shows the GAD for the same choice of sampling points as for the Fourier basis.

In all three cases, notice that the contributions from aliasing are absent. This is because the basis functions have relatively small compact support, which strongly limits their ability to alias with each other. We observe similar results for other choices of bases with relatively small compact support, such as Haar wavelets. But, unlike the other rows, in this basis data insufficiency $\|\mathbf{P}_\mathcal{N}\boldsymbol{\theta}_\mathcal{M}\|$ becomes nonzero long before the interpolation threshold. This is also at least partly due to the relatively small compact support—the basis functions are each zero throughout most of the domain, so they have a nontrivial nullspace for our choice of sampling points.
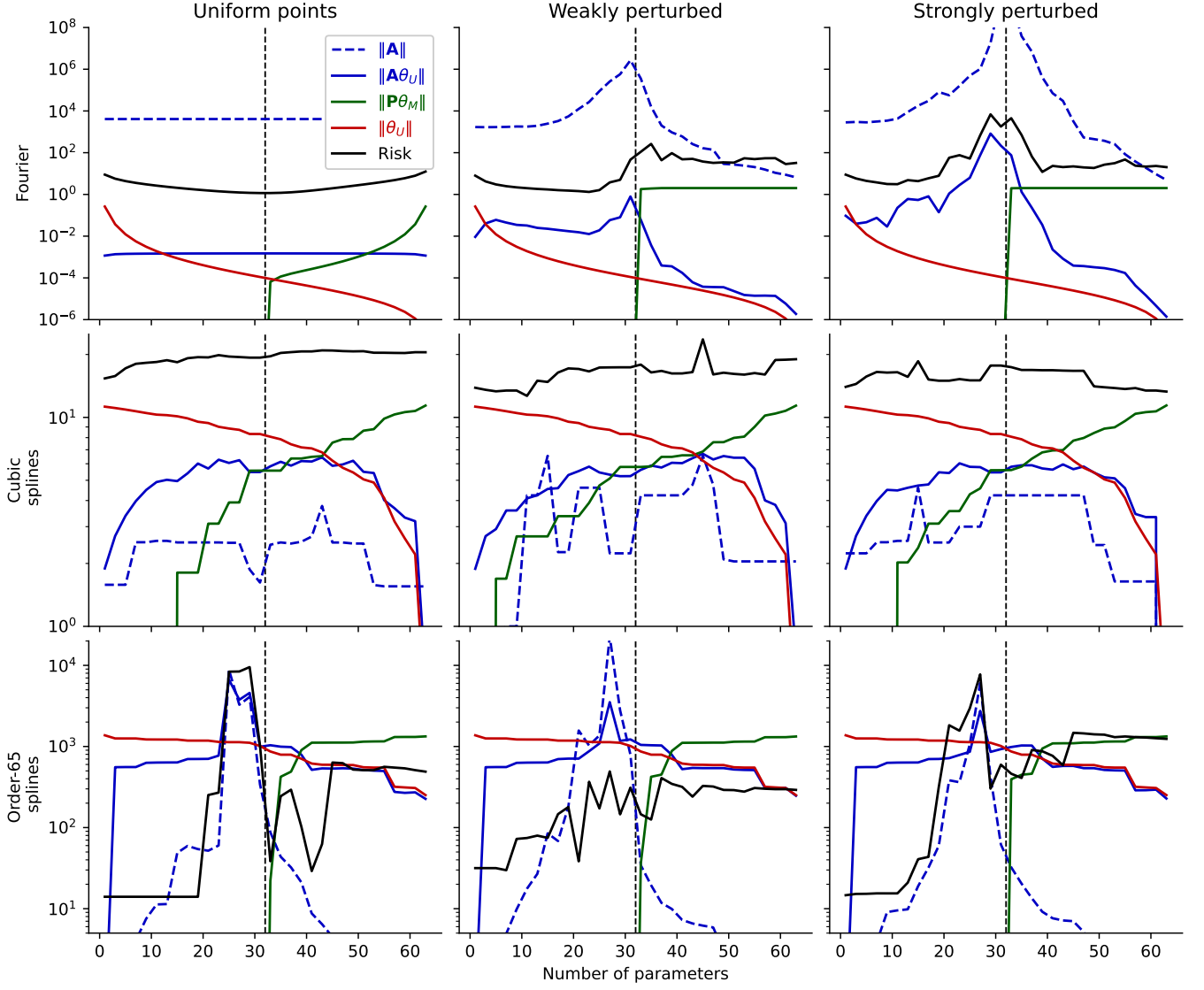
FIG. 5. Generalized aliasing decomposition for numerical solutions to the differential equation 27 using different bases and sampling schemes. In each case there are 32 collocation (training) points and 32 validation points. The dashed vertical black line marks the interpolation threshold. In the first column, the sample points (collocation and validation)are uniformly spaced. Moving to the right, the sample points are increasingly perturbed by a random amount. The first row is a Fourier-sine basis. Because these basis functions are orthonormal at the uniform points, the norm $\|\mathbf{A}\|$ of the aliasing operator is essentially constant in the upper left panel, but as the sample points are perturbed, the basis functions move away from being orthonormal and aliasing increases near the interpolation threshold. The second row corresponds to a basis of cubic splines. These have fairly narrow, compact support, so the basis functions are close to being orthogonal on all the sample points, keeping the aliasing small across the row, but the small support means model insufficiency $\|\mathbf{I}_{\mathcal{U}}\boldsymbol{\theta}_{\mathcal{U}}\|$ drops off much more slowly than in the Fourier case, and it also causes the data insufficiency $\|\mathbf{P}_{\mathcal{N}}\boldsymbol{\theta}\|$ to become significant long before the interpolation threshold. The bottom row corresponds to higher-order splines. These have support across the full domain, which makes for nontrivial aliasing for all the different choices of sample points, but the data insufficiency is very small until the interpolation threshold. These high-order splines are also highly localized, which means the model insufficiency $\|\mathbf{I}_{\mathcal{U}}\boldsymbol{\theta}_{\mathcal{U}}\|$ drops off more slowly than in the other rows. See Section III C for more details about this example.

This phenomenon is similar to that observed for the random ReLU basis on CIFAR-10 in the upper middle panel of Figure 3.

Finally, on the bottom row we apply another b-spline basis of 65th-order polynomials. As before, we remove the two basis functions that do not satisfy the boundary conditions and shuffle the remaining basis functions. While these splines are technically continuous and non-

zero throughout the entire domain, each basis function is strongly peaked around a small portion of the domain. In this case, we see some contributions from aliasing before the interpolation threshold, but it is not as prominent as with the Fourier basis and is essentially independent of the choice of sampling points.

In general, a similar analysis can be applied to other differential operators, including partial differential equations. Indeed, the effects of aliasing (and the need for dealiasing) are well known in the simulation of nonlinear partial differential equations (see [37] for the original reference or [22] for a more thorough discussion). The current decomposition applies there as well, and some results are also known for nonlinear equations, as we now summarize.

The generic setup for a quadratic nonlinearity would be of the form

$$\frac{d\boldsymbol{y}}{dt} \approx \boldsymbol{y} \odot \boldsymbol{y},$$

where $\odot$ is the Hadamard (entrywise) product. In this setting, the labels $\boldsymbol{y}$ denote a spatially and temporally dependent function described by the basis functions in the design matrix $\mathbf{M}$, i.e. $\boldsymbol{y} = \mathbf{M}\boldsymbol{\theta}$. To solve this system, we note that the differential equation can be written as:

$$\mathbf{M}\dot{\boldsymbol{\theta}} \approx (\mathbf{M}\boldsymbol{\theta}) \odot (\mathbf{M}\boldsymbol{\theta}),$$

where the $\dot{}$ refers to the time derivative. If the entire basis $\mathbf{M}$ could be used, then the solution is obtained by multiplying on the left by the appropriate pseudoinverse $\mathbf{M}^+$. In reality, all of $\mathbf{M}$ is not available, and so we decompose the system just as before, leading to a term on the right-hand side that resembles the aliasing operator, but now with a quadratic dependence on the unmodeled terms $\mathbf{M}_{\mathcal{TU}}$ which leads to a famous "3/2 rule" for pseudospectral methods [37].

D. Material Discovery: Cluster Expansion

As a final example we consider a problem from computational discovery of new materials. The *cluster expansion* is an extremely efficient model for exploring new possible phases. The cluster expansion model is a generalized Ising model [38–46] that in typical applications has hundreds to thousands of data points and a dozen to hundreds of inferred parameters. The prototypical application of the cluster expansion is predicting the energy E of an alloy as a function of elemental composition and configuration $\vec{\sigma}$.

$$E(\vec{\sigma}) = J_0 + \sum_i J_1 \xi_i + \sum_{\beta}^{\text{pairs}} \sum_{i,j} J_\beta \xi_i \xi_j +$$
$$\sum_{\gamma}^{\text{triplets}} \sum_{i,j,k} J_\gamma \xi_i \xi_j \xi_k + \sum_{\nu}^{\text{quads}} \sum_{i,j,k,l} J_4 \xi_i \xi_j \xi_k \xi_l + \cdots$$
$$= \sum_{\alpha}^{\text{clusters}} J_\alpha \Phi_\alpha(\vec{\sigma}) \qquad (28)$$

where the indices run over all the possible sites, pairs of sites, triples, and so on. The J's are expansion coefficients (inferred parameters, analogous to the θs in the notation above). The products of $\xi(\vec{\sigma})$ functions[47] form an orthogonal basis $\{\Phi_\alpha\}_\alpha$ in the discrete vector space of all possible atomic configurations.

This model has a physically intuitive interpretation as representing chemical bonds between groups of atoms. For example, a product of two functions, $\xi_i \xi_j$, represents a pairwise interaction between atoms on sites i and j. The strength of the interaction is the magnitude $|J_{ij}|$, and the sign of J_{ij} determines whether like or unlike atoms prefer to be ij-neighbors.

The CE interactions J_α are typically inferred from quantum-mechanical energies. There are no obvious strategies for picking which alloy configurations to include in the training set. As to this question—the horizontal partitioning of $\mathbf{M}$ (deciding the which rows of $\mathbf{M}$ should be in $\mathbf{M}_{\mathcal{TM}}$ and which should be in $\mathbf{M}_{\mathcal{PM}}$ in Eq. (8))—one often included the "usual suspects," configurations that often occur in actual alloys; but this rarely provides enough information to generate a model with small generalization error.

Choosing which basis functions to include in the model is even more difficult; it is difficult to know which physical interactions are negligible. (The vertical partition of $\mathbf{M}$ in Eq. (8) divides the important interactions from those that are assumed to be negligible.) Many different strategies to address these two challenges have been employed in the CE community [43–45, 48–56].

Using the generalized aliasing decomposition, CE practitioners can now reason more effectively about how to make these two difficult choices—which configurations to sample and which basis functions to include in the modeling. Furthermore, the GAD elegantly explains the complex risk curves of a typical alloy system (see, e.g., Figure 6). One can enumerate all possible configurations (up to some maximum number of atoms) [57–59], identifying all the rows of the universe matrix $\mathbf{M}$ for this problem. (In principle the number of rows is countably infinite, but under mild assumptions, one can enumerate all configurations up to a size that effectively includes all configurations that are likely to appear in nature.) It is also feasible to determine a complete set of basis functions [39] for the enumerated configurations. [60]

We demonstrate in the following study of a Pt-Cu alloy. Choosing a realistic model size, we explain the re-
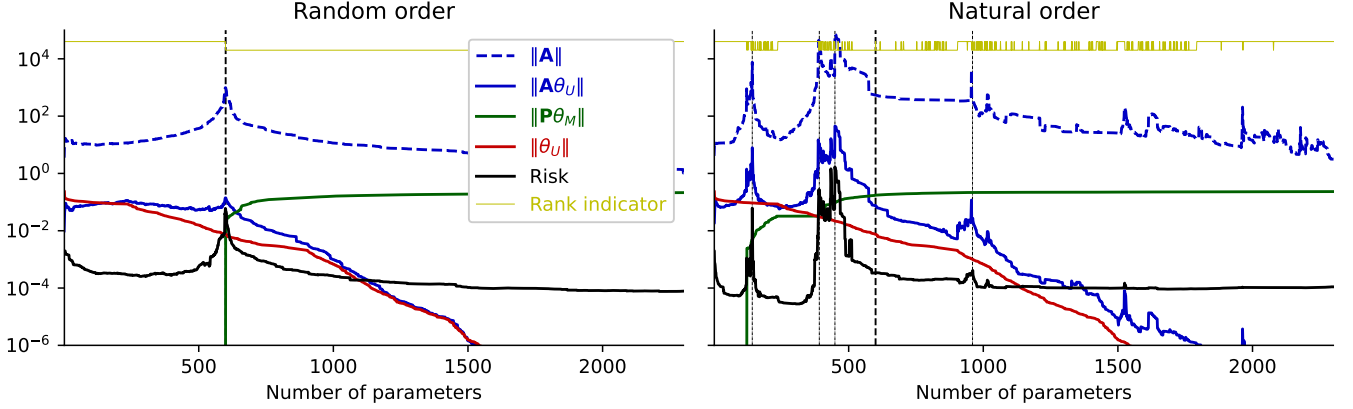
FIG. 6. Norm of the operators and true risk (black) of the cluster expansion model of Section III D as model complexity is increased (i.e., as parameters are added) with 600 training points. The interpolation threshold is indicated by a vertical dashed line. The yellow at the top of each panel is an indicator function that is high when the added basis function is linearly independent of the previous columns (restricted to training points) and low when it is linearly dependent. In the left panel rows and columns of $\mathbf{M}$ have been randomly ordered. In the right panel the rows and columns of the design matrix are given a "natural" ordering, resulting in multiple peaks (indicated by dash-dot vertical lines) and valleys for the risk and aliasing. Note also how the data insufficiency $\|\mathbf{P}_{\mathcal{N}}\boldsymbol{\theta}_{\mathcal{M}}\|$ becomes significant only at the interpolation threshold for the random ordering, but it becomes significant long before the interpolation threshold in the natural ordering.

sulting generalization curve through the lens of the GAD. A binary alloy model containing up to ten unique atomic sites has 2346 unique configurations.[61] Figure 6 shows the norms of the aliasing matrix, the model insufficiency, and the data insufficiency as a function of increasing basis size for a fixed number of training points.

Unlike the Fourier example, where a natural ordering is obvious, in this setting the "right" ordering is not clear. At first, we impose no assumptions about natural ordering to either the data or the basis functions parameters, randomizing rows and columns of $\mathbf{M}$. This is similar to the random feature models in Section III A. The risk behavior (left panel of Fig. 6) exhibits the prototypical double descent. Before the interpolation threshold, the behavior of the risk curve is the expected U-shape of the classical bias–variance trade-off. And beyond the interpolation threshold, the risk drops again. As predicted in the random feature discussion in Section II C 5, the lowest-risk model is not at the classical sweet spot but in the asymptotic limit as $m \to \infty$.

But cluster expansion practitioners do have some intuition about the natural order for the sample points and basis functions. In their preferred ordering for cluster expansion, pair-wise interactions precede triplet interactions, and all triplet interactions come before any quadruplets, and so forth. Furthermore, the terms are ordered in each class by diameter—short pairs before long pairs, small-diameter triplets before extended triplets, etc. This ordering is motivated by physical arguments that the strongest interactions are short-range and low body-order. For the ordering of the sample points (atomic configurations, defining rows of $\mathbf{M}$), there is the coarse guideline of ordering by "size," denoted by the number of

atoms in each configuration, but within each size class a natural ordering is not obvious.

This natural ordering of n-body/short-long was used to arrange columns and rows in $\mathbf{M}$ for the risk curve shown in the right panel of Fig. 6. The aliasing $\|\mathbf{A}\|$ and the risk move essentially in unison and show a complicated behavior, neither the typical U-shape of classical bias–variance trade-off nor the basic double descent. Rather, the generalization curve has multiple peaks and valleys, whose positions correspond to locations where added basis functions transition from linear independence to linear dependence (marked by the yellow "indicator function" at the top of the plot). Vertical dash-dot lines are included to clarify the connection between peaks in the operator norm $\|\mathbf{A}\|$ (dashed blue) and peaks in the risk (solid black). Surprisingly (for the bias–variance paradigm), the interpolation threshold does not seem to play any role in the generalization curve for this naturally ordered case.

Finally, note also that for the naturally ordered case, the optimal risk occurs at a classical "sweet-spot," with a low number of parameters, and is better (lower) than the optimal risk in the randomly ordered case, which occurs in the asymptotic limit as the number of parameters grows large.

IV. DISCUSSION

We have demonstrated the utility of the GAD for explaining complicated, non-monotonic risk curves in a variety of different settings. Now we turn our attention to using the GAD for improved model development and the

pursuit of more efficient and accurate representations of the data. We will discuss the impact this decomposition has on modeling and sampling decisions, and the influence those decisions have on the shape of generalized risk curves.

A. General Insights into Modeling

The formal analysis provided above, as well as the examples demonstrating the GAD give practical, intuitive guidance for formulating models which we discuss here.

1. Choosing the Basis

If the n training points are known and fixed, a modeler can control the norm of $\mathbf{M}_{\mathcal{TM}}^+$ and $\mathbf{A}$ (and hence generically control the magnitude of the risk) by strategically choosing the basis functions, without knowing anything about the labels $\boldsymbol{y}$.

For example, consider what happens when we choose the first n basis functions ϕ_k so that, when evaluated at the points $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n$, the resulting vectors $\boldsymbol{\varphi}_k = (\phi_k(\boldsymbol{t}_1), \ldots, \phi_k(\boldsymbol{t}_n))$ are orthonormal. If the columns of $\mathbf{M}_{\mathcal{TM}}$ are the first $m \leq n$ of these vectors, then the inverse $\mathbf{M}_{\mathcal{TM}}^+$ always has induced norm $\|\mathbf{M}_{\mathcal{TM}}^+\| = 1$. In this situation the norm is constant as m increases up to n; and then for $m > n$, no matter which additional columns are added, the norm $\|\mathbf{M}_{\mathcal{TM}}^+\|$ cannot increase and will eventually shrink to 0 (almost surely). Thus, the product $\|\mathbf{M}_{\mathcal{TM}}^+\|\|\mathbf{M}_{\mathcal{TU}}\|$ in the upper bound

$$\|\mathbf{A}\| \leq \|\mathbf{M}_{\mathcal{TM}}^+\|\|\mathbf{M}_{\mathcal{TU}}\|$$

on the norm of $\mathbf{A}$ also can never increase with m. Given a prior on the coefficients $\boldsymbol{\theta}$, if the basis functions are ordered to reflect the expected magnitudes of the corresponding coefficients, then we expect there to be no peak in $\|\mathbf{A}\boldsymbol{\theta}_{\mathcal{U}}\|$ at all—only descent.

In the discrete Fourier series example of Section III B, the norm of $\mathbf{A}$ is always 1 and does not decrease to 0 because the columns of $\mathcal{M}$ are specially tuned to the training set to make $\mathbf{M}_{\mathcal{TU}}$ consist of infinitely many copies of $\mathbf{M}_{\mathcal{TM}}$. This aligning of the basis functions to sample points explains why extreme over-fitting is rarely a problem in discrete Fourier transforms, in spite of it being formally equivalent to ordinary least squares regression at the interpolation threshold.

2. Choosing Training Points

If the basis functions are given and fixed, but the modeler has control over the choice of the training points, then they can control the norm $\|\mathbf{A}\|$ by strategically choosing the points $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n$. Again, this requires no knowledge of the labels $\boldsymbol{y}$.
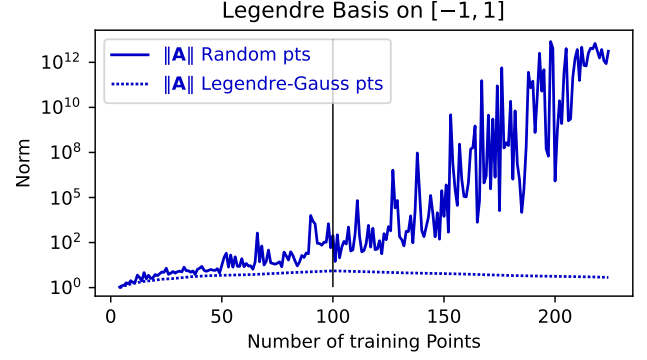

FIG. 7. The induced norm $\|\mathbf{A}\|$ of the aliasing operator for the Legendre polynomial basis with the model consisting of the first $m = 100$ Legendre polynomials. The norms are plotted as functions of the number n of training points, and the vertical black line indicates the interpolation threshold (note that this is inverted from the plots in the previous figures where n is fixed and m is variable). The solid blue shows the norm of the aliasing operator when the training points are chosen randomly (drawn uniformly from $[-1, 1]$), while the dotted blue shows the norm of the aliasing operator when the n training points are chosen to be the Legendre–Gauss points (the zeros of the nth Legendre polynomial). The norm $\|\mathbf{A}\|$ for randomly chosen training points rapidly grows to be many orders of magnitude larger than for the Legendre–Gauss points.

For example, consider the case of fitting polynomial functions on the interval $[-1, 1]$ with the Legendre basis, consisting of polynomials $\{P_k\}_{k \in \mathbb{N}}$ which are orthogonal with respect to the inner product $\langle f, g \rangle = \int_{-1}^{1} f(t)g(t)\, dt$, with P_k of degree k and $P_k(1) = 1$ for all k. For a given number m of model parameters (the first m Legendre polynomials), if we are able to choose n points at which to evaluate the basis functions, then choosing the points to be the n Legendre–Gauss points, which are the zeros of P_n, gives much better results than choosing the points randomly (drawn uniformly). This is shown in Figure 7, where the randomly chosen training points make $\|\mathbf{A}\|$ many orders of magnitude larger than with the specially chosen Legendre–Gauss points. In this case a judicious choice of training points makes a huge difference to $\|\mathbf{A}\|$. Except for very special choices of $\boldsymbol{\theta}$, this means the risk Eq. (16) will also be substantially larger when the model is trained on random points than when it is trained on the specially chosen Legendre–Gauss points.

3. Conditioning of $\mathbf{M}$

If $\mathbf{M}$ is poorly conditioned, then it is possible to have a relatively small error $\mathbf{E}_{\boldsymbol{\theta}}\boldsymbol{\theta}$ in the parameters that corresponds to a large error in the signal (large risk). Thus it is desirable to select a basis that makes the full transformation $\mathbf{M}$ well conditioned.

For polynomial approximation with the standard monomial basis $\{1, t, t^2, \dots\}$, the transformation $\mathbf{M}$ is a generalized Vandermonde matrix, which is very badly conditioned and generally should not be used with real-valued inputs [62]. But polynomial approximation for real inputs in the interval $[-1, 1]$ is well conditioned with the Chebyshev polynomial basis or the Legendre polynomial basis.

B. Regularization

It has been observed that L^2-regularization (ridge regression) generally reduces the size of the peak in risk at the interpolation threshold, but it can also increase risk elsewhere along the curve [35, 63, 64]. This can be understood in terms of the impact of regularization on the GAD and the pseudoinverse of the design matrix.

For a given decomposition $\Theta = \mathcal{M} \oplus \mathcal{U}$ of the space Θ with $m = \dim \mathcal{M}$ model parameters, ridge regression amounts to changing the objective from minimizing risk to minimizing

$$\frac{1}{n}\|\boldsymbol{y} - \mathbf{M}_{\mathcal{TM}}\boldsymbol{\theta}_{\mathcal{M}}\|_2^2 + \lambda\|\boldsymbol{\theta}_{\mathcal{M}}\|_2^2, \qquad (29)$$

where n is the number of training points and λ is a user-chosen parameter. It is straightforward to verify that the objective to minimize with L_2-regularization can be written as $\frac{1}{n}\left\|\widetilde{\boldsymbol{y}} - \widetilde{\mathbf{M}}_{\mathcal{TM}}\boldsymbol{\theta}_{\mathcal{M}}\right\|_2^2$, where $\widetilde{\mathbf{M}}_{\mathcal{TM}} = \begin{bmatrix} \mathbf{M}_{\mathcal{TM}} \\ \sqrt{n\lambda}\mathbf{I}_m \end{bmatrix}$ and $\widetilde{\boldsymbol{y}} = \begin{bmatrix} \boldsymbol{y} \\ \mathbf{0} \end{bmatrix}$. This changes the GAD to to

$$\widetilde{\mathbf{E}}_{\boldsymbol{\theta}} = \begin{bmatrix} I_{\mathcal{M}} - \widetilde{\mathbf{M}}_{\mathcal{TM}}^+ \mathbf{M}_{\mathcal{TM}} & -\widetilde{\mathbf{M}}_{\mathcal{TM}}^+ \mathbf{M}_{\mathcal{TM}} \\ 0 & I_{\mathcal{U}} \end{bmatrix},$$

so aliasing $\mathbf{A}$ becomes $\widetilde{\mathbf{A}} = \widetilde{\mathbf{M}}_{\mathcal{TM}}^+ \mathbf{M}_{\mathcal{TU}}$, and data insufficiency $\mathbf{P}_{\mathcal{N}}$ becomes $\widehat{\mathbf{P}}_{\mathcal{N}} = I_{\mathcal{M}} - \widetilde{\mathbf{M}}_{\mathcal{TM}}^+ \mathbf{M}_{\mathcal{TM}}$, while $\mathbf{I}_{\mathcal{U}}$ remains unchanged. Expanding the product

$$\widetilde{\mathbf{M}}_{\mathcal{TM}}\widetilde{\mathbf{M}}_{\mathcal{TM}}^{\mathsf{T}} = \mathbf{M}_{\mathcal{TM}}\mathbf{M}_{\mathcal{TM}}^{\mathsf{T}} + n\lambda\mathbf{I}_m,$$

shows that every eigenvalue of $\mathbf{M}_{\mathcal{TM}}\mathbf{M}_{\mathcal{TM}}^{\mathsf{T}}$ is now increased by $n\lambda$ in this product. Therefore, the singular values of $\mathbf{M}_{\mathcal{TM}}$ are all increased by $\sqrt{n\lambda}$ in $\widetilde{\mathbf{M}}_{\mathcal{TM}}$, and

$$\begin{aligned}\|\widetilde{\mathbf{M}}_{\mathcal{TM}}^+\| &= \frac{1}{\frac{1}{\|\mathbf{M}_{\mathcal{TM}}^+\|} + \sqrt{n\lambda}} \\ &= \frac{\|\mathbf{M}_{\mathcal{TM}}^+\|}{1 + \sqrt{n\lambda}\|\mathbf{M}_{\mathcal{TM}}^+\|} \leq \frac{1}{\sqrt{n\lambda}}.\end{aligned}$$

This bound is independent of both m and $\mathbf{M}_{\mathcal{TM}}$, and it essentially removes the impact of any small singular values of $\mathbf{M}_{\mathcal{TM}}$ on the norms of $\widetilde{\mathbf{M}}_{\mathcal{TM}}^+$ and $\widetilde{\mathbf{A}}$. This explains why there is no significant peak in the risk at the interpolation threshold (or anywhere else, for that matter) for L_2-regularized (ridge regression) problems, provided λ is sufficiently large. If $\sqrt{n\lambda} > \|\mathbf{M}_{\mathcal{TU}}\|$, then

the norm of $\widetilde{\mathbf{A}}$ is smaller than the norms of data insufficiency and model insufficiency, which then dominate the parameter error.

The contribution $\|\widetilde{\mathbf{P}}_{\mathcal{N}}\boldsymbol{\theta}_{\mathcal{M}}|$ of data insufficiency to parameter error, however, can increase with regularization because it is no longer the projection of $\boldsymbol{\theta}_{\mathcal{M}}$ onto the null space $\mathcal{N}$ of $\mathbf{M}_{\mathcal{TM}}$ or $\widetilde{\mathbf{M}}_{\mathcal{TM}}$ but instead is

$$\|\widetilde{\mathbf{P}}_{\mathcal{N}}\boldsymbol{\theta}_{\mathcal{M}}\| = \|(\mathbf{I}_{\mathcal{M}} - \widetilde{\mathbf{M}}_{\mathcal{TM}}^+ \mathbf{M}_{\mathcal{TM}})\boldsymbol{\theta}_{\mathcal{M}}\|.$$

When λ is large, the fact that $\|\widetilde{\mathbf{M}}_{\mathcal{TM}}^+ \mathbf{M}_{\mathcal{TM}}\| \leq \frac{\|\mathbf{M}_{\mathcal{TM}}\|}{\sqrt{n\lambda}}$, means that the data insufficiency operator approaches $\mathbf{I}_M$ and the contribution to parameter error from data insufficiency approaches $\|\boldsymbol{\theta}_{\mathcal{M}}\|$, which is generally larger than the projection $\|\mathbf{P}_{\mathcal{N}}\boldsymbol{\theta}_{\mathcal{M}}\|$ onto the kernel $\mathcal{N}$ of $\mathbf{M}_{\mathcal{TM}}$.

Nevertheless, the norm of $\widetilde{\mathbf{P}}_{\mathcal{N}}$, while no longer necessarily bounded by 1, is still bounded by

$$\|\widetilde{\mathbf{P}}_{\mathcal{N}}\| \leq 1 + \frac{\|\mathbf{M}_{\mathcal{TM}}\|}{\sqrt{n\lambda}}.$$

C. How to think about the unmodeled signal $\mathbf{M}_{\mathcal{TU}}$?

The fundamental ansatz of generalized aliasing is the decomposition in Eq. (8) that decomposes the signal into both modeled and unmodeled components. Our conception of unmodeled signal is similar to "noise" as understood in classical and modern statistics. Indeed, in comparing Eqs. (1) and (10), the unmodeled modes naively correspond to the noise in classical regression. This decomposition may initially seem unnatural since any unmodeled components are unknown and consequently, difficult to reason about. For some readers, this decomposition may seem an unnecessary introduction at best or an untractable complication at worst. However, the concept is useful for distinguishing nuances in the modeling processes that are obscured by the traditional conception of statistical noise.

First, the heart of the GAD is recognizing that model representations are embedded within a universal function space. This way of thinking is strongly motivated by signal processing, in which a signal is often assumed to have contributions from all modes, even if only a subset can be extracted from a finite sampling. This enables us to quantify the trade-off between the modeled and the unmodeled contributions and their relative informativity. In this way, the GAD naturally quantifies the intuition that as the model capacity grows, unmodeled signal necessarily shrinks. In contrast, in the classical formulation, noise is modeled as a random variable whose scale parameter is not necessarily tied to the complexity of the model except as a tunable hyperparameter. Thus, by quantifying the trade-off between the modeled and the unmodeled, we quantify the informational relationship between the data and the model.

Furthermore, recognizing the unmodeled allows flexibility in solving problems where random variables are not

the natural representation. For example, approximating the solution to differential equations is formally equivalent to regression. However, considerable information is known about the analytic nature of these solutions, and it is often more natural to represent the unmodeled piece as another continuous signal from a set for which there is no natural measure. In many applications, such as robust control, one is interested in worst-case scenarios. In such cases, one takes the extremal case over the allowed set of unmodeled signals rather than an expectation value over a random variable.

Finally, our conception of unmodeled signal encompasses any limitations in representing either models or data. Something as insipid as finite-precision arithmetic is a form of unmodeled signal that is not commonly equated with statistical "noise." For example, consider the representation of a band-limited signal. The Fourier sampling theorem guarantees its finite Fourier representation can be reconstructed from finite samples. And when the signal is sampled at generic, random points with an infinite precision representation, such a signal can still be exactly reconstructed. In finite precision, however, the represented signal is no longer band-limited: round-off error introduces small, high frequency contributions. Even if the high-frequency components introduced by the finite precision are bounded, aliasing greatly magnifies their impact on the reconstructed signal, and the ill-conditioning of the aliasing operator leads to large errors in the inferred signal.

This final point reflects a much deeper philosophical issue when modeling data, which we summarize as *fidelity* and *sensitivity to representation*. The concept of fidelity can be understood as the extent to which a representation is faithful to the real physical process. Again, consider the example of Fourier analysis. The utility of Fourier representations are often attributed to the fact that smooth functions have rapidly decaying Fourier series. Consequently, a truncated Fourier representation of a smooth signal is faithful to the truth, in the sense that they are nearby in Fourier space. Although the truncated series is formally wrong, its representational error is bounded.

In contrast, it often possible to apply inaccurate approximations to models that nevertheless make accurate predictions. When this occurs, a model exhibits insensitivity to the representation. Arguably, the most famous and important example of this is the concept of *irrelevance* in renormalization in statistical physics. In renormalization, approximations are made not because they are accurate but because they do not affect observables of interest. In Kadanoff's block-spin renormalization of the Ising model, groups of spins are aggregated into a single block-spin; that is, they are approximated as being perfectly correlated. While such approximations are inaccurate for modeling spin correlations, they lead to good approximations of phase transitions. The details of microscopic correlations are said to be *irrelevant* to macroscopic observables. On the other hand, the phase diagram is very sensitive to the so-called *relevant* parameters, such as the applied field or temperature. Small variations in these variables significantly impact the macroscopic order parameter.

The concepts of relevance in renormalization and sensitivity to unmodeled signals in generalized aliasing have conceptual similarity. A Fourier representation reconstructed from uniformly spaced samples is useful, not only because it is dominated by low frequency modes, but also because the reconstructed signal is insensitive to high-frequency, unmodeled contributions. In contrast, Fourier series from random samples exhibit strong sensitivity to unmodeled modes that distorts the coarse trends in the reconstructed signal. In the language of renormalization, high-frequency modes are irrelevant for uniform samples, but random samples render the high-frequency modes relevant.

More recent work has informed similar conclusions about the general nature of predictive modeling. Within the formalism of so-called *sloppy models*, microscopic details of complicated, multi-parameter models can be safely ignored because observables of interest are insensitive to large variations in these parameters [4, 7, 65]. Indeed, it has been found that many useful approximations may be derived by taking parameters to extreme values [66]. Even more fundamentally, evolutionary psychology has shown that psychological representations that maximize fitness are often not faithful to physical reality [67]. That is, human psychological representations of reality may be dictated more by the sensitivity of evolutionary fitness to the representation than by fidelity to reality. All of this suggests that when building a physical model, sensitivity to the unmodeled must be accounted for at least as much as fidelity to known physics.

D. Outlook

Successful model building involves numerous technical decisions related to the selection of model class, experimental design, learning algorithm, regularization, and other factors that can strongly impact the model's predictive performance. Best practices are more often art, tuned to experience, rather than science guided by formal reasoning. The generalized aliasing decomposition (Eq. (17)) facilitates reasoning about key modeling decisions in a way that is both formal and intuitive. In the context of linear regression, the approach is fully rigorous while imbuing practitioners with intuition about model performance in both the classical and modern regimes. Because the aliasing operator norm can be computed without knowing labels, practitioners can also make informed choices about data collection and experimental design for target applications.

Although our formal analysis has been restricted to linear regression, there are reasons to be optimistic that the core approach generalizes to the nonlinear regime. First, the concepts of aliasing and invertibility (or projection to

the kernel) extend formally to nonlinear operators and can be approximated through local linearization. Furthermore, many cases of practical importance may be tractable in the present framework. Results for weak, quadratic nonlinearities already exist for pseudospectral methods in partial differential equations [37]. Neural tangent kernel techniques demonstrate that wide networks are linear in their models throughout training [68]. In addition, information geometry techniques applied to large, sloppy models have shown that most nonlinearity is "parameter-effects" and removable, in principle, through an appropriate, nonlinear reparameterization [69].

An important open question is: Under what conditions is the asymptotic risk less than that of the classical "sweet spot"? The preceding analysis has sharpened that question to: When will data insufficiency be larger than the error at the classical sweet spot? While this remains an open question in general, we have begun to explore it for two broad cases. Random feature models, such as in Figure 3, but presumably also neural networks and other machine learning models, often do not exhibit large data insufficiency and are generically most effective in the over-parameterized, modern regime. In contrast, we have argued that physics-based models are most effective in the classical regime, where they leverage prior knowledge.

Framing the question in this way clarifies why classical statistics historically missed these interesting phenomena, in spite of the essential elements being known to diverse communities for decades [13]. It also apparently partitions predictive modeling into two philosophically distinct camps: physical models using classical statistics and unstructured models in the modern, interpolating regime. In our cluster expansion example, the former approach gave the model with the least risk. Although perhaps expected, as physics-based modeling leverages prior information, this benefit comes after considerable effort from the materials science community. However, it remains unclear if these are inherently irreconcilable philosophies or two points on a broad landscape just beginning to be explored.

Indeed, our work demonstrates how the theoretical and technical challenges posed by modern data science overlap with those in other fields, including signal processing, control theory, and statistical physics. We hope that the perspectives advanced here will inspire theorists and practitioners alike to better understand and leverage the relationship between data science and the broader scientific milieu.

[1] N. Goldenfeld, Simple lessons from complexity, Science **284**, 87 (1999).

[2] E. P. Hoel, L. Albantakis, and G. Tononi, Quantifying causal emergence shows that macro can beat micro, Proceedings of the National Academy of Sciences **110**, 19790 (2013).

[3] J. P. Crutchfield, The dreams of theory, WIREs Computational Statistics , 75.

[4] M. K. Transtrum, B. B. Machta, K. S. Brown, B. C. Daniels, C. R. Myers, and J. P. Sethna, Perspective: Sloppiness and emergent theories in physics, biology, and beyond, The Journal of Chemical Physics **143**, 010901 (2015).

[5] H. H. Mattingly, M. K. Transtrum, M. C. Abbott, and B. B. Machta, Maximizing the information learned from finite data selects a simple model, Proceedings of the National Academy of Sciences **115**, 1760 (2018).

[6] P. Chvykov and E. Hoel, Causal Geometry, Entropy **23**, 24.

[7] K. N. Quinn, M. C. Abbott, M. K. Transtrum, B. B. Machta, and J. P. Sethna, Information geometry for multiparameter models: new perspectives on the origin of simplicity, Reports on Progress in Physics **86**, 035901.

[8] M. Belkin, D. J. Hsu, S. Ma, and S. Mandal, Reconciling modern machine-learning practice and the classical bias-variance trade-off, Proceedings of the National Academy of Sciences **116**, 15849 (2019).

[9] Z. Yang, Y. Yu, C. You, J. Steinhardt, and Y. Ma, Rethinking bias-variance trade-off for generalization of neural networks, in *International Conference on Machine Learning* (PMLR, 2020) pp. 10767–10777.

[10] Y. Dar, V. Muthukumar, and R. G. Baraniuk, A farewell to the bias-variance tradeoff? an overview of the theory of overparameterized machine learning, arXiv preprint arXiv:2109.02355 (2021).

[11] S. Geman, E. Bienenstock, and R. Doursat, Neural networks and the bias/variance dilemma, Neural computation **4**, 1 (1992).

[12] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, Surprises in high-dimensional ridgeless least squares interpolation, The Annals of Statistics **50**, 949 (2022).

[13] M. Loog, T. Viering, A. Mey, J. H. Krijthe, and D. M. Tax, A brief prehistory of double descent, Proceedings of the National Academy of Sciences **117**, 10625 (2020).

[14] L. Chen, Y. Min, M. Belkin, and A. Karbasi, Multiple descent: Design your own generalization curve, in *Advances in Neural Information Processing Systems*, Vol. 34, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, Inc., 2021) pp. 8898–8912.

[15] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, Deep double descent: Where bigger models and more data hurt, Journal of Statistical Mechanics: Theory and Experiment **2021**, 124003 (2021).

[16] S. d'Ascoli, M. Refinetti, G. Biroli, and F. Krzakala, Double trouble in double descent: Bias and variance (s) in

the lazy regime, in *International Conference on Machine Learning* (PMLR, 2020) pp. 2280–2290.

[17] B. Adlam and J. Pennington, Understanding double descent requires a fine-grained bias-variance decomposition, Advances in neural information processing systems **33**, 11022 (2020).

[18] E. H. Lee and V. Cherkassky, Vc theoretical explanation of double descent, arXiv preprint arXiv:2205.15549 (2022).

[19] L. Oneto, S. Ridella, and D. Anguita, Do we really need a new theory to understand the double-descent?, in *ESANN* (2022).

[20] R. Schaeffer, M. Khona, Z. Robertson, A. Boopathy, K. Pistunova, J. W. Rocks, I. R. Fiete, and O. Koyejo, Double descent demystified: Identifying, interpreting & ablating the sources of a deep learning puzzle, arXiv preprint arXiv:2303.14151 (2023).

[21] M. Lafon and A. Thomas, Understanding the double descent phenomenon in deep learning, arXiv preprint arXiv:2403.10459 (2024).

[22] J. P. Boyd, *Chebyshev and Fourier spectral methods* (Courier Corporation, 2001).

[23] B. Neal, S. Mittal, A. Baratin, V. Tantia, M. Scicluna, S. Lacoste-Julien, and I. Mitliagkas, A modern take on the bias-variance tradeoff in neural networks, arXiv preprint arXiv:1810.08591 (2018).

[24] Although fitting monomials is the canonical pedagogical example, the ill conditioning of this Vandermonde matrix makes this basis ill-suited for practical applications.

[25] J. Humpherys, T. J. Jarvis, and E. J. Evans, *Foundations of applied mathematics. Vol. 1* (Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017) pp. xx+689, mathematical analysis.

[26] X. Sheng and T. Wang, An iterative method to compute moore-penrose inverse based on gradient maximal convergence rate, Filomat **27**, 1269 (2013).

[27] R. M. Gower and P. Richtárik, Randomized quasi-newton updates are linearly convergent matrix inversion algorithms, SIAM Journal on Matrix Analysis and Applications **38**, 1380 (2017).

[28] Any real analytic function is determined by its values on a dense set, so we may limit ourselves to only considering rational points t in the interval $[a, b]$.

[29] L. Wasserman, *All of statistics*, Springer Texts in Statistics (Springer-Verlag, New York, 2004) pp. xx+442, a concise course in statistical inference.

[30] F. P. Gantmacher and M. G. Krein, *Oscillation matrices and kernels and small vibrations of mechanical systems*, revised ed. (AMS Chelsea Publishing, Providence, RI, 2002) pp. viii+310, translation based on the 1941 Russian original, Edited and with a preface by Alex Eremenko.

[31] J. R. Bunch and C. P. Nielsen, Updating the singular value decomposition, Numer. Math. **31**, 111 (1978/79).

[32] R. C. Thompson, The behavior of eigenvalues and singular values under perturbations of restricted rank, Linear Algebra Appl. **13**, 69 (1976), collection of articles dedicated to Olga Taussky Todd.

[33] J. H. Wilkinson, *The algebraic eigenvalue problem* (Clarendon Press, Oxford, 1965) pp. xviii+662.

[34] M. Rudelson and R. Vershynin, Non-asymptotic theory of random matrices: extreme singular values, in *Proceedings of the International Congress of Mathematicians. Volume III* (Hindustan Book Agency, New Delhi, 2010) pp. 1576–1602, https://www.math.uci.edu/ rvershyn/papers/rv-

ICM2010.pdf.

[35] S. Mei and A. Montanari, The generalization error of random features regression: Precise asymptotics and the double descent curve, Communications on Pure and Applied Mathematics **75**, 667 (2022), https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.22008.

[36] R. A. Roberts and C. T. Mullis, *Digital signal processing* (Addison-Wesley Longman Publishing Co., Inc., 1987).

[37] S. A. Orszag, On the elimination of aliasing in finite-difference schemes by filtering high-wavenumber components, Journal of Atmospheric Sciences **28**, 1074 (1971).

[38] J. M. Sanchez, F. Ducastelle, and D. Gratias, Generalized cluster description of multicomponent systems, Physica A: Statistical Mechanics and its Applications **128**, 334 (1984).

[39] J. Sanchez, Cluster expansions and the configurational energy of alloys, Physical review B **48**, 14013 (1993).

[40] J. Sanchez, Cluster expansion and the configurational theory of alloys, Physical Review B–Condensed Matter and Materials Physics **81**, 224202 (2010).

[41] A. Zunger, P. Turchi, and A. Gonis, Statics and dynamics of alloy phase transformations, NATO ASI Series. Series B, Physics **319** (1994).

[42] A. Van De Walle, M. Asta, and G. Ceder, The alloy theoretic automated toolkit: A user guide, Calphad **26**, 539 (2002).

[43] T. Mueller and G. Ceder, Bayesian approach to cluster expansions, Physical Review B—Condensed Matter and Materials Physics **80**, 024103 (2009).

[44] D. Lerch, O. Wieckhorst, G. L. Hart, R. W. Forcade, and S. Müller, Uncle: a code for constructing cluster expansions for arbitrary lattices with minimal user-input, Modelling and Simulation in Materials Science and Engineering **17**, 055003 (2009).

[45] M. Ångqvist, W. A. Muñoz, J. M. Rahm, E. Fransson, C. Durniak, P. Rozyczko, T. H. Rod, and P. Erhart, Icet–a python library for constructing and sampling alloy cluster expansions, Advanced Theory and Simulations **2**, 1900015 (2019).

[46] A. Seko, K. Yuge, F. Oba, A. Kuwabara, and I. Tanaka, Prediction of ground-state structures and order-disorder phase transitions in ii-iii spinel oxides: A combined cluster-expansion method and first-principles study, Physical Review B–Condensed Matter and Materials Physics **73**, 184117 (2006).

[47] The *site functions* themselves are usually discrete Cheybschev polynomials or a Fourier basis. Any functions that form an orthonormal set over the discrete values of σ_i are suitable.

[48] G. L. Hart, V. Blum, M. J. Walorski, and A. Zunger, Evolutionary approach for determining first-principles hamiltonians, Nature materials **4**, 391 (2005).

[49] V. Blum, G. L. W. Hart, M. J. Walorski, and A. Zunger, Using genetic algorithms to map first-principles results to model hamiltonians: Application to the generalized ising model for alloys, Phys. Rev. B **72**, 165113 (2005).

[50] A. Seko, Y. Koyama, and I. Tanaka, Cluster expansion method for multicomponent systems based on optimal selection of structures for density-functional theory calculations, Physical Review B—Condensed Matter and Materials Physics **80**, 165122 (2009).

[51] L. J. Nelson, V. Ozoliņš, C. S. Reese, F. Zhou, and G. L. Hart, Cluster expansion made easy with bayesian compressive sensing, Physical Review B—Condensed Matter

and Materials Physics **88**, 155105 (2013).

[52] L. J. Nelson, G. L. Hart, F. Zhou, and V. Ozoliņš, Compressive sensing as a paradigm for building physics models, Physical Review B—Condensed Matter and Materials Physics **87**, 035125 (2013).

[53] J. M. Sanchez, Renormalized interactions in truncated cluster expansions, Phys. Rev. B **99**, 134206 (2019).

[54] A. van de Walle, M. D. Asta, and G. Ceder, The Alloy Theoretic Automated Toolkit: A user guide, Calphad **26**, 539 (2002).

[55] B. Puchala and A. Van der Ven, Thermodynamics of the zr-o system from first-principles calculations, Physical Review B—Condensed Matter and Materials Physics **88**, 094108 (2013).

[56] Z. Leong and T. L. Tan, Robust cluster expansion of multicomponent systems using structured sparsity, Physical Review B **100**, 134108 (2019).

[57] G. L. Hart and R. W. Forcade, Algorithm for generating derivative structures, Physical Review B–Condensed Matter and Materials Physics **77**, 224115 (2008).

[58] G. L. Hart and R. W. Forcade, Generating derivative structures from multilattices: Algorithm and application to hcp alloys, Physical Review B–Condensed Matter and Materials Physics **80**, 014120 (2009).

[59] G. L. Hart, L. J. Nelson, and R. W. Forcade, Generating derivative structures at a fixed concentration, Computational Materials Science **59**, 101 (2012).

[60] Until recently, enumerating all linearly independent basis functions, without also generating many linearly dependent basis functions, was an outstanding problem. This new algorithm is not yet published.

[61] This formation enthalpy data were generated by "unrelaxed" Density Functional Theory calculations for configurations of platinum and copper, which were then adjusted by a linear regression using slight regularization to smooth out noise.

[62] The Vandermonde matrix *is* well conditioned in the special case that the inputs t all lie on the unit circle $U^1 = \{t \in \mathbb{C} : |z| = 1\}$, but it is badly conditioned if the inputs t do not have unit modulus.

[63] F. F. Yilmaz and R. Heckel, Regularization-wise double descent: Why it occurs and how to eliminate it, in *2022 IEEE International Symposium on Information Theory (ISIT)* (2022) pp. 426–431.

[64] P. Nakkiran, P. Venkat, S. Kakade, and T. Ma, Optimal Regularization Can Mitigate Double Descent, arXiv e-prints , arXiv:2003.01897 (2020), arXiv:2003.01897 [cs.LG].

[65] B. B. Machta, R. Chachra, M. K. Transtrum, and J. P. Sethna, Parameter space compression underlies emergent theories and predictive models, Science **342**, 604 (2013).

[66] M. K. Transtrum and P. Qiu, Model reduction by manifold boundaries, Physical Review Letters **113**, 098701 (2014).

[67] D. D. Hoffman, M. Singh, and C. Prakash, The interface theory of perception, Psychonomic bulletin & review **22**, 1480 (2015).

[68] J. Lee, L. Xiao, S. S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington, Wide neural networks of any depth evolve as linear models under gradient descent [*], Journal of Statistical Mechanics: Theory and Experiment **2020**, 124002 (2020).

[69] M. K. Transtrum, B. B. Machta, and J. P. Sethna, Geometry of nonlinear least squares with applications to sloppy models and optimization, Physical Review E **83**, 036701 (2011).