Harmless interpolation of noisy data in regression

Vidya Muthukumar  Kailas Vodrahalli  Anant Sahai

BLISS and ML4Wireless
Department of Electrical Engineering and Computer Sciences, UC Berkeley

March 22, 2019

Abstract
A continuing mystery in understanding the empirical success of deep neural networks has been in their ability to achieve zero training error and yet generalize well, even when the training data is noisy and there are more parameters than data points. We investigate this "overparametrization" phenomena in the classical underdetermined linear regression problem, where all solutions that minimize training error interpolate the data, including noise. We give a bound on how well such interpolative solutions can generalize to fresh test data, and show that this bound generically decays to zero with the number of extra features, thus characterizing an explicit benefit of overparameterization. For appropriately sparse linear models, we provide a hybrid interpolating scheme (combining classical sparse recovery schemes with harmless noise-fitting) to achieve generalization error close to the bound on interpolative solutions.

1 Introduction
In the standard statistical machine learning setup, we have high-dimensional data in the form of \( n \) covariate-response pairs \( (X_i, Y_i)_{i=1}^n \). When training parametric models for functions fitting covariate to response, the traditional wisdom \( [1] \) is to select function classes with a number of parameters \( d < n \). In classification problems (i.e. when the labels \( Y \) are discrete), the scaling of the test error with respect to \( n \) is determined by characterizations of the VC-dimension \( [2] \)/Rademacher complexity \( [3] \) of the function class, which in the worst case increases with its number of parameters. In regression (i.e. when the labels \( Y \) are continuous), the mean-squared error of the ordinary least-squares estimate is characterized by the condition number of the regression matrix, which is reasonable for appropriate ratios of \( d/n \) but tends to increase astronomically as \( d \) approaches \( n \). The qualitative fear is the same: if the function class is too complex, it starts to overfit noise and can generalize poorly to unseen test data.

But there is a gap between "can" and "will" — and indeed this conventional wisdom has been challenged by the recent advent of deeper and deeper neural networks. In particular, a thought-provoking paper \( [4] \) noted that several deep neural networks generalize well despite achieving zero or close to zero training error, and being so expressive that they even have the ability to fit pure noise. As they put it, "understanding deep learning requires rethinking generalization". How can we reconcile the fact that good interpolative solutions exist with the classical bias-variance tradeoff?

These phenomena are being actively investigated in a statistical sense \( [5,6] \) and a computational sense \( [7,9] \) in classification problems and/or noiseless models. (We include additional discussion of related work \( [10,18] \) in Appendix \( A \).) Surprisingly, they have not been investigated for overparametrized linear regression in the presence of noise, which is both naturally defined and has been well studied under the assumption of a sparse linear generative model. Can we interpolate the noise and retain favorable generalization guarantees? Is there a price for interpolation?

In this paper, we analyze linear regression in an overparameterized setting and provide constructive answers to the above questions using elementary machinery. Our contributions are:

1 This paper was presented at ITA in San Diego on February 15, 2019 after being submitted to ISIT 2019.
1. We give a fundamental limit for the excess MSE of any interpolative solution and show that this excess converges to 0 as the number of features goes to infinity under a general sub-Gaussian assumption on the random covariates.

2. We show the performance of the \( \ell_2 \)-norm interpolator depends on the structure of the features used.

3. We provide a hybrid estimator that recovers the sparse signal using known estimators and uses the ideal interpolative solution to fit noise, thus achieving generalization performance close to the bound for the optimal interpolator. We also observe that sparse signal recovery methods run as if in a noiseless environment perform surprisingly well.

Thus, we conclude that the right set of features can be conducive to finding harmless interpolative solutions.

2 Problem Setting

We consider covariate-response pairs \((X, Y) \in \mathbb{R}^p \times \mathbb{R}\) and generative model \(Y = \langle a(X), \alpha^* \rangle + W\) for “lifted” feature vector \(a(X) \in \mathbb{R}^d\) and Gaussian noise \(W \sim \mathcal{N}(0, \sigma^2)\) independent of \(X\). The regressor \(\alpha^*\) is unknown apriori to an estimator. We also assume a distribution on \(X \in \mathbb{R}^p\), which induces a distribution on the \(d\)-dimensional feature vector \(a(X)\). Let \(\Sigma = \mathbb{E}[a(X)a(X)\top]\) denote the covariance matrix of the feature vector under this induced distribution. We assume that \(\Sigma\) is invertible; therefore it is positive definite and its square-root-inverse \(\Sigma^{-1/2}\) exists.

We define shorthand notation for the training data:

\[
A_{\text{train}} := \begin{bmatrix} a(X_1)\top & a(X_2)\top & \ldots & a(X_n)\top \end{bmatrix} \top
\]

denotes the design matrix, and let \(Y_{\text{train}}, W_{\text{train}} \in \mathbb{R}^n\) denote the output and noise vectors respectively.

We will primarily consider the interpolative regime, i.e. where \(d > n\), and thus are interested in solutions \(\alpha\) that satisfy the following feasibility condition for interpolation:

\[
A_{\text{train}}\alpha = Y_{\text{train}} \tag{1}
\]

Observe that the set \(\{\alpha \in \mathbb{R}^d : A_{\text{train}}\alpha = Y_{\text{train}}\}\) is non-empty as long as \(A_{\text{train}}\) is full-rank (i.e. of rank \(n\)), which we will assume.

For any solution \(\hat{\alpha} \in \mathbb{R}^d\), we define the expected mean squared test error below.

**Definition 1.** The expected mean squared test error (MSE) of any estimator \(\hat{\alpha}((X_i, Y_i)_{i=1}^n)\) is given by

\[
\mathcal{E}_{\text{test}}(\hat{\alpha}) := \mathbb{E}[(Y - \langle a(X), \hat{\alpha} \rangle)^2]
\]

where the expectation is taken only over the joint distribution on the fresh test sample \((X, Y)\).

3 The ideal interpolative solution

The first question we posed is whether interpolation can ever lead to a desirable guarantee on the test MSE \(\mathcal{E}_{\text{test}}\). The constraint in Equation (1) is sufficiently restrictive to not allow trivial solutions of the form \(\alpha = \alpha^*\) — so this is a surprisingly well-posed problem. In fact, we can define the ideal interpolator below.

**Definition 2.** The ideal interpolator \(\hat{\alpha}_{\text{ideal}}\) is defined as:

\[
\hat{\alpha}_{\text{ideal}} := \arg \min \mathcal{E}_{\text{test}}(\alpha) \quad \text{subject to} \quad A_{\text{train}}\alpha = Y_{\text{train}}.
\]

We also denote the ideal test error as \(\mathcal{E}_{\text{test}}^*\). This is a lower bound on the expected test error of any scheme that interpolates these training points using these features.
Theorem 1. For any joint distribution on \((X,Y)\), we have
\[
\mathcal{E}^*_{\text{test}} = W_{\text{train}}^T B_{\text{train}}^T (B_{\text{train}} B_{\text{train}}^T)^{-1} W_{\text{train}} + \sigma^2,
\] (2)
where \(B_{\text{train}} := A_{\text{train}} \Sigma^{-1/2}\).

The proof of Theorem 1 is outlined in Section 3.2. It proceeds by characterizing exactly the interpolative solution that minimizes test MSE and calculating its performance. Theorem 1 provides an explicit expression for a fundamental limit on the generalization ability of interpolation. Thus, we can easily evaluate it (numerically) for a number of commonly used feature sets, listed below.

1. iid Gaussian features: \(a(X) \sim N(0, I)\).
2. correlated Gaussian features: \(a(X) \sim N(0, \Sigma)\) for some invertible \(\Sigma\).
3. random Fourier features \cite{19}: \(X \in \mathbb{R}^d, a(X)_j = e^{i \langle v_j, X \rangle}\) for some \(v_j \sim N(0, I)\).
4. polynomial Vandermode features: \(a(X) = \begin{bmatrix} 1 & X & X^2 & \ldots & X^{d-1} \end{bmatrix}\).
5. Legendre polynomial features (form an orthonormal basis).

Figure 3 evaluates the ideal interpolative error \(\mathcal{E}^*_{\text{test}}\) as a function of \(d\) for iid Gaussian features. We observe a spike at \(d = n\), and a decay in the generalization error as \(d >> n\), implying that favorable noise-fitting is indeed possible for these features. Appendix ?? includes additional empirical results showing favorable noise-fitting for random Fourier features.

3.1 Conditions for favorable noise fitting

We can theoretically explain the favorable noise-fitting ability that we empirically observed for Gaussian and random Fourier features. Noteworthy from the expression in Equation (2) is that the test MSE naturally depends on the singular value spectrum of \(B_{\text{train}}^T\) (which can be thought of as a whitened version of \(A_{\text{train}}\)) and in the worst case its minimum singular value. This matrix has independent columns and more importantly, becomes taller as the number of parameters \(d\) increases. Thus, if the features are distributed such that they themselves have favorable concentration properties, we expect that the matrix \(B_{\text{train}}^T\) becomes better and better conditioned and the error arising solely from fitting noise reduces. This is stated formally in the following corollary.

Corollary 1. Suppose one of the following conditions holds:

1. \(a(X) \sim N(0, \Sigma)\) (Gaussian design)
2. \(a(X)\) is a sub-Gaussian random vector (definitions in Appendix B.1) with parameter \(K\) and its rows are normalized, i.e. \(\|a(X)\|_2 = \sqrt{d}\).

Then, we can bound the ideal test error as
\[
\mathcal{E}^*_{\text{test}} \leq \left( \frac{n(1 + \delta)}{(\sqrt{d} - C_K \sqrt{n})^2} + 1 \right) \sigma^2
\]
with probability greater than or equal to \((1 - e^{-c_K n} - e^{-\delta^2 n/8})\) for constants \(C_K, c_K > 0\) depending on the sub-Gaussian parameter \(K\).

Corollary 1 is true for Gaussian features and random Fourier features (\(a(X)\) is normalized and bounded, therefore sub-Gaussian). Corollary 1 implies that \(\mathcal{E}^*_{\text{test}} \to \sigma^2\) in probability as \(d \to \infty\). Thus, we need not be paying anything extra by fitting the noise other than the noise variance itself (which is of course irreducible). Thus, there always exists an interpolative solution that fits noise in such a manner that the effect of fitting this noise on test error decays to 0 as the number of features goes to infinity.

---

2In general, for \(d < n\) we evaluate the test error of the unique least-squares solution as Equation (1) is no longer feasible.
3.2 Proof of Theorem 1 and Corollary 1

The usual argument gives us

\[ E_{\text{test}}(\hat{a}) = \mathbb{E}[(a(X), a^*) + W - (a(X), \hat{a}))^2] \]
\[ = \mathbb{E}[(a(X), \hat{a} - a^*))^2] + \mathbb{E}[W^2] \]
\[ = \mathbb{E}[(\hat{a} - a^*)^T a(X) a(X)^T (\hat{a} - a^*)] + \sigma^2 \]
\[ = (\hat{a} - a^*)^T \Sigma (\hat{a} - a^*) + \sigma^2 \]
\[ = \|\Sigma^{1/2}(\hat{a} - a^*)\|^2 + \sigma^2, \]

Thus, we can equivalently characterize the ideal interpolating solution \( \hat{a}_{\text{ideal}} := \arg\min E_{\text{test}}(\hat{a}) \) as:

\[ \hat{a}_{\text{ideal}} := \arg\min \|\Sigma^{1/2}(a - a^*)\|_2 \]

subject to Equation (1) holding.

Observe that Equation (1) can be rewritten as

\[ A_{\text{train}}(a - a^*) = W_{\text{train}} \]
\[ \implies A_{\text{train}}\Sigma^{-1/2}\Sigma^{1/2}(a - a^*) = W_{\text{train}} \]

Then we have a closed form expression for the minimum norm solution, denoting \( B_{\text{train}} = A_{\text{train}}\Sigma^{-1/2} \):

\[ \Sigma^{1/2}(\hat{a}_{\text{ideal}} - a^*) = B_{\text{train}}^\dagger(B_{\text{train}}B_{\text{train}}^\dagger)^{-1}W_{\text{train}} \]

(note that \( B_{\text{train}}^\dagger := B_{\text{train}}^\dagger(B_{\text{train}}B_{\text{train}}^\dagger)^{-1} \) is just the right Moore-Penrose pseudoinverse of \( B_{\text{train}} \)).

Substituting this expression into the test error calculation:

\[ E_{\text{test}}^* = E_{\text{test}}(\hat{a}_{\text{ideal}}) = \|B_{\text{train}}^\dagger(B_{\text{train}}B_{\text{train}}^\dagger)^{-1}W_{\text{train}}\|_2^2 + \sigma^2 \]
\[ = W_{\text{train}}^\dagger(B_{\text{train}}B_{\text{train}}^\dagger)^{-1}W_{\text{train}} + \sigma^2 \]

Plugging this into the expression of \( E_{\text{test}}^* \) gives us Equation (2), thus proving Theorem 1.

Next, we use results from matrix concentration theory [20] to prove Corollary 1. Denoting \( B_{\text{train}}^\dagger = [b_1 \ldots b_n] \), we can upper bound Equation (2) as

\[ E_{\text{test}}^* \leq \frac{\|W_{\text{train}}\|_2^2}{\lambda_{\text{min}}((B_{\text{train}}^\dagger)^{-1}B_{\text{train}}^\dagger)} + \sigma^2. \]

Observe that the columns of \( B_{\text{train}}^\dagger \) are now isotropic, i.e. \( \mathbb{E}[b_i b_i^\top] = I \) for all \( i = 1, 2, \ldots, n \), and independent. Thus, we can apply Theorem 5.58 from [20] (restated in Appendix B.1) directly - this is a concentration result on the minimum singular value of \( B_{\text{train}}^\dagger \). That gives us

\[ E_{\text{test}}^* \leq \frac{\|W_{\text{train}}\|_2^2}{(\sqrt{d} - C_{\text{K}}\sqrt{n})^2} + \sigma^2 \]

with probability greater than or equal to \((1 - e^{-CKn})\).

Further, we have \( \|W_{\text{train}}\|_2^2 = \sum_{i=1}^n W_i^2 \) and a standard tail bound on chi-squared random variables gives us

\[ \|W_{\text{train}}\|_2^2 \leq n\sigma^2(1 + \delta) \]

with probability greater than or equal to \((1 - e^{-n\delta^2/8})\). Using the union bound on these high-probability statements completes the proof.
### 3.3 The interpolation threshold and regularity of training data

We observe that Corollary 1 is meaningful primarily for the heavily overparameterized regime where \( d \gg n \) (more formally, if we vary \( d \) as a function of \( n \), we have \( \lim_{n \to \infty} \frac{n}{d(n)} = 0 \)). It does not mathematically explain the magnitude of the interpolation peak at \( d \sim n \) that we observe in Figure 2, which reflects the mean of the minimum possible test MSE that results purely from fitting noise at the interpolation threshold.

It is well-known that the behavior of the minimum singular value of a random matrix with iid Gaussian entries is very different when the matrix is approximately square – a "hard-edge" phenomenon arises \(^{21,22}\) and the minimum singular value actually exhibits heavy-tailed behavior \(^{23}\). It is unclear whether the same phenomena hold for lifted features on data such as Fourier, Legendre or even Vandermonde features. It thus makes sense to consider the statistics of the test MSE at the interpolation threshold more carefully.

---

**Figure 1.** Test MSE of the ideal interpolator (and other practical interpolation schemes) of Legendre and Fourier features for regularly spaced and randomly drawn training points, true fit is a 2-degree polynomial fit.

Figure 1 shows the generalization performance of various interpolative solutions on approximately "whitened" Legendre and Fourier features on data. In particular, we consider two generative assumptions for the covariates \( \{X_i\}_{i=1}^n \):

1. **Regularly spaced data:** \( X_i = -1 + \frac{2(i-1)}{(n-1)} \) for all \( i = 1, 2, \ldots, n \).

2. **Randomly drawn data:** \( X_i \sim \text{Unif}[−1, 1] \).

and the generative model for the data to be \( Y_i = f^*(X_i) + W_i \) where \( f^* : [−1, 1] \to \mathbb{R} \) is a polynomial of degree 2. We make 50 random draws of both the covariates (in the case of randomly drawn covariates) and noise, and plot both the mean and the median test MSE evaluated over \( 10^4 \) test points, together with 85% confidence intervals. The results depicted in Figure 1 have several implications. In particular, the distributional assumptions on the covariates make a significant difference: regularly spaced data approximately
preserves the orthogonality of the design matrix $A_{\text{train}} \approx B_{\text{train}}$. On the other hand, randomly drawn training points constitute a more complex random matrix $A_{\text{train}}$, for which the minimum singular value is much lower and moreover exhibits heavy-tailed behavior. This can be observed from the large size of the confidence intervals, as well as the much lower median test MSE (which is still higher than the test MSE incurred from regularly spaced training data). Randomly drawn training data is what we use in the conventional ML step, and this analysis tells us that we truly need to overparameterize to ensure harmless fitting of noise, unlike in logistic regression.

The performance of practical interpolators in Figure 1 is also interesting: in particular, the minimum $\ell_2$-norm interpolator does poorly even as the level of overparameterization increases. Schemes like OMP to completion (which intuitively takes advantage of the sparsity of the model) appear to have lower test MSE. We will now discuss these practical interpolation schemes.

4 Practical interpolation schemes

![Figure 2](image)

**Figure 2.** Test MSE for polynomial and Gaussian features. Polynomials – here, data is sampled from $\text{Unif}[-1,1]$ with $n = 10$, $k = 2$ (a degree 1 polynomial), and $W \sim N(0,10^{-4})$. The unwhitened case uses Vandermonde features. The whitened version of these features correspond to the Legendre polynomial features. Gaussian – we sample from $N(0,1)$ and use the same settings as for polynomials. Notice that with the appropriate whitening, even polynomial interpolation can benefit from overparameterization but that the scale clearly indicates that polynomial features suffer more from noise-fitting than Gaussian features.

The proof of Theorem 1 tells us that the ideal interpolator fits what can be thought of as the ideal residual, $\alpha - \alpha^*$, to noise in such a way to minimize its $\ell_2$-norm weighted by the (square root of the) covariance matrix $\Sigma$. This is well and truly an ideal interpolator, as it requires knowledge of the true signal $\alpha^*$ to implement. We turn to practically used interpolating schemes and characterize their performance.

4.1 $\ell_2$-minimizing solutions

We first consider the practically used minimum $\ell_2$-norm interpolator. Note that gradient descent run on the overparameterized linear regression problem provably converges to the minimum norm solution [24].

**Proposition 1.** Let the minimum norm interpolator be

$$\hat{\alpha}_2 := \arg \min \| \alpha \|_2 \text{ subject to } \alpha \text{ satisfying Equation (1)}$$
Figure 3. Test MSE of a variety of sparse recovery methods for Gaussian data sampled from \(N(0, 1)\). Here, \(n = 5000\) and \(k = 500\) and noise \(W \sim N(0, \sigma^2)\) for different choices of variance \(\sigma^2\).
Then its test error satisfies

\[ \mathcal{E}_{\text{test}}(\hat{\alpha}_2) - \sigma^2 \leq \mathcal{E}_{\text{test}}(\hat{\alpha}_2; \text{noiseless}) + \mathcal{E}_{\text{test}}(\hat{\alpha}_2; \text{noise}) \]

where

\[ \mathcal{E}_{\text{test}}(\hat{\alpha}_2; \text{noiseless}) = \| \Sigma^{1/2}(A_{\text{train}}^\top A_{\text{train}} - I)\alpha^* \|_2^2 \]

\[ \mathcal{E}_{\text{test}}(\hat{\alpha}_2; \text{noise}) = W_{\text{train}}^\top (A_{\text{train}}^\top \Sigma A_{\text{train}}) W_{\text{train}}. \]

The proof of Proposition 1 is provided in Appendix B.3 for completeness. The expression shows a natural split for all \( \ell_2 \)-norm-minimizing interpolators between the error they incur in a noiseless setting on recovering \( \alpha^* \), and error arising solely from fitting noise. This latter quantity corresponds to the test MSE that would be obtained if we chose the solution that minimized \( \| \alpha - \alpha^* \|_2 \).

It is particularly instructive to consider the ramifications of this expression in the context of an example of features that are highly correlated. Consider data \( X \sim \text{Unif}[{-1, 1}] \) and labels \( Y \in \mathbb{R} \) that are truly generated from a 1-degree polynomial model with noise. Provided with 10 samples, we investigate the performance of interpolative linear regressors using Vandermonde features, which are highly correlated. Figure 2 shows that the overparameterized test error using the solution that fits the noise to minimize \( \| \hat{\alpha} - \alpha^* \|_2 \) is substantially higher than that of the ideal solution (which effectively “whitens" the Vandermonde feature vector to the orthonormalized Legendre polynomial basis); indeed, it does much worse at the interpolation threshold as compared to Gaussian features. This highlights that there is a critical difference between the performance of the minimum \( \ell_2 \)-norm and weighted \( \ell_2 \)-norm interpolators, or equivalently, the choice of features in their effects on fitting noise.

On the other hand, even if the features are whitened, the \( \ell_2 \)-norm solution can incur substantial error due to its poor signal recovery in a noiseless setting - this is a well-known observation in signal processing [26]. Figure 2 displays that the error arising purely from signal recovery in the absence of noise using the whitened Legendre polynomial basis is substantial. To understand why this is the case, consider the noiseless error term for the weighted \( \ell_2 \)-norm minimizing solution. After whitening the features, the effective covariance matrix is the identity and bounding the MSE is equivalent to bounding the estimation error rate on the parameter vector \( \alpha^* \). It is well known that in the absence of underlying sparse structure, we unavoidably have non-zero noiseless error in the asymptotic high dimensional regime [27,28].

This implies that the minimum \( \ell_2 \)-norm solution is not a definite indicator for good generalization, and the “double-descent" curves it traces in practice [6] are likely due to an approximation-theoretic benefit.

### 4.2 Sparsity seeking methods

From the results obtained so far, we can think of any successful interpolator as recovering the signal in the noiseless high-dimensional regime and then fit noise optimally as defined in Theorem 1. Proposition 1 showed that the \( \ell_2 \)-interpolator does poorly in the former goal of signal recovery. If the signal \( \alpha^* \) is dense, this is difficult to do regardless of the estimator chosen. On the other hand, if there is underlying sparsity in the signal we know that recovery is possible. However, the estimators used do not interpolate. This subsection revisits the sparse regime through the lens of understanding guarantees on interpolative solutions in the presence of noise.

#### 4.2.1 A hybrid scheme that achieves close to ideal interpolation

Consider \( \alpha^* \) being \( k \)-sparse, i.e. \( |\text{supp}(\alpha^*)| = k \). For this section, also assume that \( \Sigma = I \), i.e. the matrix is iid Gaussian. We propose a natural approach that combines sparse signal recovery guarantees with the optimal noise-fitting described in Theorem 1 and achieves close to the ideal MSE:

1. Use a suitable estimator that knows either the sparsity level \( k \), or the noise variance \( \sigma^2 \), and provides an estimate of the signal \( \hat{\alpha}_1 \).

\[ \text{Notice that in expectation (over iid training noise), these two terms are uncorrelated and hence the inequality would hold with equality.} \]

\[ \text{We are grateful to a concurrent analysis of the minimum } \ell_2 \text{ interpolator for inspiring this experiment [25].} \]
2. Compute the residual $W'_\text{train} = Y_{\text{train}} - A_{\text{train}}\hat{\alpha}_1$.

3. Return $\hat{\alpha}_\text{hybrid} := \arg \min \| \alpha - \hat{\alpha}_1 \|_2$ subject to $\alpha$ satisfying Equation (1).

Observe that the feasibility constraint is just a rewriting of Equation (1), ensuring that the estimator $\hat{\alpha}_\text{hybrid}$ interpolates the data. The generalization guarantee is described in the following proposition below:

**Proposition 2.** For a given sparse-signal estimator $\hat{\alpha}_1$, the hybrid estimator $\hat{\alpha}_\text{hybrid}$ has test MSE

$$E_{\text{test}}(\hat{\alpha}_\text{hybrid}) \leq \| \hat{\alpha}_1 - \alpha^* \|_2^2 + \frac{\| W'_\text{train} \|_2^2}{\lambda_{\min}(A_{\text{train}}A_{\text{train}}^\top)^{-1}} + \sigma^2 \tag{3}$$

As an (informally stated) corollary, we have the following explicit high-probability bounds for the Lasso and OMP estimators with appropriate levels of regularization and stopping condition respectively:

$$E_{\text{test}} \leq \sigma^2 \left( 1 + O\left( \frac{n}{d} \right) + O\left( \frac{k \log d}{d} \right) + O\left( \frac{k \log d}{n} \right) \right)$$

with probability greater than or equal to $(1 - c'e^{-cn})$ for constants $c,c',C,C' > 0$ that depend on the choice of algorithm, but not on $d$ and $n$.

Proposition 2 shows a natural split in the error of such hybrid interpolators in terms of two quantities: the error that arises from signal recovery, and the error that arises from fitting noise. The proof of Equation (3) and the corollaries for Lasso and OMP are stated formally in Appendix B.2. The signal recovery guarantees are slightly different for the Lagrangian Lasso and OMP - for the former, we restate a direct bound [29] on the $\ell_2$-norm of the signal recovery error, which scales as given in the informal equation above. For the latter, our error bound derives from a much stronger variable selection guarantee [30] which implicitly requires the SNR to scale as $\log d$. Thus for both estimators, the qualitative story is the same: there is a tradeoff in how to overparameterize, i.e. how to set $d >> n$. As we increase the number of overparameterizing features, the error arising from fitting noise goes down linearly - but these “fake features" also have the potential to be falsely discovered, thus driving up the signal recovery error rate logarithmically. This logarithmic-linear tradeoff still ensures that the best test error is achieved when we sizeably overparameterize, even if not at $d = \infty$ like in the case of the ideal interpolator. Figure 3 considers an example with iid Gaussian design and true sparsity level $500$ for different noise levels. For all the levels of noise variance considered, we observe that the test MSE closely tracks that of the ideal interpolator’s MSE, perhaps because the values of $d$ plotted were not large enough to start to affect signal recovery.

### 4.3 Sparsity seeking methods run to completion

Note that the hybrid approach described above is not practically viable - the first stage of signal recovery requires knowledge of the noise variance or the sparsity level; moreover with this knowledge we would not bother to fit the noise! A reasonable practical interpolator for the sparse regime might be to use an estimator meant for the sparse linear model to fit the signal as well as noise. For example:

1. Orthogonal matching pursuit (OMP) to completion.

2. Basis pursuit (BP): $\arg \min \| \alpha \|_1$ subject to Equation (1).

**It is important to remember that there is non-zero noise in the problem, i.e. $\sigma > 0$. The above methods are meant to be used in the noiseless setting, and their performance guarantees when they are also used to fit the noise is unknown.**

The results are shown in Figure 3. We observe that OMP run to completion and BP are both reasonably competitive with the ideal interpolator. This result is somewhat surprising; these methods appear to favor sparsity to recover the signal and additionally fit noise harmlessly.

7This same idea inspired recent work on explicitly constructing fake “knockoff" features to draw away energy from features that have the potential to be falsely discovered [31].
5 Conclusions and Future Directions

We analyzed interpolative solutions to overparameterized linear regression in the presence of noise. We showed that

1. There exists an ideal interpolation for which the excess MSE vanishes as the overparameterization increases. This shows that we can fit noise and yet generalize well.

2. We can use algorithms for sparse signal recovery in conjunction with optimal noise-fitting to interpolate with good generalization.

3. The performance of practical interpolators can vary. In addition to not guaranteeing signal recovery, the minimum $\ell_2$-norm interpolator can fit noise harmfully for certain feature sets. The minimum $\ell_1$-norm interpolator is empirically observed to generalize well in the sparse setting even though it also interpolates noise.

The benefit of overparameterization that is observed in real-world examples is very likely solely due to an approximation-theoretic benefit. A precise quantitative characterization of this benefit and conditions under which it arises is naturally desired in future work.

Acknowledgment

We thank Peter L Bartlett for insightful discussions. We would like to acknowledge the staff of EE16A and EE16B at Berkeley for in part inspiring this idea. Last but not least, we acknowledge the support of the ML4Wireless center member companies and NSF grants AST-144078 and ECCS-134398.

References

[1] J. Friedman, T. Hastie, and R. Tibshirani, The elements of statistical learning, vol. 1, no. 10.
[2] V. N. Vapnik, “An overview of statistical learning theory,” IEEE transactions on neural networks, vol. 10, no. 5, pp. 988–999, 1999.
[3] P. L. Bartlett, O. Bousquet, S. Mendelson et al., “Local rademacher complexities,” The Annals of Statistics, vol. 33, no. 4, pp. 1497–1537, 2005.
[4] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, “Understanding deep learning requires rethinking generalization,” arXiv preprint arXiv:1611.03539, 2016.
[5] M. Belkin, A. Rakhlin, and A. B. Tsybakov, “Does data interpolation contradict statistical optimality?” arXiv preprint arXiv:1806.09471, 2018.
[6] M. Belkin, D. Hsu, S. Ma, and S. Mandal, “Reconciling modern machine learning and the bias-variance trade-off,” arXiv preprint arXiv:1812.11118, 2018.
[7] M. Soltanolkotabi, A. Javannard, and J. D. Lee, “Theoretical insights into the optimization landscape of over-parameterized shallow neural networks,” IEEE Transactions on Information Neural Networks, 2018.
[8] Y. Li, T. Ma, and H. Zhang, “Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations,” in Conference On Learning Theory, 2018, pp. 2–47.
[9] Z. Allen-Zhu, Y. Li, and Y. Liang, “Learning and generalization in overparameterized neural networks, going beyond two layers,” arXiv preprint arXiv:1811.04918, 2018.
[10] A. Miller, Subset selection in regression. Chapman and Hall/CRC, 2002.
[11] D. Hsu, S. M. Kakade, and T. Zhang, “Random design analysis of ridge regression,” in Conference on Learning Theory, 2012, pp. 9–1.

[12] Y. C. Pati, R. Rezaifar, and P. S. Krishnaprasad, “Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition,” in Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on. IEEE, 1993, pp. 40–44.

[13] D. L. Donoho, Y. Tsaig, I. Drori, and J.-L. Starck, “Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit,” IEEE Transactions on Information Theory, vol. 58, no. 2, pp. 1094–1121, 2012.

[14] S. van de Geer, “On tight bounds for the lasso,” arXiv preprint arXiv:1804.00989, 2018.

[15] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht, “The marginal value of adaptive gradient methods in machine learning,” in Advances in Neural Information Processing Systems, 2017, pp. 4148–4158.

[16] G. Raskutti, M. J. Wainwright, and B. Yu, “Restricted eigenvalue properties for correlated gaussian designs,” Journal of Machine Learning Research, vol. 11, no. Aug, pp. 2241–2259, 2010.

[17] J. A. Tropp and A. C. Gilbert, “Signal recovery from random measurements via orthogonal matching pursuit,” IEEE Transactions on information theory, vol. 53, no. 12, pp. 4655–4666, 2007.

[18] A. K. Fletcher, S. Rangan, and V. K. Goyal, “Necessary and sufficient conditions for sparsity pattern recovery,” IEEE Transactions on Information Theory, vol. 55, no. 12, pp. 5758–5772, 2009.

[19] A. Rahimi and B. Recht, “Random features for large-scale kernel machines,” in Advances in neural information processing systems, 2008, pp. 1177–1184.

[20] R. Vershynin, “Introduction to the non-asymptotic analysis of random matrices,” arXiv preprint arXiv:1011.3027, 2010.

[21] A. Edelman, “Eigenvalues and condition numbers of random matrices,” SIAM Journal on Matrix Analysis and Applications, vol. 9, no. 4, pp. 543–560, 1988.

[22] S. J. Szarek, “Condition numbers of random matrices,” Journal of Complexity, vol. 7, no. 2, pp. 131–149, 1991.

[23] M. Rudelson and R. Vershynin, “The littlewood–offord problem and invertibility of random matrices,” Advances in Mathematics, vol. 218, no. 2, pp. 600–633, 2008.

[24] V. Shah, A. Kyrillidis, and S. Sanghavi, “Minimum norm solutions do not always generalize well for over-parameterized problems,” arXiv preprint arXiv:1811.07055, 2018.

[25] P. L. Bartlett, “Eecs technical report.”

[26] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” SIAM review, vol. 43, no. 1, pp. 129–159, 2001.

[27] M. J. Wainwright, “Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting,” IEEE Transactions on Information Theory, vol. 55, no. 12, pp. 5728–5741, 2009.

[28] S. Aeron, V. Saligrama, and M. Zhao, “Information theoretic bounds for compressed sensing,” IEEE Transactions on Information Theory, vol. 56, no. 10, pp. 5111–5130, 2010.

[29] P. J. Bickel, Y. Ritov, A. B. Tsybakov et al., “Simultaneous analysis of lasso and dantzig selector,” The Annals of Statistics, vol. 37, no. 4, pp. 1705–1732, 2009.
[30] T. T. Cai and L. Wang, “Orthogonal matching pursuit for sparse signal recovery with noise,” IEEE Transactions on Information theory, vol. 57, no. 7, pp. 4680–4688, 2011.

[31] R. F. Barber, E. J. Candès et al., “Controlling the false discovery rate via knockoffs,” The Annals of Statistics, vol. 43, no. 5, pp. 2055–2085, 2015.

[32] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro, “The implicit bias of gradient descent on separable data,” The Journal of Machine Learning Research, vol. 19, no. 1, pp. 2822–2878, 2018.

[33] C. Wei, J. D. Lee, Q. Liu, and T. Ma, “On the margin theory of feedforward neural networks,” arXiv preprint arXiv:1810.05369, 2018.

[34] T. Liang and A. Rakhlin, “Just interpolate: Kernel" ridgeless" regression can generalize,” arXiv preprint arXiv:1808.00387, 2018.
Appendix A  Related work

Linear regression has been studied for decades in the underdetermined regime \[10\], where error is controlled by the condition number of the design matrix. Relevant concepts for our purposes include ridge regression, which can be used to improve the conditioning of the regressor matrix - elegant analysis of this for a random design is considered in \[11\]. These analyses do not apply directly to the overdetermined regime, because for any positive value of ridge regularization they do not yield interpolative solutions. In fact, signal recovery in the overdetermined regime, even in the absence of noise, is only possible when there is sparsity in the coefficients \[27\] \[28\]. There is rich literature on sparsity-seeking algorithms that are motivated by iterative perspectives from signal processing (orthogonal matching pursuit (OMP) \[12\] and stagewise OMP \[13\]) and convex relaxation (basis pursuit/Lasso \[26\]). In the presence of noise, these algorithms are run until the residual (training error) scales as the noise variance - thus, they do not interpolate. The recovery guarantees on these algorithms are not well-understood if they are run to completion in the presence of noise, as the setting is no longer hard-sparse \[14\].

The motivation for understanding overparameterization comes from the recent empirical observation that deeper neural networks, on which it is possible to train optimization algorithms to reach a global minimum, generalize better\[^8\]. The optimization algorithm used is key: adaptive methods in optimization need not always improve generalization for specially constructed examples in overparameterized linear regression \[15\]. On the other hand, on a different set of examples \[24\] adaptive methods converge to better solutions (which are not minimum \(\ell_2\)-norm) than gradient descent. Differing conclusions for different examples suggests that there is not a “one-size-fits-all” optimization algorithm that converges to desirable interpolative solutions - but even more fundamentally, there may not even exist a universal solution that generalizes well for all overdetermined linear regression problems. Empirical and theoretical attempts have also been made to directly explain the statistical guarantees on global minima, which correspond to interpolative solutions. In logistic regression, a phenomenon of implicit bias in gradient descent \[32\] exists - in particular, the global optimum that is found by gradient descent is the minimum \(\ell_2\)-norm interpolator. Moreover, the minimum \(\ell_2\)-norm solution can correspond to the solution with the maximum margin, and if the data is perfectly separable by a more complex class of models, it is natural that this maximum margin is non-decreasing with an increase in model complexity; this has been observed classically for AdaBoost \[1\] and recently in the context of deep neural networks \[33\]. The existence of special interpolative solutions that generalize well has also been considered for kernel SVM-based classification \[5\].

For regression with squared loss, the margin-based analyses do not apply and there is actually less consensus on the benefits of overparameterization; and particularly the ramifications of overfitting noise. Very pertinent to this work is the analysis of the test MSE of “ridgeless” minimum-Hilbert-norm kernel interpolators; it is shown that these can generalize well under appropriate conditions on the eigenvalues of the kernel as well as the data matrix \[34\]. Most recently, a double-descent curve on the test error as a function of the number of parameters of several parametric models (that presumably approximate non-parametric functions) was displayed on several common datasets \[6\], where the minimum \(\ell_2\)-norm interpolating solution is used. Among these models was the random Fourier feature model which is known to approximate Gaussian kernel machines \[19\]. The experiments showed that the effect of increased parameterization on the test MSE is more nuanced: at the interpolation threshold, fitting noise has an adverse effect on the test MSE; but as the number of parameters (feature vectors) becomes many more than the number of data points, the test MSE decays and approaches the test MSE of the minimum Hilbert norm kernel interpolator. The observed benefit of overparameterization is linked to improved approximability of the underlying non-parametric model, but the quantitative benefits and the conditions under which they are actually realized remain unclear. We show that interpolation of noise is harmless when the features are overparameterized, as well as harmful when they are not - and that both phenomena can be explained through a combination of signal processing and random matrix theory.

\[^8\]Overparameterization in, for e.g., deep neural network training, also has the coincidental benefit of making non-convex optimization landscapes more attractive \[7\] \[9\], in the sense that optimization algorithms like SGD are more likely to actually converge to a global minimum.
Appendix B  Supplemental proofs

B.1 Technical lemmas

Here, we formally state the critical matrix concentration lemmas that are used in the proof of Corollary 1.

Before that, we define a sub-Gaussian random variable and a sub-Gaussian random vector.

Definition 3. A random variable $X$ is sub-Gaussian with parameter at most $K < \infty$ if for all $p \geq 1$, we have

$$p^{-1/2}E[|X|^p]^{1/p} \leq K.$$  

Further, a random vector $X$ is sub-Gaussian with parameter at most $K$ if for every (fixed) vector $v$, the random variable $\langle X, v \rangle \|v\|_2$ is sub-Gaussian with parameter at most $K$.

We cite the following lemma for sub-Gaussian matrix concentration.

Lemma 1 (Theorem 5.58, [20]). Let $B$ be a $d \times n$ matrix with $d \geq n$ whose columns $b_j$ are independent sub-Gaussian isotropic random vectors in $\mathbb{R}^d$ with $\|b_j\|_2 = \sqrt{d}$ almost surely. Then, for every $t > 0$ we have the following inequality:

$$\sqrt{d} - C_K \sqrt{n} - t \leq \sigma_{\min}(B) \leq \sigma_{\max}(B) \leq \sqrt{d} + C_K \sqrt{n} + t$$

with probability at least $1 - 2e^{-c_K t^2}$, where $C_K, c_K$ depend only on the sub-Gaussian parameter $K$ of the columns.

A direct application of the lower tail bound (i.e. the bound on $\sigma_{\min}(B)$) is used in the proof of Corollary 1.

B.2 Proof of Proposition 2

In this section, we provide the proof of Proposition 2. We first provide the proof of Equation (3). Recall that we have assumed $E[a(X)a(X)^\top] = I$. Using the triangle inequality, we have

$$E[\mathcal{E}_{test}(\hat{\alpha}_{hybrid})] = \|\hat{\alpha}_{hybrid} - \alpha^*\|_2^2 + \sigma^2$$

$$\leq \|\hat{\alpha}_1 - \alpha^*\|_2^2 + \|\hat{\alpha}_{hybrid} - \hat{\alpha}_1\|_2^2 + \sigma^2.$$  

From the definition of the estimator $\hat{\alpha}_{hybrid}$ and a similar argument as in the proof of Theorem 1, we have $\hat{\alpha}_{hybrid} - \hat{\alpha}_1 = A_{\text{train}}^\top W_{\text{train}}$ and thus,

$$\|\hat{\alpha}_{hybrid} - \hat{\alpha}_1\|_2^2 \leq \frac{\|W_{\text{train}}\|_2^2}{\lambda_{\min}(A_{\text{train}} A_{\text{train}}^\top)}.$$  

Substituting this into the equation for the test error above completes the proof of Equation (3).

Recall that $\hat{\alpha}_1$ is a suitable estimator that uses the sparsity level $k$ or the noise variance $\sigma^2$ to estimate $\alpha^*$ directly. So $\hat{\alpha}_1$ can be provided by a suitable estimator used for sparse recovery like Lasso or OMP. We provide suitable corollaries for these cases below.

B.2.1 Recovery guarantee for Lasso

We formalize the version of Lasso that we use below.

Definition 4 (Lagrangian Lasso in presence of noise [29]). We define the Lagrangian lasso with regularization parameter $\lambda_n$ (which can in general depend on $\sigma^2$ and $n$) as the estimator that solves the following optimization problem:

$$\hat{\alpha}_{1, \text{Lass}} := \arg\min_{\alpha} \frac{1}{2n} \|Y_{\text{train}} - A_{\text{train}} \alpha\|_2^2 + \lambda_n \|\alpha\|_1.$$  

9Common examples of sub-Gaussian random variables are Gaussian, Bernoulli and all bounded random variables.
For recovery guarantees on the Lagrangian Lasso, in addition to \( k \)-sparsity, we require the following restricted eigenvalue condition on the design matrix:

**Definition 5.** The matrix \( A_{\text{train}} \) satisfies the restricted eigenvalue condition over set \( \text{supp}(\alpha^*) \) with parameters \((\kappa, \alpha)\) if

\[
\frac{1}{n} \| A_{\text{train}} \Delta \|_2^2 \geq \kappa \| \Delta \|_2^2 \quad \text{for all } \Delta \in C_\alpha(\text{supp}(\alpha^*))
\]

where for any subset \( S \), we define

\[
C_\alpha(S) := \{ \Delta \in \mathbb{R}^d : \| \Delta_{S^c} \|_1 \leq \alpha \| \Delta_S \|_1 \}.
\]

It turns out that the iid Gaussian matrix \( A_{\text{train}} \) satisfies this property with high probability as restated in the following lemma.

**Lemma 2.** [16] The iid Gaussian matrix \( A_{\text{train}} \) satisfies the restricted eigenvalue condition \((\kappa, \alpha = 3)\) for some constant \( \kappa > 0 \) with probability greater than or equal to \((1 - 2e^{-cn})\), provided that \( n = \Omega(k \log d) \).

We restate the following upper bound on the error \( \hat{\alpha}_{1,\text{Las}} \) below.

**Theorem 2** ([29]). Let \( A_{\text{train}} \) satisfy the restricted eigenvalue condition with parameters \((\kappa, \alpha = 3)\). Then, any solution of the Lagrangian Lasso with regularization parameter \( \lambda_n \geq 2 \| A_{\text{train}}^\top W_{\text{train}} \|_\infty \) satisfies

\[
\| \hat{\alpha}_{1,\text{Las}} - \alpha^* \|_2^2 \leq \frac{9k \lambda_n^2}{\kappa}.
\]

As a corollary, if the entries of \( A_{\text{train}} \) are column normalized, i.e. \( \max_{j \in [d]} \| a_j \|_2 \leq C' \), then for any \( \delta > 0 \) and regularization parameter choice \( \lambda_n = 2C' \sigma \left( \sqrt{\frac{2 \log d}{n}} + \delta \right) \) we have

\[
\| \hat{\alpha}_{1,\text{Las}} - \alpha^* \|_2^2 \leq \frac{36C'k \sigma^2}{\kappa^2} \left( \frac{2 \log d}{n} + \delta' \right)
\]

with probability greater than or equal to \((1 - 2e^{-n\delta'/2})\).

Putting these two results together, we get the following formal corollary for the Lagrangian Lasso.

**Corollary 2.** Consider the hybrid estimator \( \hat{\alpha}_{\text{hybrid}} \) that uses sparse signal estimate \( \hat{\alpha}_{1,\text{Las}}(A_{\text{train}}, Y_{\text{train}}) \). For every \( d > n \), this estimator has test error

\[
\mathcal{E}_{\text{test}}(\hat{\alpha}_{\text{hybrid}}) \leq \frac{\sigma^2}{(\sqrt{d} - C\sqrt{n})^2} \left( 1 + \frac{C'k \log d}{n} + C'' \delta \right)
\]

\[
\quad + \frac{C'k \log d}{n} + C'' \delta
\]

with probability greater than or equal to \((1 - 2e^{-cn\delta})\), provided that \( n = \Omega(k \log d) \).

**B.2.2 Recovery guarantee for OMP**

The first theoretical guarantees on OMP were proved for noiseless recovery, when OMP is run to completion in the absence of noise [17]. Subsequent work [18] improved the sampling requirement, also in the noiseless setting. Recovery guarantees in the presence of noise were then proved for an appropriate stopping condition [30]. We formalize the version of OMP that we use below.

\[^{10}\text{In fact the original result in the paper shows this property for generally correlated Gaussian design.}\]
For every $S$.

**Corollary 3.** Consider the hybrid estimator $\hat{\alpha}_{\text{hybrid}}$ that uses sparse signal estimate $\hat{\alpha}_{1,\text{OMP}}(A_{\text{train}}, Y_{\text{train}})$. For every $d > n$, this estimator has test error

$$
E_{\text{test}}(\hat{\alpha}_{\text{hybrid}}) \leq \sigma^2 \left( 1 + \frac{n}{(\sqrt{d} - C\sqrt{n})^2} \left( 1 + \frac{k}{n} \right) + \frac{k}{n} \right)
$$

with probability greater than or equal to $\left( 1 - \frac{k}{d^{3/2} \sqrt{\log d}} \right)$ for any $k$-sparse $\alpha^*$ satisfying the SNR condition in Equation (4), and provided that $n = \Omega(k^2 \log d)$. 

**Definition 6** (OMP for recovery in presence of Gaussian noise [30]). Denote that $j^{th}$ column of $A_{\text{train}}$ by $a_j$. The orthogonal matching pursuit algorithm is defined iteratively according to the following steps:

1. (Iteration $t = 1$.) Initialize residual $r_0 = Y_{\text{train}}$ and initialize the set of selected variables $S = \emptyset$.
2. Select variable $s_t := \arg\max_{i} |\langle a_i, r_{t-1} \rangle|$, and add it to $S$.
3. Denote $A_{\text{train}}(S)$ as the submatrix of $A_{\text{train}}$ whose columns are in $S$. Let $P_t = A_{\text{train}}(S)A_{\text{train}}(S)^\dagger$ denote the projection of $Y_{\text{train}}$ onto the linear space spanned by the elements of $A_{\text{train}}(S)$. Update $r_t = (I - P_t)Y_{\text{train}}$.
4. If $\|A_{\text{train}}r_t\|_\infty \leq \sigma \sqrt{2(1 + \eta) \log d}$ for algorithmic parameter $\eta > 0$, stop the algorithm. (Alternatively, we can use the deterministic stopping condition of $k$ steps.) Otherwise, set $t = t + 1$ and go back to Step 2.
5. (Once algorithm has terminated) return $\hat{\alpha}_{1,\text{OMP}}(A_{\text{train}}, Y_{\text{train}}) = \hat{\alpha}_{\text{OLS}}(A_{\text{train}}(S), Y_{\text{train}})$.

We also define pairwise incoherence

$$
\mu(A_{\text{train}}) := \max_{j \neq j'} \frac{|\langle a_j, a_{j'} \rangle|}{\|a_j\|_2 \|a_{j'}\|_2}
$$

In general for recovery guarantees, we would like this quantity to be small. We state the following well-known lemma for the entries of $A_{\text{train}}$ being iid $\mathcal{N}(0, 1)$:

**Lemma 3.** For iid Gaussian $n \times d$ matrix $A_{\text{train}}$, we have $\mu(A_{\text{train}}) < \frac{1}{2k-1}$ with high probability as long as $n = \Omega(k^2 \log d)$.

We now restate the result that is most relevant for our purposes, and holds for any design matrix $A_{\text{train}}$.

**Theorem 3** (Theorem 8 from [30]). Suppose that $\mu < \frac{1}{2k-1}$ and all the non-zero coefficients $\alpha^*_i, i \in \text{supp}(\alpha^*)$ satisfy

$$
|\alpha^*_i| \geq \frac{2\sigma \sqrt{2(1 + \eta) \log d}}{1 - (2k - 1) \mu(A_{\text{train}})}
$$

for some $\eta > 0$. Then, the above version of OMP (with either choice of stopping condition) will recover $S = \text{supp}(\alpha^*)$ with probability greater than equal to $(1 - \frac{k}{d^{3/2} \sqrt{\log d}})$. This directly implies that

$$
\|\hat{\alpha}_1 - \alpha^*\|_2^2 \leq \frac{k\sigma^2}{n}
$$

with probability greater than or equal to $(1 - \frac{k}{d^{3/2} \sqrt{\log d}})$. 

We observe that the obtained guarantees for OMP in the noisy setting are support recovery guarantees, which are much stronger than directly bounding the $\ell_2$-norm of the error. While the expression in Equation (5) does not have an implicit dependence on the dimension $d$, it requires the absolute values of the coefficients $|\alpha^*_i|$ to be bounded above by $\Omega(\sigma \sqrt{\log d})$. This can effectively be thought of as requiring the SNR to be high enough to ensure signal recovery - the more we overparameterized the number of features $d$, the more stringent this condition becomes.

Using Theorem 3 directly, we obtain the following corollary for our hybrid interpolator with OMP used as the sparse signal estimator:

**Corollary 3.** Consider the hybrid estimator $\hat{\alpha}_{\text{hybrid}}$ that uses sparse signal estimate $\hat{\alpha}_{1,\text{OMP}}(A_{\text{train}}, Y_{\text{train}})$. For every $d > n$, this estimator has test error

$$
E_{\text{test}}(\hat{\alpha}_{\text{hybrid}}) \leq \sigma^2 \left( 1 + \frac{n}{(\sqrt{d} - C\sqrt{n})^2} \left( 1 + \frac{k}{n} \right) + \frac{k}{n} \right)
$$

with probability greater than or equal to $\left( 1 - \frac{k}{d^{3/2} \sqrt{\log d}} \right)$ for any $k$-sparse $\alpha^*$ satisfying the SNR condition in Equation (4), and provided that $n = \Omega(k^2 \log d)$. 

16
B.3 Proof of Proposition

In this section, we provide the proof of Proposition for completeness. Recall that

\[ \hat{\alpha}_2 := \arg \min_{\alpha} \| \alpha \|_2 \]
subject to
\[ A_{\text{train}} \alpha = Y_{\text{train}}. \]

An elementary linear algebraic argument gives

\[ \hat{\alpha}_2 := A_{\text{train}}^\dagger Y_{\text{train}} \]
\[ = A_{\text{train}}^\dagger (A_{\text{train}} \alpha^* + W_{\text{train}}) \]

where \( A_{\text{train}}^\dagger \) is the right Moore-Penrose pseudoinverse of \( A_{\text{train}} \).

Thus, we have

\[ \hat{\alpha}_2 - \alpha^* = (A_{\text{train}}^\dagger A_{\text{train}} - I) \alpha^* + A_{\text{train}}^\dagger W_{\text{train}}, \]

and substituting this into the expression we have already derived for test error in the proof of Theorem, we have

\[ \mathcal{E}_{\text{test}}(\hat{\alpha}_2) = \| \Sigma_{1/2}(\hat{\alpha}_2 - \alpha^*) \|_2^2 + \sigma^2 \]
\[ \leq \mathcal{E}_{\text{test}}(\hat{\alpha}_2; \text{noiseless}) + \mathcal{E}_{\text{test}}(\hat{\alpha}_2; \text{noise}) + \sigma^2 \]

where we define

\[ \mathcal{E}_{\text{test}}(\hat{\alpha}_2; \text{noiseless}) := \| \Sigma_{1/2}(A_{\text{train}}^\dagger A_{\text{train}} - I) \alpha^* \|_2^2 \]
\[ \mathcal{E}_{\text{test}}(\hat{\alpha}_2; \text{noise}) := \| \Sigma_{1/2} A_{\text{train}}^\dagger W_{\text{train}} \|_2^2 \]

and we just used the triangle inequality. This completes the proof. \( \Box \)