Overparameterization and generalization error: weighted trigonometric interpolation*

Yuege Xie†, Hung-Hsu Chou‡, Holger Rauhut‡, and Rachel Ward† §

Abstract. Motivated by surprisingly good generalization properties of learned deep neural networks in overparameterized scenarios and by the related double descent phenomenon, this paper analyzes the relation between smoothness and low generalization error in an overparameterized linear learning problem. We study a random Fourier series model, where the task is to estimate the unknown Fourier coefficients from equidistant samples. We derive exact expressions for the generalization error of both plain and weighted least squares estimators. We show precisely how a bias towards smooth interpolants, in the form of weighted trigonometric interpolation, can lead to smaller generalization error in the overparameterized regime compared to the underparameterized regime. This provides insight into the power of overparameterization, which is common in modern machine learning.

Key words. overparameterization, generalization error, weighted optimization, smoothness

AMS subject classifications. 42A15, 65T40

1. Introduction.

Consider the regression/interpolation problem: Given training data $(\mathbf{x}_j, y_j) \in \mathcal{D} \times \mathbb{C}$, $j = 1, \dots, n$, corresponding to samples of an unknown function $y = f(\mathbf{x})$ and the sampling points drawn from $\mathcal{D} \subset \mathbb{R}^d$, we would like to fit the data to a hypothesis class $\mathcal{H} := \{f_{\boldsymbol{\theta}}(\mathbf{x}) : \mathbb{R}^d \to \mathbb{C}, \boldsymbol{\theta} \in \mathbb{C}^p\}$ by solving for parameters minimizing the empirical ℓ_2-risk

$$\boldsymbol{\theta}_{opt} \in \arg\min_{\boldsymbol{\theta} \in \mathbb{C}^p} \sum_{j=1}^{n} |f_{\boldsymbol{\theta}}(\boldsymbol{x}_j) - y_j|^2. \tag{1.1}$$

Traditionally, the number of parameters p is restricted to be smaller than the number of training samples, i.e., $p \leq n$, to avoid overfitting. For $p > n$, the solution $\boldsymbol{\theta}_{opt}$ is often not unique and traditional wisdom says that explicit regularization such as weight decay must be added to ensure that the solution is stable or meaningful. However, such wisdom has been challenged by modern machine learning practice, where small generalization error is achieved with massively *overparameterized* ($p \gg n$) hypothesis classes $\mathcal{H}$ such as deep neural networks, without any explicit regularization. This implies that in such settings, the optimization method used for (1.1) has a favorable *implicit* bias towards a particular choice of $\boldsymbol{\theta}_{opt} \in \mathcal{H}$ among all empirical risk minimizers. As neural networks can be trained with a particularly simple algorithm, (stochastic) gradient descent, a flurry of research in the past several years, starting with [15, 9, 16, 5], has been devoted to answering the question:

*This paper is a modified version of our previous arxiv preprint titled "Weighted optimization: better generalization by smoother interpolation".

Funding: This work was funded by AFOSR 2018 MURI Award, DAAD grant 57417829 and Excellence Initiative of the German federal and state governments.

†Oden Institute, University of Texas at Austin, Austin TX 78712 USA (yuege@oden.utexas.edu).

‡Chair for Mathematics of Information Processing, RWTH Aachen University, Pontdriesch 10, 52056 Aachen, Germany (chou@mathc.rwth-aachen.de, rauhut@mathc.rwth-aachen.de).

§Mathematics Department, University of Texas at Austin, Austin TX 78712 USA (rward@math.utexas.edu).

How and when does the implicit bias of gradient descent interact favorably with the structure of a particular problem to achieve better performance in the interpolation regime?

The papers [4] and [11] were the first to observe that the power of overparameterization is not limited to neural networks, and can even be found in *linear* interpolation models, where the feature basis $\{\psi_k\}_{k=1}^p$ is fixed and the empirical risk is a quadratic function of the parameters: $\|\Psi\boldsymbol{\theta}-\mathbf{y}\|^2 = \sum_{j=1}^n(\sum_{k=1}^p \theta_k\psi_k(\mathbf{x}_j)-y_j)^2$. In this setting, the implicit bias of gradient descent is well understood: by applying (stochastic) gradient descent to the empirical loss (1.1) with initialization belonging to the range of the feature matrix Ψ, the solution converges to the parameter solution $\boldsymbol{\theta}_{\min}$ of minimal ℓ_2-norm among all interpolating solutions.[1]

The seminal work [2] claimed that improvement in generalization error is due to the connection between small ℓ_2-norm of a parameter solution $\boldsymbol{\theta}_{opt}$ and *smoothness* of the corresponding interpolating function $f_{\boldsymbol{\theta}_{opt}}$. This connection was highlighted through the example of linear interpolation with random Fourier features [14] where the features basis $\psi_k : \mathbb{R}^d \to \mathbb{C}$ are random complex exponentials $\psi_k(\mathbf{x}) = e^{i\langle \boldsymbol{w}_k,\mathbf{x}\rangle}$ with $\boldsymbol{w}_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$, and which can be viewed as a class of two-layer neural networks with fixed weights in the first layer. As the number of features $p \to \infty$, this basis converges to that of the reproducing kernel Hilbert space (RKHS) of smooth functions corresponding to the Gaussian kernel, and the interpolating solution by gradient descent converges to the smooth function with minimal RKHS norm.

However, while this connection between the minimal ℓ_2-norm and smoothness of the solution is intriguing, there is no rigorous analysis. In this paper, we initiate this analysis by deriving exact non-asymptotic expressions for the generalization error and corresponding confidence bounds in a random Fourier series model. We show precisely how a bias towards smooth interpolants, in the form of weighted trigonometric interpolation, results in a smaller generalization error in the overparameterized regime compared to the underparameterized regime. We note that in the linear setting, randomness in the model is required (here in the form of random Fourier coefficients), as there may be pathological counter-examples. Our analysis provides insight into the power of overparameterization in modern machine learning.

1.1. Main Contribution and Outline. In this paper, we consider the method of weighted ℓ_2-norm trigonometric interpolation in Fourier feature space, where the weight on a particular feature is proportional to the qth power of its gradient norm to encourage lower-frequency features. This additional degree of freedom, which is possible only in the overparameterized regime, enables us to reduce the risk.

We consider the weighted ℓ_2-norm interpolation in Fourier feature space to equispaced training data $(x_j, f_{\boldsymbol{\theta}}(x_j))$ from functions with sharply decaying Fourier series $\boldsymbol{\theta}$. Our key theoretical results are as follows. In Theorem 3.1 and 4.1, we derive analytic expressions for the risk $\mathbb{E}\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\|_2^2$ both in overparameterized and underparameterized regimes. The expression in overparameterized regime is particularly representative since $\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}$ concentrates well around its expectation, as shown in Theorem 3.8. We then derive more informative upper bounds for the risk in Theorem 3.5. Moreover, Theorem 4.4 states that with sufficient decay, the solution in the overparameterized regime is strictly better than that in the underparameterized regime. We illustrate the trends of empirical risks in Figure 1 and detailed numerical results with

[1] Observe that the gradient descent iterates $\boldsymbol{\theta}_t$ remain in the row span of the feature matrix Ψ if $\boldsymbol{\theta}_0$ is in the row span, and the minimal-norm solution is the unique solution in the row span of the feature matrix.

different settings can be found in Figure 2.

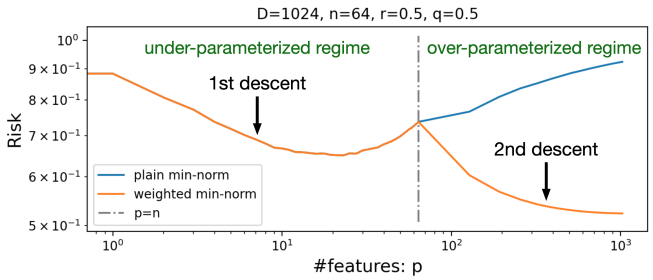

Figure 1. *Above is a demonstration of empirical risks ($\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\|^2$, average of 100 runs) of plain and weighted min-norm (with $q = 0.5$) estimators. Here, $\boldsymbol{\theta}$ is sampled from the distribution in Def. 2.1 for $D = 1024, r = 0.5$.*

In Section 2 we introduce the notation and formulation of the problem. In Section 3 and 4 we analyze the risks in both over- and under-parameterized regime, respectively. We show the numerical results in Section 5 and discuss the importance of randomness in Section 6.

1.2. Previous work on generalization and overparameterization. The work of [2] initiated the study of the extended bias-variance trade-off curve, and showed that double descent behavior is often exhibited, where the risk in the overparameterized regime $p \gg n$ can decrease to a point below the best possible risk in the underparameterized regime. They illustrated that this behavior also occurs in kernel regression/interpolation problems.

Subsequently, several works derived quantitative bounds on the risk in the interpolating setting but required that (a) the features are random (so that random matrix theory can be leveraged) or (b) p and n are in the asymptotic regime and go to infinity at a comparable rate (or both). In contrast, our results are for deterministic Fourier features and hold for any p and n. The precise high-dimensional asymptotic risk for a general random model with correlated covariates was derived in [8]. The work [1] derived sharp bounds on the risk in general linear regression problems with non-isotropic subgaussian covariates and highlighted the importance of selecting features according to higher-variance covariates. Other works include [7, 12, 6, 13].

Prior to the above line of work, [3]—which was a large inspiration for us—considered the discrete Fourier series model we consider, but with a theoretical analysis only for randomly chosen Fourier features, unweighted optimization, and isotropic covariates, in the asymptotic setting. Empirical evidence pointing to improved generalization using weighted optimization and truncated Fourier series instead of random Fourier frequencies was provided, but without theoretical analysis. In this sense, our paper can be viewed as answering an open question regarding the role of weighted optimization in Fourier series interpolation from [3].

We acknowledge a concurrent preprint [10] that also studies generalization error of minimum weighted norm interpolation, but it focuses on providing upper bounds in the asymptotic setting from an approximation theory viewpoint. Compared to it, our work provides exact non-asymptotic expressions and bounds in probability for the generalization error.

2. Formulation. Smooth functions are characterized by the rate of decay in their Fourier series coefficients—the smoother the function, the faster the decay.[2] Drawing inspiration from

[2]The classical Sobolev spaces are Hilbert spaces defined in terms of Fourier series whose coefficients decay sufficiently rapidly. For square-integrable complex-valued functions f on the circle $\mathbb{T}$, consider the space of

this connection, we consider as a model for random smooth periodic functions the class of trigonometric polynomials $f : [-\pi, \pi] \to \mathbb{C}$ with random r-decaying Fourier series coefficients:

Definition 2.1. ***(Random Fourier series with r-decaying coefficients)*** *Fix $D \in \mathbb{N}$ and $r \geq 0$. We say that a function is a random Fourier series with r-decaying coefficients if*[3]

$$f_{\boldsymbol{\theta}}(x) = \sum_{k=0}^{D-1} \theta_k e^{ikx}, \quad 0 \leq x \leq 2\pi, \tag{2.1}$$

where $\boldsymbol{\theta} \in \mathbb{C}^D$ is a random vector satisfying $\mathbb{E}[\boldsymbol{\theta}] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{\theta}\boldsymbol{\theta}^*] = c_r \Sigma^{2r}$, where $\Sigma := diag\left((k+1)^{-1}, k = 0, \dots, D-1\right) \in \mathbb{R}^{D \times D}$ and $c_r = (\sum_{k=0}^{D-1}(k+1)^{-2r})^{-1}$ s.t. $\mathbb{E}[\|\boldsymbol{\theta}\|^2] = 1$.

We observe $n \leq D$ training samples $(x_j, y_j)_{j=0}^{n-1} = (x_j, f_{\boldsymbol{\theta}}(x_j))_{j=1}^{n}$ of such a function $f_{\boldsymbol{\theta}}$ at equispaced points on the domain, $x_j = \frac{2\pi j}{n}$, $j \in [n]$, where $[n]$ denotes the set $\{0, \dots, n-1\}$ for notation simplicity. We can express the observation vector $\mathbf{y} \in \mathbb{C}^n$ concisely as $\mathbf{y} = \boldsymbol{F}\boldsymbol{\theta}$ in terms of the *sample discrete Fourier matrix* $\boldsymbol{F} \in \mathbb{C}^{n \times D}$ whose entries are $(\boldsymbol{F})_{j,k} = e^{ikx_j} = e^{2\pi ijk/n}$, $j \in [n]$, $k \in [D]$. If D is a multiple of n, i.e., $D = \tau n$ for $\tau \in \mathbb{N}$, then we can write $\boldsymbol{F} = [\boldsymbol{F}^{(n)} | \boldsymbol{F}^{(n)} | \cdots | \boldsymbol{F}^{(n)}]$, where $\boldsymbol{F}^{(n)} \in \mathbb{C}^{n \times n}$ is the discrete Fourier matrix in dimension n.

We fit the training samples to a degree-p trigonometric polynomial $f_{\widehat{\boldsymbol{\theta}}}(x) = \sum_{k=0}^{p-1} \widehat{\theta}_k e^{ikx}$ such that $\hat{f}_{\widehat{\boldsymbol{\theta}}}(x_j) \approx f_{\boldsymbol{\theta}}(x_j)$, $j \in [n]$, i.e., $\mathbf{y} \approx \boldsymbol{F}_T \widehat{\boldsymbol{\theta}}_T$ and $\boldsymbol{\theta}_{T^c} = \mathbf{0}$, where $\boldsymbol{F}_T \in \mathbb{C}^{n \times p}$ is the matrix containing the first p columns of $\boldsymbol{F}$, indexed by $T = [p]$. We solve for $\widehat{\boldsymbol{\theta}}_T$ as the least squares fitting vector in the regression regime $p < n$, and as the solution of minimal weighted ℓ_2 norm in the interpolation regime $p > n$:

$$\widehat{\boldsymbol{\theta}}_T = \left\{ \begin{array}{ll} \arg\min_{\boldsymbol{w} \in \mathbb{C}^p} \|\boldsymbol{F}_T \boldsymbol{w} - \boldsymbol{y}\|_2^2; & p \leq n \\ \arg\min_{\boldsymbol{w} \in \mathbb{C}^p} \|\Sigma_T^{-q} \boldsymbol{w}\|_2^2 \quad s.t.\ \boldsymbol{F}_T \mathbf{w} = \mathbf{y} & p > n \end{array} \right\}, \tag{2.2}$$

where $\Sigma_T \in \mathbb{R}^{p \times p}$ is the diagonal matrix as in Def. 2.1 restricted to its first p rows and p columns and $q \geq 0$ controls the rate of growth of the weight. Note that the weight matrix Σ_T^{-q} has no influence on the estimator in the underparameterized regime $p \leq n$. Denoting by $\boldsymbol{A}^\dagger$ the Moore-Penrose pseudo-inverse of a matrix $\boldsymbol{A}$, we can write the solution in both the under- and overparameterized case as $\widehat{\boldsymbol{\theta}}_T = \Sigma_T^q (\boldsymbol{F}_T \Sigma_T^q)^\dagger \mathbf{y}$.

We will derive sharp non-asymptotic expressions for the risk of the estimator $f_{\widehat{\boldsymbol{\theta}}}$ in terms of p, n, D, r, and q. The risk in this setting is defined as

$$\text{risk} = \text{risk}_q = \mathbb{E}_{\boldsymbol{\theta}} \left[\int_{-\pi}^{\pi} |f_{\boldsymbol{\theta}}(x) - f_{\widehat{\boldsymbol{\theta}}}(x)|^2 dx \right] = \mathbb{E}_{\boldsymbol{\theta}} \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2^2, \tag{2.3}$$

where the last equality follows from Parseval's identity.

functions $H^r(\mathbb{T}) = \{f \in L^2(\mathbb{T}) : \|f\|_{r,2}^2\} := \sum_{j=-\infty}^{\infty} (1+|j|^2)^r |\hat{f}(j)|^2 < \infty\}$, $r \in \mathbb{R}, r \geq 0$, where $\hat{f}(j)$ is the jth Fourier series coefficient of f. If $r \in \mathbb{N}$, by duality between differentiation in time and multiplication in frequency, the Sobolev norm is equivalently defined in terms of the rth derivative $f^{(r)}$: $\|f\|_{r,2}^2 = \|f\|_{L_2}^2 + \|f^{(r)}\|_{L_2}^2$.

[3]For ease of exposition we only work with positive frequencies, although it may seem more natural to work with trigonometric polynomials of the form $f(x) = \sum_{k=-D}^{D} \theta_k e^{ikx}$. All our results can be formulated within that setting, by symmetrically extending the weights to negative indices k and by replacing D with $2D-1$, and similarly for n and p. The notation, however, will be more heavy and then make the comparison with other works less straightforward.

3. Risk Analysis for Decaying Fourier Series Model in the Overparameterized Setting. First, we derive non-asymptotic expressions, asymptotic rate, and concentration for the risk via plain and weighted ℓ_2 regression with Fourier series features in the overparameterized regime ($p \geq n$). For ease of exposition, we restrict below to the case where p is an integer multiple of n, but note that the result can be extended to the general case. Moreover, we introduce the notation $t_j = (j+1)^{-1}$, $j \in [D]$, $\Sigma = \mathrm{diag}([t_0, \dots, t_{D-1}])$, and $c_r = 1/\sum_{j=0}^{D-1} t_j^{2r}$.

Theorem 3.1. ***(Risk in overparameterized regime)*** *Assume $D = \tau n$ and $p = ln$ for $\tau \geq l$, $\forall \tau, l \in \mathbb{N}_+ := \{1, 2, \dots\}$ ($p \geq n$). Let the feature vector $\boldsymbol{\theta}$ be drawn from a distribution with $\mathbb{E}[\boldsymbol{\theta}] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{\theta}\boldsymbol{\theta}^*] = c_r \Sigma^{2r}$. Then the risks ($\mathbb{E}[\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^2]$) of the regression coefficients $\hat{\boldsymbol{\theta}}$ fitted by plain min-norm estimator ($q = 0$) and weighted min-norm estimator are*

$$\mathrm{risk}_0 = 1 - \frac{n}{p} + \frac{2n}{p} \cdot c_r \sum_{j=p}^{D-1} t_j^{2r}, \tag{3.1}$$

$$\mathrm{risk}_q = 1 - \underbrace{c_r \sum_{k=0}^{n-1} \frac{\sum_{\nu=0}^{l-1} t_{k+n\nu}^{2q+2r}}{\sum_{\nu=0}^{l-1} t_{k+n\nu}^{2q}}}_{\mathcal{P}_q} + \underbrace{c_r \sum_{k=0}^{n-1} \frac{(\sum_{\nu=0}^{l-1} t_{k+n\nu}^{4q})(\sum_{\nu=l}^{\tau-1} t_{k+n\nu}^{2r})}{(\sum_{\nu=0}^{l-1} t_{k+n\nu}^{2q})^2}}_{\mathcal{Q}_q}. \tag{3.2}$$

Remark 3.2. While the general risk expressions are difficult to parse, special cases are straightforward: if $p = D$, then $\mathrm{risk}_0 = 1 - \frac{n}{D}$ and $\mathrm{risk}_q = 1 - \mathcal{P}_q$ since $\mathcal{Q}_q = 0$. Moreover, if $n = p = D$ ($l = \tau = 1$), then $\mathcal{P}_q = c_r \sum_{n=0}^{D-1} (t_s^{2q+m}/t_s^{2q}) = 1$ so that $\mathrm{risk}_0 = \mathrm{risk}_q = 0$.

Using the expressions in Theorem 3.1, we can quantify how smoothness (as reflected in the rate of decay $r > 0$ in the underlying Fourier series coefficients) can be exploited by setting the weights accordingly ($q = r$) to reduce the risk in the overparameterized setting.

3.1. Proof of Theorem 3.1. The proof is based on the following lemmas. Below, the matrix $\boldsymbol{F}_{T^c} \in \mathbb{C}^{n \times (D-p)}$ is the submatrix of $\boldsymbol{F}$ with the columns in $T^c = [D] \setminus T$.

Lemma 3.3. ***(Risks of estimators in overparameterized regime)*** *Assume $p \geq n$ and let the feature vector $\boldsymbol{\theta} \in \mathbb{C}^D$ be drawn from a distribution with $\mathbb{E}[\boldsymbol{\theta}] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{\theta}\boldsymbol{\theta}^*] = \boldsymbol{K}$, where $\boldsymbol{K}$ is a diagonal matrix. The risk of the weighted min-norm estimator with $q \geq 0$ is $\mathrm{risk}_q = \mathbb{E}[\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^2] = tr(\boldsymbol{K}) - \mathcal{P}_q + \mathcal{Q}_q$, where $\mathcal{P}_q = tr(\boldsymbol{F}_T \Sigma_T^{2q} \boldsymbol{K}_T \boldsymbol{F}_T{}^* (\boldsymbol{F}_T \Sigma_T^{2q} \boldsymbol{F}_T{}^*)^{-1})$ and $\mathcal{Q}_q = tr(\boldsymbol{F}_T \Sigma_T^{4q} \boldsymbol{F}_T{}^* (\boldsymbol{F}_T \Sigma_T^{2q} \boldsymbol{F}_T{}^*)^{-1} \boldsymbol{F}_{T^c} \boldsymbol{K}_{T^c} \boldsymbol{F}_{T^c}{}^* (\boldsymbol{F}_T \Sigma_T^{2q} \boldsymbol{F}_T{}^*)^{-1})$.*

Proof of Lemma 3.3. Using the re-parameterization $\boldsymbol{\beta} = \Sigma^{-q}\boldsymbol{\theta}$, the weighted min-norm estimator is $\hat{\boldsymbol{\beta}}_T := \tilde{\boldsymbol{F}}_T^\dagger \boldsymbol{y}, \hat{\boldsymbol{\beta}}_{T^c} := \mathbf{0}$, where $\boldsymbol{y} = \tilde{\boldsymbol{F}}_T \boldsymbol{\beta}_T + \tilde{\boldsymbol{F}}_{T^c} \boldsymbol{\beta}_{T^c}$ and $\tilde{\boldsymbol{F}} = \boldsymbol{F}\Sigma^q$. Since $\tilde{\boldsymbol{F}}_T$ has full rank, the matrix $\tilde{\boldsymbol{F}}_T \tilde{\boldsymbol{F}}_T^* = \boldsymbol{F}_T \Sigma_T^{2q} \boldsymbol{F}_T{}^*$ is invertible and $\tilde{\boldsymbol{F}}_T^\dagger = \tilde{\boldsymbol{F}}_T^* (\tilde{\boldsymbol{F}}_T \tilde{\boldsymbol{F}}_T^*)^{-1}$. Then,

$$\begin{aligned}
\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2^2 &= \|\Sigma_T^q (\boldsymbol{\beta}_T - \hat{\boldsymbol{\beta}}_T)\|^2 + \|\Sigma_T^q \Sigma_{T^c} (\boldsymbol{\beta}_{T^c} - \hat{\boldsymbol{\beta}}_{T^c})\|^2 \\
&= \|\Sigma_T^q \boldsymbol{\beta}_T - \Sigma_T^q \tilde{\boldsymbol{F}}_T^\dagger (\tilde{\boldsymbol{F}}_T \boldsymbol{\beta}_T + \tilde{\boldsymbol{F}}_{T^c} \boldsymbol{\beta}_{T^c})\|^2 + \|\Sigma_{T^c}^q \boldsymbol{\beta}_{T^c}\|^2 \\
&= \|\Sigma_T^q (\boldsymbol{I} - \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_T) \boldsymbol{\beta}_T - \Sigma_T^q \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_{T^c} \boldsymbol{\beta}_{T^c}\|^2 + \|\Sigma_{T^c}^q \boldsymbol{\beta}_{T^c}\|^2 \\
&= \|\Sigma_T^q (\boldsymbol{I} - \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_T) \boldsymbol{\beta}_T\|^2 + \|\Sigma_T^q \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_{T^c} \boldsymbol{\beta}_{T^c}\|^2 + \|\Sigma_{T^c}^q \boldsymbol{\beta}_{T^c}\|^2 \\
&\quad - \underbrace{2\,\mathrm{Re}(\boldsymbol{\beta}_T^* (\boldsymbol{I} - \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_T) \Sigma_T^{2q} \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_{T^c} \boldsymbol{\beta}_{T^c})}_{=:\mathcal{C}_1}.
\end{aligned} \tag{3.3}$$

Since $\tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_T$ is Hermitian, we have

$$\|\Sigma_T^q\Big(\boldsymbol{I} - \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_T\Big)\boldsymbol{\beta}_T\|^2 = \left\|\Sigma_T^q \boldsymbol{\beta}_T\right\|^2 + \|\Sigma_T^q \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_T \boldsymbol{\beta}_T\|^2 - \underbrace{2(\boldsymbol{\beta}_T^* \Sigma_T^{2q} \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_T \boldsymbol{\beta}_T)}_{=:\mathcal{C}_2}. \tag{3.4}$$

Combining (3.3) and (3.4) and taking expectation yields

$$\mathbb{E}[\|\Sigma^q(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|^2] = \mathbb{E}[\|\Sigma^q \boldsymbol{\beta}\|^2] + \mathbb{E}[\|\Sigma_T^q \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_T \boldsymbol{\beta}_T\|^2] + \mathbb{E}[\|\Sigma_T^q \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_{T^c} \boldsymbol{\beta}_{T^c}\|^2] - \mathbb{E}[\mathcal{C}_1] - \mathbb{E}[\mathcal{C}_2]. \tag{3.5}$$

The "trace trick" and $\tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_T = \tilde{\boldsymbol{F}}_T^*(\tilde{\boldsymbol{F}}_T \tilde{\boldsymbol{F}}_T^*)^{-1} \tilde{\boldsymbol{F}}_T$ give

$$\begin{aligned}\mathbb{E}[\|\Sigma_T^q \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_T \boldsymbol{\beta}_T\|^] &= \mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{\beta}_T^* \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_T \Sigma_T^{2q} \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_T \boldsymbol{\beta}_T\right)\right] = \operatorname{tr}\left(\tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_T \Sigma_T^{2q} \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_T \mathbb{E}[\boldsymbol{\beta}_T \boldsymbol{\beta}_T^*]\right) \\ &= \operatorname{tr}\left(\Sigma_T^{2q} \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_T \Sigma_T^{-q} \boldsymbol{K}_T \Sigma_T^{-q}\right) = \operatorname{tr}\left(\tilde{\boldsymbol{F}}_T \boldsymbol{K}_T \tilde{\boldsymbol{F}}_T^* (\tilde{\boldsymbol{F}}_T \tilde{\boldsymbol{F}}_T^*)^{-1}\right) = \operatorname{tr}\left(K_T \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_T\right).\end{aligned}$$

Moreover,

$$\begin{aligned}\mathbb{E}[\|\Sigma_T^q \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_{T^c} \boldsymbol{\beta}_{T^c}\|^2] &= \operatorname{tr}\left(\tilde{\boldsymbol{F}}_{T^c}^* (\tilde{\boldsymbol{F}}_T^\dagger)^* \Sigma_T^{2q} \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_{T^c} \mathbb{E}[\boldsymbol{\beta}_{T^c} \boldsymbol{\beta}_{T^c}^*]\right) \\ &= \operatorname{tr}\left(\Sigma_T^{2q} \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_{T^c} \Sigma_{T^c}^{-q} \boldsymbol{K}_{T^c} \Sigma_{T^c}^{-q}\right) \tilde{\boldsymbol{F}}_{T^c}^* (\tilde{\boldsymbol{F}}_T^\dagger)^* \\ &= \operatorname{tr}\left(\Sigma_T^{2q} \tilde{\boldsymbol{F}}_T^* (\tilde{\boldsymbol{F}}_T \tilde{\boldsymbol{F}}_T^*)^{-1} \boldsymbol{F}_{T^c} \boldsymbol{K}_{T^c} \boldsymbol{F}_{T^c}{}^* (\tilde{\boldsymbol{F}}_T \tilde{\boldsymbol{F}}_T^*)^{-1} \tilde{\boldsymbol{F}}_T\right).\end{aligned}$$

Since $\mathbb{E}[\boldsymbol{\beta}_{T^c} \boldsymbol{\beta}_T^*] = \Sigma_{T^c}^{-q} \mathbb{E}[\boldsymbol{\theta}_{T^c} \boldsymbol{\theta}_T^*] \Sigma_T^{-q} = 0$ we have $\mathbb{E}[\mathcal{C}_1] = 0$. Furthermore, since $\boldsymbol{K}$ commutes with Σ^{-q} by diagonality, we have $\Sigma_T^{2q} \mathbb{E}\left[\boldsymbol{\beta}_T \boldsymbol{\beta}_T^*\right] = \Sigma_T^{2q} \mathbb{E}[\Sigma^{-q} \boldsymbol{\theta} \boldsymbol{\theta}^* \Sigma^{-q}] = \boldsymbol{K}$ so that

$$\mathbb{E}[\mathcal{C}_2] = 2 \operatorname{tr}\left(\tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_T \Sigma_T^{2q} \mathbb{E}\left[\boldsymbol{\beta}_T \boldsymbol{\beta}_T^*\right]\right) = 2 \operatorname{tr}\left(\boldsymbol{K}_T \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_T\right).$$

Plugging all terms into (3.5), we have

$$\begin{aligned}\operatorname{risk}_q = \mathbb{E}[\|\Sigma^q(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|^2] &= \operatorname{tr}(\boldsymbol{K}) - \operatorname{tr}\left(\boldsymbol{F}_T \Sigma_T^{2q} \boldsymbol{K}_T \boldsymbol{F}_T{}^* (\boldsymbol{F}_T \Sigma_T^{2q} \boldsymbol{F}_T{}^*)^{-1}\right) \\ &\quad + \operatorname{tr}\left(\boldsymbol{F}_T \Sigma_T^{4q} \boldsymbol{F}_T{}^* (\boldsymbol{F}_T \Sigma_T^{2q} \boldsymbol{F}_T{}^*)^{-1} \boldsymbol{F}_{T^c} \boldsymbol{K}_{T^c} \boldsymbol{F}_{T^c}{}^* (\boldsymbol{F}_T \Sigma_T^{2q} \boldsymbol{F}_T{}^*)^{-1}\right).\end{aligned}$$

The risk of the plain min-norm estimator corresponds to $q = 0$, which gives

$$\operatorname{risk}_0 = \mathbb{E}[\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^2] = \operatorname{tr}(\boldsymbol{K}) - \operatorname{tr}\left(\boldsymbol{F}_T \boldsymbol{K}_T \boldsymbol{F}_T{}^* (\boldsymbol{F}_T \boldsymbol{F}_T{}^*)^{-1}\right) + \operatorname{tr}\left(\boldsymbol{F}_{T^c} \boldsymbol{K}_{T^c} \boldsymbol{F}_{T^c}{}^* (\boldsymbol{F}_T \boldsymbol{F}_T{}^*)^{-1}\right). \blacksquare$$

Lemma 3.4. ***(Properties of $\boldsymbol{F}_T$)*** *Assume that $D = \tau n$ and $p = nl$ for $\tau, l \in \mathbb{N}_+$. Then, $\boldsymbol{F}_T \boldsymbol{F}_T{}^* = p \boldsymbol{I}_n$. For $u \in \mathbb{N}_+$, define $\boldsymbol{A}_u := \boldsymbol{F}_T \Sigma_T^u \boldsymbol{F}_T{}^*$ and $\boldsymbol{C}_u := \boldsymbol{F}_{T^c} \Sigma_{T^c}^u \boldsymbol{F}_{T^c}{}^*$, where Σ is a diagonal matrix (e.g., the diagonal matrix in Def. 2.1). Then, $\boldsymbol{A}_u$ and $\boldsymbol{C}_u$ are circulant matrices.*

Proof of Lemma 3.4. For $u \geq 0$, we set $\boldsymbol{A}_u = \boldsymbol{F}_T \Sigma_T^u \boldsymbol{F}_T{}^*$ and $\boldsymbol{C}_u = \boldsymbol{F}_{T^c} \Sigma_{T^c}^u \boldsymbol{F}_{T^c}{}^*$, and define $\omega_n = \exp(-\frac{2\pi i}{n})$. Since $p = nl$, we have, for $j_1, j_2 \in [n]$,

$$(\boldsymbol{A}_u)_{j_1,j_2} = (\boldsymbol{F}_T \Sigma_T^u \boldsymbol{F}_T{}^*)_{j_1,j_2} = \sum_{k=0}^{p-1} t_k^u \exp\left(\frac{-2\pi i}{n}(j_2 - j_1)\cdot k\right) = \sum_{\nu=0}^{l-1}\sum_{k=0}^{n-1} t_{k+n\nu}^u \omega_n^{(j_2-j_1)k},$$
$$(\boldsymbol{C}_u)_{j_1,j_2} = (\boldsymbol{F}_{T^c} \Sigma_{T^c}^u \boldsymbol{F}_{T^c}{}^*)_{j_1,j_2} = \sum_{k=p}^{D-1} t_k^u \exp\left(\frac{-2\pi i}{n}(j_2 - j_1)\cdot k\right) = \sum_{\nu=l}^{\tau-1}\sum_{k=0}^{n-1} t_{k+n\nu}^u \omega_n^{(j_2-j_1)k}.$$

In the above equations, we use $\omega_n^{k+n\nu} = \omega_n^k$ for $\nu \in \mathbb{N}_+$.

For $j \in [n]$, let $a_j = \sum_{\nu=0}^{l-1}\sum_{k=0}^{n-1} t_{k+n\nu}^u \omega_n^{-jk}$ and $c_j = \sum_{\nu=l}^{\tau-1}\sum_{k=0}^{n-1} t_{k+n\nu}^u \omega_n^{-jk}$. Then

$$\begin{aligned}(\boldsymbol{A}_u)_{j_1,j_2} &= \sum_{\nu=0}^{l-1}\sum_{k=0}^{n-1} t_{k+n\nu}^u \omega_n^{(j_2-j_1)k} = a_{j_2-j_1} \pmod n,\\ (\boldsymbol{C}_u)_{j_1,j_2} &= \sum_{\nu=l}^{\tau-1}\sum_{k=0}^{n-1} t_{k+n\nu}^u \omega_n^{(j_2-j_1)k} = c_{j_2-j_1} \pmod n.\end{aligned} \tag{3.6}$$

Hence, for any $u \geq 0$, $\boldsymbol{A}_u$ and $\boldsymbol{C}_u$ are circulant matrices.

For $u = 0$ we use again $p = nl, l \in \mathbb{N}_+$ to obtain that, for $j_1, j_2 \in [n]$,

$$(\boldsymbol{F}_T \boldsymbol{F}_T{}^*)_{j_1,j_2} = \sum_{k=0}^{p-1} \omega_n^{(j_2-j_1)k} = \begin{cases} p, & \text{if} \quad j_1 = j_2,\\ 0, & \text{if} \quad j_1 \neq j_2.\end{cases}$$

Hence, $\boldsymbol{F}_T \boldsymbol{F}_T{}^* = p\boldsymbol{I}_n$ as claimed. ∎

Proof of the risk of the plain min-norm estimator. Lemma 3.3 with $\boldsymbol{K} = c_r \Sigma^{2r}$ gives

$$\text{risk}_0 = 1 - c_r \text{tr}\left(\boldsymbol{F}_T \Sigma_T^{2r} \boldsymbol{F}_T{}^* (\boldsymbol{F}_T \boldsymbol{F}_T{}^*)^{-1}\right) + c_r \text{tr}\left(\boldsymbol{F}_{T^c} \boldsymbol{K}_{T^c} \boldsymbol{F}_{T^c}{}^* (\boldsymbol{F}_T \boldsymbol{F}_T{}^*)^{-1}\right). \tag{3.7}$$

By Lemma 3.4, we have $\boldsymbol{F}_T \boldsymbol{F}_T{}^* = p\boldsymbol{I}_p$, so that

$$\mathcal{P}_0 = \frac{1}{p}\text{tr}\left(c_r \boldsymbol{F}_T \Sigma_T^{2r} \boldsymbol{F}_T{}^*\right) = \frac{c_r}{p}\text{tr}\left(\boldsymbol{A}_{2r}\right); \mathcal{Q}_0 = \frac{1}{p}\text{tr}\left(c_r \boldsymbol{F}_{T^c} \Sigma_{T^c}^{2r} \boldsymbol{F}_{T^c}{}^*\right) = \frac{c_r}{p}\text{tr}\left(\boldsymbol{C}_{2r}\right). \tag{3.8}$$

The diagonal entries of $\boldsymbol{A}_{2r}$ and $\boldsymbol{C}_{2r}$ are given by

$$\boldsymbol{A}_{2r}^{(i,i)} = \sum_{j=0}^{p-1} t_j^{2r} \exp(0) = \sum_{j=0}^{p-1} t_j^{2r} \quad \text{and} \quad \boldsymbol{C}_{2r}^{(i,i)} = \sum_{j=0}^{D-1} t_j^{2r} \exp(0) = \sum_{j=p}^{D-1} t_j^{2r}, \quad i \in [n]. \tag{3.9}$$

It follows that $\mathcal{P}_0 = \frac{c_r}{p}\text{tr}\left(\boldsymbol{A}_{2r}\right) = \frac{nc_r}{p}\sum_{j=0}^{p-1} t_j^{2r}$ and $\mathcal{Q}_0 = \frac{c_r}{p}\text{tr}\left(\boldsymbol{C}_{2r}\right) = \frac{nc_r}{p}\sum_{j=p}^{D-1} t_j^{2r}$ so that the risk is given by

$$\text{risk}_0 = \mathbb{E}\left[\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^2\right] = 1 - \frac{nc_r}{p}\left(\frac{1}{c_r} - 2\sum_{j=p}^{D-1} t_j^{2r}\right) = 1 - \frac{n}{p} + \frac{2n}{p}\cdot\frac{\sum_{j=p}^{D-1} t_j^{2r}}{\sum_{j=0}^{D-1} t_j^{2r}}.$$

∎

Proof of the risk of the weighted min-norm estimator. Since $\boldsymbol{A}_u$ and $\boldsymbol{C}_u$ are circulant matrices, we can write $\boldsymbol{A}_u = \boldsymbol{U}_n \Lambda_{a,u} \boldsymbol{U}_n^*$ and $\boldsymbol{C}_u = \boldsymbol{U}_n \Lambda_{c,u} \boldsymbol{U}_n^*$, where $\boldsymbol{U}_n$ is the unitary discrete Fourier matrix of size n, and $\Lambda_{a,u}$ and $\Lambda_{c,u}$ are diagonal matrices with eigenvalues of $\boldsymbol{A}_u$ and $\boldsymbol{C}_u$ on the diagonal, respectively. The eigenvalues can be calculated by taking discrete Fourier transform of the first column of $\boldsymbol{A}_u$ or $\boldsymbol{C}_u$.

Let $\boldsymbol{F}_n$ be the nth order discrete Fourier matrix, i.e., $(\boldsymbol{F}_n)_{s,j} = \omega_n^{sj}$, then for any $s \in [n]$, the sth diagonal element (eigenvalue) of $\Lambda_{a,u}$ or $\Lambda_{c,u}$ is

$$\lambda_{a,u}^{(s)} = \boldsymbol{F}_n[s,:]\boldsymbol{A}_u[:,0] = \sum_{j=0}^{n-1} \omega_n^{sj} \left(\sum_{\nu=0}^{l-1} \sum_{k=0}^{n-1} t_{k+n\nu}^u \omega_n^{-jk} \right) = \sum_{k=0}^{n-1} \left(\sum_{\nu=0}^{l-1} t_{k+n\nu}^u \right) \left(\sum_{j=0}^{n-1} \omega_n^{(s-k)j} \right)$$

$$\lambda_{c,u}^{(s)} = \boldsymbol{F}_n[s,:]\boldsymbol{C}_u[:,0] = \sum_{j=0}^{n-1} \omega_n^{sj} \left(\sum_{\nu=l}^{\tau-1} \sum_{k=0}^{n-1} t_{k+n\nu}^u \omega_n^{-jk} \right) = \sum_{k=0}^{n-1} \left(\sum_{\nu=l}^{\tau-1} t_{k+n\nu}^u \right) \left(\sum_{j=0}^{n-1} \omega_n^{(s-k)j} \right)$$

For $s, k \in [n]$ we define

$$e_{s,k}^{(n)} := \sum_{j=0}^{n-1} \omega_n^{(s-k)j} = \begin{cases} n, & \text{if } k = s, \\ 0, & \text{otherwise.} \end{cases} \tag{3.10}$$

If the random Fourier series has r-decaying coefficients, i.e., $\boldsymbol{K} = c_r \Sigma^{2r}$ $(r \geq 0)$, then by Lemma 3.3,

$$\begin{aligned} \mathcal{P}_q &= c_r \mathrm{tr} \left(\boldsymbol{F}_T \Sigma_T^{2q+2r} \boldsymbol{F}_T{}^* (\boldsymbol{F}_T \Sigma_T^{2q} \boldsymbol{F}_T{}^*)^{-1} \right) = c_r \mathrm{tr} \left(\boldsymbol{U}_n \Lambda_{a,2q+2r} \boldsymbol{U}_n^* \boldsymbol{U}_n \Lambda_{a,2q}^{-1} \boldsymbol{U}_n^* \right) \\ &= c_r \mathrm{tr} \left(\Lambda_{a,2q+2r} \Lambda_{a,2q}^{-1} \right) = c_r \sum_{s=0}^{n-1} \frac{\sum_{k=0}^{n-1} \left(\sum_{\nu=0}^{l-1} t_{k+n\nu}^{2q+2r} \right) e_{s,k}^{(n)}}{\sum_{k=0}^{n-1} \left(\sum_{\nu=0}^{l-1} t_{k+n\nu}^{2q} \right) e_{s,k}^{(n)}} = \frac{1}{\sum_{j=0}^{D-1} t_j^{2r}} \sum_{k=0}^{n-1} \frac{\sum_{\nu=0}^{l-1} t_{k+n\nu}^{2q+2r}}{\sum_{\nu=0}^{l-1} t_{k+n\nu}^{2q}}, \end{aligned}$$

and

$$\begin{aligned} \mathcal{Q}_q &= c_r \mathrm{tr} \left(\boldsymbol{F}_{T^c} \Sigma_{T^c}^{2r} \boldsymbol{F}_{T^c}{}^* (\boldsymbol{F}_T \Sigma_T^{2q} \boldsymbol{F}_T{}^*)^{-1} \boldsymbol{F}_T \Sigma_T^{4q} \boldsymbol{F}_T{}^* (\boldsymbol{F}_T \Sigma_T^{2q} \boldsymbol{F}_T{}^*)^{-1} \right) \\ &= c_r \mathrm{tr} \left(\boldsymbol{U}_n \Lambda_{c,2r} \boldsymbol{U}_n^* \boldsymbol{U}_n \Lambda_{a,2q}^{-1} \boldsymbol{U}_n^* \boldsymbol{U}_n \Lambda_{a,4q} \boldsymbol{U}_n^* \boldsymbol{U}_n \Lambda_{a,2q}^{-1} \boldsymbol{U}_n^* \right) = c_r \mathrm{tr} \left(\Lambda_{c,2r} \Lambda_{a,2q}^{-1} \Lambda_{a,4q} \Lambda_{a,2q}^{-1} \right) \\ &= c_r \sum_{s=0}^{n-1} \frac{\left(\sum_{k=0}^{n-1} \left(\sum_{\nu=0}^{l-1} t_{k+n\nu}^{4q} \right) e_{s,k}^{(n)} \right) \left(\sum_{k=0}^{n-1} \left(\sum_{\nu=l}^{\tau-1} t_{k+n\nu}^{2r} \right) e_{s,k}^{(n)} \right)}{\left(\sum_{k=0}^{n-1} \left(\sum_{\nu=0}^{l-1} t_{k+n\nu}^{2q} \right) e_{s,k}^{(n)} \right)^2} \\ &= c_r \sum_{k=0}^{n-1} \frac{\left(\sum_{\nu=0}^{l-1} t_{k+n\nu}^{4q} \right) \left(\sum_{\nu=l}^{\tau-1} t_{k+n\nu}^{2r} \right)}{\left(\sum_{\nu=0}^{l-1} t_{k+n\nu}^{2q} \right)^2}. \end{aligned}$$

Therefore, the risk satisfies

$$\mathrm{risk}_q = 1 - \mathcal{P}_q + \mathcal{Q}_q = 1 - c_r \sum_{k=0}^{n-1} \frac{\sum_{\nu=0}^{l-1} t_{k+n\nu}^{2q+2r}}{\sum_{\nu=0}^{l-1} t_{k+n\nu}^{2q}} + c_r \sum_{k=0}^{n-1} \frac{\left(\sum_{\nu=0}^{l-1} t_{k+n\nu}^{4q} \right) \left(\sum_{\nu=l}^{\tau-1} t_{k+n\nu}^{2r} \right)}{\left(\sum_{\nu=0}^{l-1} t_{k+n\nu}^{2q} \right)^2}. \qquad \blacksquare$$

3.2. Asymptotic Rate of Weighted Min-norm Risk. In this section, we will derive an informative upper bound of equation 3.2 in Theorem 3.5 that demonstrates the asymptotic behavior of the risk and its relation to the parameters n, p, and r.

Theorem 3.5. ***(Asymptotic rate of weighted min-norm risk)*** *In the overparameterized setting of Theorem 3.1, if $q = r > 1/2$ and $p = nl$ with $l \in \mathbb{N}_+, l \geq 2$, then the risk of weighted optimization satisfies*

$$\text{risk}_q \leq an^{-2r+1} + bn^{-2r}p^{-2r+1}$$

with $a = \frac{2+d_r n^{-2r}}{(1+d_r n^{-2r})(1-(D+1)^{-2r+1})}$, $b = \frac{d_r}{(1+d_r n^{-2r})(1-(D+1)^{-2r+1})}$, *and* $d_r = \frac{2^{-2r+1}-(l+1)^{-2r+1}}{2r-1}$.

Remark 3.6. For sufficiently large D and n, the constants in the above theorem satisfy $a \leq 2$ and $b \leq 2^{-2r+1}/(2r-1)$ so that then

$$\text{risk}_q \leq 2n^{-2r+1} + \frac{2}{2r-1}(2n)^{-2r}p^{-2r+1}.$$

To prove Theorem 3.5, we start with a lemma that provides an explicit bound for the summation, and in particular c_r.

Lemma 3.7. ***(Bounds for the summation)*** *For $n_1 < n_2$ and $\alpha > 1$,*

$$\sum_{n=n_1}^{n_2}(an+b)^{-\alpha} \geq \frac{1}{a(\alpha-1)}((an_1+b)^{-\alpha+1} - (a(n_2+1)+b)^{-\alpha+1})$$
$$\sum_{n=n_1}^{n_2}(an+b)^{-\alpha} \leq (an_1+b)^{-\alpha} + \frac{1}{a(\alpha-1)}((an_1+b)^{-\alpha+1} - (an_2+b)^{-\alpha+1}).$$

Consequently, for $r > 1/2$, the constant $c_r = (\sum_{j=0}^{D-1}(j+1)^{-2r})^{-1}$ satisfies

$$\frac{2r-1}{2r-D^{-2r+1}} \leq c_r \leq \frac{2r-1}{1-(D+1)^{-2r+1}}.$$

Proof of Lemma 3.7. By comparison of the sum to an integral, we have that

$$\sum_{n=n_1}^{n_2}(an+b)^{-\alpha} \geq \int_{n_1}^{n_2+1}(ax+b)^{-\alpha}dx = \frac{1}{a(\alpha-1)}((an_1+b)^{-\alpha+1} - (a(n_2+1)+b)^{-\alpha+1})$$
$$\sum_{n=n_1}^{n_2}(an+b)^{-\alpha} = (an_1+b)^{-\alpha} + \sum_{n=n_1+1}^{n_2}(an+b)^{-\alpha} \leq (an_1+b)^{-\alpha} + \int_{n_1}^{n_2}(ax+b)^{-\alpha}dx$$
$$= (an_1+b)^{-\alpha} + \frac{1}{a(\alpha-1)}((an_1+b)^{-\alpha+1} - (an_2+b)^{-\alpha+1}).$$

In particular, for $\alpha = 2r$, $a = 1$, $b = 0$, $n_1 = 1$, $n_2 = D$,

$$c_r^{-1} \geq \frac{1}{2r-1}(1-(D+1)^{-2r+1})$$
$$c_r^{-1} \leq 1 + \frac{1}{2r-1}(1-D^{-2r+1}) = \frac{1}{2r-1}(2r-D^{-2r+1}).$$

■

Proof of Theorem 3.5. For $k \in [n]$ and $\alpha \in \mathbb{R}$, we define

$$A(k,\alpha) := \sum_{\nu=0}^{l-1} t_{k+n\nu}^{\alpha} \quad \text{and} \quad B(k,\alpha) = \sum_{\nu=1}^{l-1} t_{k+n\nu}^{\alpha} = A(k,\alpha) - \frac{1}{(1+k)^{\alpha}},$$

where we understand that $B(k,\alpha) = 0$ if $l = 1$. By Theorem 3.1 we can write

$$\begin{aligned} 1 - \mathcal{P}_q &= c_r \left(c_r^{-1} - \sum_{k=0}^{n-1} \frac{A(k,2q+2r)}{A(k,2q)} \right) \\ &= c_r \sum_{k=0}^{n-1} \underbrace{\left(\frac{1}{(1+k)^{2r}} - \frac{A(k,2q+2r)}{A(k,2q)} \right)}_{=:\gamma_k} + c_r \sum_{k=n}^{D-1} \frac{1}{(1+k)^{2r}}. \end{aligned}$$

We have

$$\begin{aligned} \gamma_{k-1} &= \frac{1}{k^{2r}} - \frac{\frac{1}{k^{2q+2r}} + B(k-1,2q+2r)}{\frac{1}{k^{2q}} + B(k-1,2q)} = \frac{1}{k^{2r}} - \frac{1 + k^{2q+2r} B(k-1,2q+2r)}{k^{2r} + k^{2r+2q} B(k-1,2q)} \\ &= \frac{1 + k^{2q} B(k-1,2q) - 1 - k^{2q+2r} B(k-1,2q+2r)}{k^{2r}(1 + k^{2q} B(k-1,2q))} \\ &= \frac{k^{2q} B(k-1,2q) - k^{2q+2r} B(k-1,2q+2r))}{k^{2r}(1 + k^{2q} B(k-1,2q))}. \end{aligned}$$

Furthermore, if $l = 1$ (i.e., $p = n$) then the numerator in the last expression vanishes and for $l > 1$ it satisfies

$$\begin{aligned} k^{2q} B(k-1,2q) - k^{2q+2r} B(k-1,2q+2r) &= \sum_{\nu=1}^{l-1} \left(\frac{k}{k+n\nu} \right)^{2q} - \sum_{\nu=1}^{l-1} \left(\frac{k}{k+n\nu} \right)^{2q+2r} \\ = \sum_{\nu=1}^{l-1} \left(\frac{k}{k+n\nu} \right)^{2q} \left(1 - \left(\frac{k}{k+n\nu} \right)^{2r} \right). \end{aligned}$$

Altogether, we have that $1 - \mathcal{P}_q = c_r \sum_{k=n+1}^{D} k^{-2r}$ if $l = 1$ and for $l > 1$ it holds

$$1 - \mathcal{P}_q = c_r \sum_{\nu=1}^{l-1} \sum_{k=1}^{n} \frac{k^{2q-2r}}{1 + k^{2q} B(k-1,2q)} \frac{1}{(k+n\nu)^{2q}} \left(1 - \left(\frac{k}{k+n\nu} \right)^{2r} \right) + c_r \sum_{k=n+1}^{D} \frac{1}{k^{2r}}. \tag{3.11}$$

For $r > 1/2$, the last term can be upper bounded by $c_r (2r-1)^{-1} n^{-2r+1}$ according to lemma 3.7. Similarly,

$$B(k-1,2q) \geq \frac{1}{n(2q-1)} \left((k+n)^{-2q+1} - (k+\ell n)^{-2q+1} \right),$$

we obtain the lower bound
where we use the fact that $q > 1/2$. Since the last expression is decreasing with $k \in \{1,\dots,n\}$,

$$B(k-1,2q) \geq \frac{1}{n(2q-1)} \left((2n)^{-2q+1} - ((l+1)n)^{-2q+1} \right) = \frac{d_q}{n^{2q}},$$

where $d_q := \frac{2^{-2q+1}-(l+1)^{-2q+1}}{2q-1}$. Hence, we have

$$1-\mathcal{P}_q \le \frac{c_r}{1+d_q n^{-2q}} \sum_{\nu=1}^{l-1}\sum_{k=1}^{n} \frac{k^{2q-2r}}{(k+n\nu)^{2q}} \left(1-\left(\frac{k}{k+n\nu}\right)^{2r}\right) + \frac{c_r}{2r-1} n^{-2r+1}$$
$$\le \frac{c_r}{1+d_q n^{-2q}} \sum_{\nu=1}^{l-1}\sum_{k=1}^{n} \frac{k^{2q-2r}}{(k+n\nu)^{2q}} + \frac{c_r}{2r-1} n^{-2r+1}.$$

If $q=r$ then the double sum above can be estimated as

$$\sum_{\nu=1}^{l-1}\sum_{k=1}^{n} \frac{1}{(k+n\nu)^{2q}} = \sum_{j=n+1}^{p} \frac{1}{j^{2q}} \le \int_n^p \frac{1}{x^{2q}} dx = \frac{1}{2q-1}\left(n^{-2q+1}-p^{-2q+1}\right).$$

Altogether, for $r=q>1/2$ and $p=ln$ for $l\ge 2$,

$$1-\mathcal{P}_q \le \frac{c_r}{2r-1}\left(\frac{n^{-2r+1}-p^{-2r+1}}{1+d_r n^{-2r}} + n^{-2r+1}\right)$$
$$\le \frac{1}{1-(D+1)^{-2r+1}}\left(\frac{n^{-2r+1}-p^{-2r+1}}{1+d_r n^{-2r}} + n^{-2r+1}\right),$$

where we have used Lemma 3.7 in the last step.

It remains to bound $\mathcal{Q}_q$ from above. Towards this goal, we observe the simple inequality $\sum_{\nu=0}^{l-1} t_{k+n\nu}^{4q} \le \left(\sum_{\nu=0}^{l-1} t_{k+n\nu}^{2q}\right)^2$. Thus by Lemma 3.7, we have the immediate bound

$$\mathcal{Q}_q \le c_r \sum_{k=0}^{n-1}\sum_{\nu=l}^{\tau-1} t_{k+n\nu}^{2r} = c_r \sum_{k=p+1}^{D} \frac{1}{k^{2r}} \le \frac{p^{-2r+1}}{1-(D+1)^{-2r+1}}.$$

Altogether, for $r=q>1/2$ and $p=ln$ with $l\ge 2$,

$$\begin{aligned}\mathrm{risk}_q &= 1-\mathcal{P}_q+\mathcal{Q}_q \\ &\le \frac{1}{1-(D+1)^{-2r+1}}\left(\left(\frac{1}{1+d_r n^{-2r}}+1\right)n^{-2r+1} + \left(1-\frac{1}{1+d_r n^{-2r}}\right)p^{-2r+1}\right) \\ &= \frac{1}{1-(D+1)^{-2r+1}}\left(\frac{2+d_r n^{-2r}}{1+d_r n^{-2r}} n^{-2r+1} + \frac{d_r}{1+d_r n^{-2r}} n^{-2r} p^{-2r+1}\right). \qquad (3.12)\end{aligned}$$

The previous expression can be bounded by Cn^{-2r+1} for a suitable constant C.

If $l=1$ so that $p=n$ then the above derivations give

$$\mathrm{risk}_q = 1-\mathcal{P}_q+\mathcal{Q}_q \le \frac{1}{1-(D+1)^{-2r+1}}\left(n^{-2r+1}+p^{-2r+1}\right) = \frac{2}{1-(D+1)^{-2r+1}} n^{-2r+1},$$

which gives the statement of the theorem also in this case. ∎

3.3. Concentration of Error. In this section, we will justify that equation 3.2 is indeed representative due to the concentration of the error $\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}$, i.e., it is close to its expectation with high probability.

Theorem 3.8. ***(Probabilistic Bound)*** *In the overparameterized setting of Theorem 3.1, if $r \geq q \geq \frac{1}{2}$ and $\boldsymbol{\theta}$ has independent sub-Gaussian coordinates with $\|\boldsymbol{\theta}_k\|_{\psi_2} = \sqrt{c_r} k^{-r}$, where $\|X\|_{\psi_2} := \inf\{t > 0 : \mathbb{E}(exp(X^2/t^2)) > 2\}$ is the sub-Gaussian norm. For any $t > 0$,*

$$\mathbb{P}\left(\left|\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2^2 - \mathbb{E}\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2^2\right| \leq t\right) \geq 1 - 2exp\left[-\min\left(\frac{t^2}{T^2}, \frac{t}{T}\right)\right], \tag{3.13}$$

where $T = 4(2r-1)\sqrt{\frac{q(24q^2-17q+3)}{(2q-1)^2(4q-1)}}$.

Proof of Theorem 3.8. From Lemma 3.3, we have

$$\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2^2 = \left\|\Sigma_T^q \left(\boldsymbol{I} - \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_T\right) \boldsymbol{\beta}_T\right\|^2 + \left\|\Sigma_T^q \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_{T^c} \boldsymbol{\beta}_{T^c}\right\|^2 + \left\|\Sigma_{T^c}^q \boldsymbol{\beta}_{T^c}\right\|^2 - \mathcal{C}_1, \tag{3.14}$$

where $\mathcal{C}_1 = 2\,\mathrm{Re}\left(\boldsymbol{\beta}_T^* \left(\boldsymbol{I} - \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_T\right) \Sigma_T^{2q} \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_{T^c} \boldsymbol{\beta}_{T^c}\right)$. Equation (3.14) can be expressed as

$$\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2^2 = \begin{bmatrix}\boldsymbol{\beta}_T^* & \boldsymbol{\beta}_{T^c}^*\end{bmatrix} \begin{bmatrix} D_1 & B \\ B^* & D_2 \end{bmatrix} \begin{bmatrix}\boldsymbol{\beta}_T \\ \boldsymbol{\beta}_{T^c}\end{bmatrix} = \boldsymbol{\beta}^* M \boldsymbol{\beta},$$

where $D_1 = \left(\boldsymbol{I} - \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_T\right) \Sigma_T^{2q} \left(\boldsymbol{I} - \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_T\right)$, $D_2 = \Sigma_{T^c}^{2q} + \tilde{\boldsymbol{F}}_{T^c}^* (\tilde{\boldsymbol{F}}_T^\dagger)^* \Sigma_T^{2q} \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_{T^c}$, and $B = \left(\boldsymbol{I} - \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_T\right) \Sigma_T^{2q} \tilde{\boldsymbol{F}}_T^\dagger \tilde{\boldsymbol{F}}_{T^c}$. Since $\boldsymbol{\theta}$ (and hence $\boldsymbol{\beta}$) has independent sub-Gaussian coordinates, the Hanson-Wright inequality gives

$$\mathbb{P}(|\boldsymbol{\beta}^* M \boldsymbol{\beta} - \mathbb{E}\boldsymbol{\beta}^* M \boldsymbol{\beta}| \geq t) \leq 2\exp\left[-c \min\left(\frac{t^2}{K^4 \|M\|_F^2}, \frac{t}{K^2 \|M\|}\right)\right],$$

where $K = \max_k \|\boldsymbol{\beta}_k\|_{\psi_2} = \sqrt{c_r}$ because $r \geq q$. Since $\|M\|_2 \leq \|M\|_F$, it suffices to bound

$$\|M\|_F^2 = \mathrm{tr}(D_1^* D_1 + 2B^* B + D_2^* D_2).$$

Since $\tilde{\boldsymbol{F}}_T$ has full rank, $\tilde{\boldsymbol{F}}_T^\dagger = \tilde{\boldsymbol{F}}_T^* (\tilde{\boldsymbol{F}}_T \tilde{\boldsymbol{F}}_T^*)^{-1} = \Sigma_T^q \boldsymbol{F}_T^* (\boldsymbol{F}_T \Sigma_T^{2q} \boldsymbol{F}_T^*)^{-1}$. By the circulant property from Lemma 3.4, we have

$$\begin{aligned}
\mathrm{tr}(\tilde{\boldsymbol{F}}_T \Sigma_T^u \tilde{\boldsymbol{F}}_T^\dagger) &= \mathrm{tr}(\boldsymbol{F}_T \Sigma_T^q \Sigma_T^u \Sigma_T^q \boldsymbol{F}_T^* (\boldsymbol{F}_T \Sigma_T^{2q} \boldsymbol{F}_T^*)^{-1}) = \mathrm{tr}(\Lambda_{a,2q+u} \Lambda_{a,2q}^{-1}) \\
\mathrm{tr}((\tilde{\boldsymbol{F}}_T^\dagger)^* \Sigma_T^u \tilde{\boldsymbol{F}}_T^\dagger) &= \mathrm{tr}(((\boldsymbol{F}_T \Sigma_T^{2q} \boldsymbol{F}_T^*)^{-1})^* \boldsymbol{F}_T \Sigma_T^q \Sigma_T^u \Sigma_T^q \boldsymbol{F}_T^* (\boldsymbol{F}_T \Sigma_T^{2q} \boldsymbol{F}_T^*)^{-1}) = \mathrm{tr}(\Lambda_{a,2q+u} \Lambda_{a,2q}^{-2}).
\end{aligned}$$

An explicit calculation yields

$$\begin{aligned}
\operatorname{tr}(D_1^* D_1) &= \operatorname{tr}\left((\boldsymbol{I}-\tilde{\boldsymbol{F}}_T^\dagger\tilde{\boldsymbol{F}}_T)\Sigma_T^{2q}(\boldsymbol{I}-\tilde{\boldsymbol{F}}_T^\dagger\tilde{\boldsymbol{F}}_T)^2\Sigma_T^{2q}(\boldsymbol{I}-\tilde{\boldsymbol{F}}_T^\dagger\tilde{\boldsymbol{F}}_T)\right)\\
&= \operatorname{tr}\left(\left[\Sigma_T^{2q}(\boldsymbol{I}-\tilde{\boldsymbol{F}}_T^\dagger\tilde{\boldsymbol{F}}_T)^2\right]^2\right) = \operatorname{tr}\left(\left[\Sigma_T^{2q}(\boldsymbol{I}-\tilde{\boldsymbol{F}}_T^\dagger\tilde{\boldsymbol{F}}_T)\right]^2\right)\\
&= \operatorname{tr}\left(\Sigma_T^{4q}-2\tilde{\boldsymbol{F}}_T\Sigma_T^{4q}\tilde{\boldsymbol{F}}_T^\dagger+(\tilde{\boldsymbol{F}}_T\Sigma_T^{2q}\tilde{\boldsymbol{F}}_T^\dagger)^2\right) = \operatorname{tr}\left(\Sigma_T^{4q}-2\Lambda_{a,6q}\Lambda_{a,2q}^{-1}+(\Lambda_{a,4q}\Lambda_{a,2q}^{-1})^2\right);\\
\operatorname{tr}(D_2^* D_2) &= \operatorname{tr}\left(\Sigma_{T^c}^{4q}+2\Sigma_{T^c}^{2q}\tilde{\boldsymbol{F}}_{T^c}^*(\tilde{\boldsymbol{F}}_T^\dagger)^*\Sigma_T^{2q}\tilde{\boldsymbol{F}}_T^\dagger\tilde{\boldsymbol{F}}_{T^c}+\tilde{\boldsymbol{F}}_{T^c}^*(\tilde{\boldsymbol{F}}_T^\dagger)^*\Sigma_T^{2q}\tilde{\boldsymbol{F}}_T^\dagger\tilde{\boldsymbol{F}}_{T^c}\tilde{\boldsymbol{F}}_{T^c}^*(\tilde{\boldsymbol{F}}_T^\dagger)^*\Sigma_T^{2q}\tilde{\boldsymbol{F}}_T^\dagger\tilde{\boldsymbol{F}}_{T^c}\right)\\
&= \operatorname{tr}\left(\Sigma_{T^c}^{4q}+2\tilde{\boldsymbol{F}}_{T^c}\Sigma_{T^c}^{2q}\tilde{\boldsymbol{F}}_{T^c}^*(\tilde{\boldsymbol{F}}_T^\dagger)^*\Sigma_T^{2q}\tilde{\boldsymbol{F}}_T^\dagger+(\tilde{\boldsymbol{F}}_{T^c}\tilde{\boldsymbol{F}}_{T^c}^*(\tilde{\boldsymbol{F}}_T^\dagger)^*\Sigma_T^{2q}\tilde{\boldsymbol{F}}_T^\dagger)^2\right)\\
&= \operatorname{tr}\left(\Sigma_{T^c}^{4q}+2\Lambda_{c,4q}\Lambda_{a,4q}\Lambda_{a,2q}^{-2}+(\Lambda_{c,2q}\Lambda_{a,4q}\Lambda_{a,2q}^{-2})^2\right);\\
\operatorname{tr}(B^*B) &= \operatorname{tr}\left(\tilde{\boldsymbol{F}}_{T^c}^*(\tilde{\boldsymbol{F}}_T^\dagger)^*\Sigma_T^{2q}(\boldsymbol{I}-\tilde{\boldsymbol{F}}_T^\dagger\tilde{\boldsymbol{F}}_T)^2\Sigma_T^{2q}\tilde{\boldsymbol{F}}_T^\dagger\tilde{\boldsymbol{F}}_{T^c}\right)\\
&= \operatorname{tr}\left(\tilde{\boldsymbol{F}}_{T^c}\tilde{\boldsymbol{F}}_{T^c}^*(\tilde{\boldsymbol{F}}_T^\dagger)^*\Sigma_T^{2q}(\boldsymbol{I}-\tilde{\boldsymbol{F}}_T^\dagger\tilde{\boldsymbol{F}}_T)\Sigma_T^{2q}\tilde{\boldsymbol{F}}_T^\dagger\right)\\
&= \operatorname{tr}\left(\tilde{\boldsymbol{F}}_{T^c}\tilde{\boldsymbol{F}}_{T^c}^*(\tilde{\boldsymbol{F}}_T^\dagger)^*\Sigma_T^{4q}\tilde{\boldsymbol{F}}_T^\dagger-\tilde{\boldsymbol{F}}_{T^c}\tilde{\boldsymbol{F}}_{T^c}^*(\tilde{\boldsymbol{F}}_T^\dagger)^*\Sigma_T^{2q}\tilde{\boldsymbol{F}}_T^\dagger\tilde{\boldsymbol{F}}_T\Sigma_T^{2q}\tilde{\boldsymbol{F}}_T^\dagger\right)\\
&= \operatorname{tr}\left(\Lambda_{c,2q}\Lambda_{a,6q}\Lambda_{a,2q}^{-2}-\Lambda_{c,2q}\Lambda_{a,4q}\Lambda_{a,2q}^{-2}\Lambda_{a,4q}\Lambda_{a,2q}^{-1}\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|M\|_F^2 &= \operatorname{tr}(D_1^*D_1+D_2^*D_2+BB^*+B^*B)\\
&= \operatorname{tr}\left(\Sigma^{4q}+(\Lambda_{a,4q}^2+2\Lambda_{c,4q}\Lambda_{a,4q}+2\Lambda_{c,2q}\Lambda_{a,6q})\Lambda_{a,2q}^{-2}+\Lambda_{c,2q}^2\Lambda_{a,4q}^2\Lambda_{a,2q}^{-4}\right)\\
&\quad-\operatorname{tr}\left(2\Lambda_{a,6q}\Lambda_{a,2q}^{-1}+2\Lambda_{c,2q}\Lambda_{a,4q}^2\Lambda_{a,2q}^{-3}\right).
\end{aligned}\tag{3.15}$$

We will now bound this expression using the information on Λ. Note that $\lambda_{a,q}^{(s)}$, the s-th diagonal entry of $\Lambda_{a,q}$, follows the inequality $\lambda_{a,q_1+q_2}^{(s)} \le n^{-1}\lambda_{a,q_1}^{(s)}\lambda_{a,q_2}^{(s)}$ for any $q_1, q_2 > 0$. Hence,

$$\begin{aligned}
\operatorname{tr}\left(\Lambda_{a,4q}^2\Lambda_{a,2q}^{-2}\right) &\le n^{-1}\operatorname{tr}\left(\Lambda_{a,4q}\Lambda_{a,2q}^2\Lambda_{a,2q}^{-2}\right) = n^{-1}\operatorname{tr}(\Lambda_{a,4q}) = \operatorname{tr}(\Sigma_T^{4q})\\
\operatorname{tr}\left(\Lambda_{c,4q}\Lambda_{a,4q}\Lambda_{a,2q}^{-2}\right) &\le n^{-1}\operatorname{tr}\left(\Lambda_{c,4q}\Lambda_{a,2q}^2\Lambda_{a,2q}^{-2}\right) = n^{-1}\operatorname{tr}(\Lambda_{c,4q}) = \operatorname{tr}(\Sigma_{T^c}^{4q})\\
\operatorname{tr}\left(\Lambda_{c,2q}\Lambda_{a,6q}\Lambda_{a,2q}^{-2}\right) &\le n^{-2}\operatorname{tr}\left(\Lambda_{c,2q}\Lambda_{a,2q}^3\Lambda_{a,2q}^{-2}\right) = n^{-2}\operatorname{tr}(\Lambda_{c,2q}\Lambda_{a,2q}) \le \operatorname{tr}(\Sigma_T^{2q})\operatorname{tr}(\Sigma_{T^c}^{2q})\\
\operatorname{tr}\left(\Lambda_{c,2q}^2\Lambda_{a,4q}^2\Lambda_{a,2q}^{-4}\right) &\le n^{-2}\operatorname{tr}\left(\Lambda_{c,2q}^2\Lambda_{a,2q}^4\Lambda_{a,2q}^{-4}\right) = n^{-2}\operatorname{tr}(\Lambda_{c,2q}^2) \le \operatorname{tr}(\Sigma_{T^c}^{2q})^2
\end{aligned}$$

Lemma 3.7 implies that for $\alpha > 1$,

$$\frac{1}{2(\alpha-1)} \le \operatorname{tr}(\Sigma^\alpha) \le \frac{\alpha}{\alpha-1}, \quad K^2 \le 2(2r-1),$$

and hence,

$$\begin{aligned}\|M\|_F^2 &\le \operatorname{tr}\left(\Sigma^{4q}+\Sigma_T^{4q}+2\Sigma_{T^c}^{4q}\right)+2\operatorname{tr}(\Sigma_T^{2q})\operatorname{tr}(\Sigma_{T^c}^{2q})+\operatorname{tr}(\Sigma_{T^c}^{2q})^2\\ &\le \frac{12q}{4q-1}+\frac{16q^2}{(4q-1)(2q-1)}+\frac{4q^2}{(2q-1)^2}=\frac{4q(24q^2-17q+3)}{(2q-1)^2(4q-1)}.\end{aligned}$$

The conclusion then follows by plugging all the bounds into the Hanson-Wright inequality. ■

4. Risk Analysis in the Underparameterized Setting and Benefits of Overparameterization. In order to fully understand the benefit of overparameterization, we derive the non-asymptotic risk for the estimators in the underparameterized regime ($p \le n$), where q does not have an influence on the estimators and, hence, on the risk.

Theorem 4.1. ***(Risk in underparameterized regime)*** *Suppose $D = \tau n$ for $\tau \in \mathbb{N}_+$. Suppose $p \le n$, and assume that the feature vector $\boldsymbol{\theta}$ is drawn from a distribution with $\mathbb{E}[\boldsymbol{\theta}] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{\theta}\boldsymbol{\theta}^*] = c_r\Sigma^{2r}$. Then the risk ($\mathbb{E}[\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\|^2]$) is given by*

$$\text{risk}_{under} = c_r\left(\sum_{j=p}^{D-1} t_j^{2r}+\sum_{k=1}^{\tau-1}\sum_{j=0}^{p-1} t_{kn+j}^{2r}\right). \tag{4.1}$$

Remark 4.2. When $r=0$, $\text{risk}_{under} = \frac{1}{D}(D-p+(\tau-1)p) = 1+p(\frac{1}{n}-\frac{2}{D})$ and the risk increases with p until $p=n$, provided $n < D/2$. From Figure 2, as we vary r in the range $0 \le r \le 1$, this behavior persists for a while, then changes to a U-shape curve, and lastly to a decreasing curve. For $r \ge 1$, we prove that the risk is monotonically decreasing in p.

The proof of Theorem 4.1 is based on the following lemma.

Lemma 4.3. ***(Risks of weighted and plain min-norm estimator in the underparameterized regime)*** *Let the feature vector $\boldsymbol{\theta}$ be sampled from a distribution with $\mathbb{E}[\boldsymbol{\theta}] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{\theta}\boldsymbol{\theta}^*] = \boldsymbol{K}$. In the underparameterized regime ($p \le n$), the regression coefficients $\hat{\boldsymbol{\theta}}$ are fitted by weighted least squares with Σ^q as the re-parameterization matrix, then for any $q \ge 0$, $\hat{\boldsymbol{\theta}}_T = (\boldsymbol{F}_T{}^*\boldsymbol{F}_T)^{-1}\boldsymbol{F}_T{}^*\boldsymbol{y}, \hat{\boldsymbol{\theta}}_{T^c} = \mathbf{0}$, where $\boldsymbol{y} = \boldsymbol{F}_T\boldsymbol{\theta}_T + \boldsymbol{F}_{T^c}\boldsymbol{\theta}_{T^c}$. The risk is given by*

$$\text{risk}_{under} = \mathbb{E}[\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\|^2] = tr(\boldsymbol{K}_{T_c}) + tr\left(\boldsymbol{F}_T(\boldsymbol{F}_T{}^*\boldsymbol{F}_T)^{-2}\boldsymbol{F}_T{}^*\boldsymbol{F}_{T^c}\boldsymbol{K}_{T_c}\boldsymbol{F}_{T^c}{}^*\right).$$

Proof of Lemma 4.3. In the under-parameterized setting, the error of the least squares estimator satisfies

$$\begin{aligned}\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\|^2 &= \left\|(\boldsymbol{F}_T{}^*\boldsymbol{F}_T)^{-1}\boldsymbol{F}_T{}^*(\boldsymbol{F}_T\boldsymbol{x}_T+\boldsymbol{F}_{T^c}\boldsymbol{x}_{T^c})-\boldsymbol{x}_T\right\|^2+\|\boldsymbol{x}_{T^c}\|^2\\ &= \left\|(\boldsymbol{F}_T{}^*\boldsymbol{F}_T)^{-1}\boldsymbol{F}_T{}^*\boldsymbol{F}_{T^c}\boldsymbol{x}_{T^c}\right\|^2+\|\boldsymbol{x}_{T^c}\|^2\\ &= \operatorname{tr}\left(\boldsymbol{F}_{T^c}{}^*\boldsymbol{F}_T(\boldsymbol{F}_T{}^*\boldsymbol{F}_T)^{-2}\boldsymbol{F}_T{}^*\boldsymbol{F}_{T^c}\boldsymbol{x}_{T^c}\boldsymbol{x}_{T^c}^*\right)+\|\boldsymbol{x}_{T^c}\|^2.\end{aligned}$$

Taking expectation yields

$$\mathbb{E}\left[\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\|^2\right] = c_r\operatorname{tr}\left(\Sigma_{T^c}^{2r}\right)+c_r\operatorname{tr}\left(\boldsymbol{F}_T(\boldsymbol{F}_T{}^*\boldsymbol{F}_T)^{-2}\boldsymbol{F}_T{}^*\boldsymbol{F}_{T^c}\Sigma_{T^c}^{2r}\boldsymbol{F}_{T^c}{}^*\right).$$

■

Proof of Theorem 4.1. [of Theorem 4.1] Then, denoting $\omega_n = \exp(-\frac{2\pi i}{n})$ we have, for $k_1, k_2 \in [p]$ with $p < n$,

$$(\boldsymbol{F}_T{}^*\boldsymbol{F}_T)_{k_1,k_2} = \sum_{j=0}^{n-1} \exp\left(-\frac{2\pi i}{n}k_1 \cdot j\right) \exp\left(\frac{2\pi i}{n}k_2 \cdot j\right) = \sum_{j=0}^{n-1} \omega_n^{(k_1-k_2)\cdot j} = \begin{cases} n, & \text{if } k_1 = k_2, \\ 0, & \text{otherwise.} \end{cases}$$

Moreover, for $0 \le k_1, k_2 \le D - p - 1$, we have

$$\begin{aligned}(\boldsymbol{F}_{T^c}{}^*\boldsymbol{F}_{T^c})_{k_1,k_2} &= \sum_{j=0}^{n-1} \exp\left(-\frac{2\pi i}{n}(k_1+p) \cdot j\right) \exp\left(\frac{2\pi i}{n}(k_2+p) \cdot j\right) = \sum_{j=0}^{n-1} \omega_n^{(k_1-k_2)\cdot j} \\ &= \begin{cases} n, & \text{if } \quad \exists\gamma \in \mathbb{N}, \ s.t. \ \ k_1 - k_2 = \gamma n, \\ 0, & \text{otherwise.} \end{cases}\end{aligned}$$

Since $D = \tau n$ and $0 < p < n$ it holds $v = D - p - n \cdot \lfloor \frac{D-p}{n} \rfloor = n - p$ and $p = n - v$. Introducing the matrices $\boldsymbol{M}_{n\times v} = \begin{bmatrix} \boldsymbol{I}_{v\times v} \\ \boldsymbol{O}_{p\times v} \end{bmatrix}$, $\boldsymbol{N}_{v\times n} = \begin{bmatrix} \boldsymbol{I}_{v\times v} & \boldsymbol{O}_{v\times p} \end{bmatrix}$, $\boldsymbol{I}_{n,v} = \boldsymbol{M}_{n\times v}\boldsymbol{N}_{v\times n} = \begin{bmatrix} \boldsymbol{I}_{v\times v} & \boldsymbol{O}_{v\times p} \\ \boldsymbol{O}_{p\times v} & \boldsymbol{O}_{p\times p} \end{bmatrix}$, we can write

$$\begin{aligned}(\boldsymbol{F}_{T^c}{}^*\boldsymbol{F}_{T^c})^2 &= n^2 \begin{bmatrix} \boldsymbol{I}_n & \cdots & \boldsymbol{I}_n & \boldsymbol{M}_{n\times v} \\ \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{I}_n & \cdots & \boldsymbol{I}_n & \boldsymbol{M}_{n\times v} \\ \boldsymbol{N}_{v\times n} & \cdots & \boldsymbol{N}_{v\times n} & \boldsymbol{I}_{v\times v} \end{bmatrix}^2 \\ &= n^2 \begin{bmatrix} (\tau-1)\boldsymbol{I}_n + \boldsymbol{I}_{n,v} & \cdots & (\tau-1)\boldsymbol{I}_n + \boldsymbol{I}_{n,v} & \tau\boldsymbol{M}_{n\times v} \\ \vdots & \ddots & \vdots & \vdots \\ (\tau-1)\rfloor\boldsymbol{I}_n + \boldsymbol{I}_{n,v} & \cdots & (\tau-1)\boldsymbol{I}_n + \boldsymbol{I}_{n,v} & \tau\boldsymbol{M}_{n\times v} \\ \tau\boldsymbol{N}_{v\times n} & \cdots & \tau\boldsymbol{N}_{v\times n} & \tau\boldsymbol{I}_{v\times v} \end{bmatrix} \triangleq n^2\boldsymbol{L}.\end{aligned}$$

Since $\boldsymbol{F}_T\boldsymbol{F}_T{}^* + \boldsymbol{F}_{T^c}\boldsymbol{F}_{T^c}{}^* = D\boldsymbol{I}_n$ the risk is given as

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^2] &= c_r \mathrm{tr}\left(\Sigma_{T^c}^{2r}\right) + \frac{c_r}{n^2}\mathrm{tr}\left((D\boldsymbol{I}_n - \boldsymbol{F}_{T^c}\boldsymbol{F}_{T^c}{}^*)\boldsymbol{F}_{T^c}\Sigma_{T^c}^{2r}\boldsymbol{F}_{T^c}{}^*\right) \\ &= c_r \mathrm{tr}\left(\Sigma_{T^c}^{2r}\right) + \frac{c_r}{n^2}\mathrm{tr}\left(\left(D\boldsymbol{F}_{T^c}{}^*\boldsymbol{F}_{T^c} - (\boldsymbol{F}_{T^c}{}^*\boldsymbol{F}_{T^c})^2\right)\Sigma_{T^c}^{2r}\right) \\ &= c_r(1+\tau)\mathrm{tr}\left(\Sigma_{T^c}^{2r}\right) - c_r\mathrm{tr}\left(\boldsymbol{L}\Sigma_{T^c}^{2r}\right) \\ &= c_r(1+\tau)\mathrm{tr}\left(\Sigma_{T^c}^{2r}\right) - c_r\tau\mathrm{tr}\left(\Sigma_{T^c}^{2r}\right) + c_r\sum_{k=1}^{\tau-1}\sum_{j=0}^{p-1} t_{kn+j}^{2r} \\ &= c_r\sum_{j=p}^{D-1} t_j^{2r} + c_r\sum_{k=1}^{\tau-1}\sum_{j=0}^{p-1} t_{kn+j}^{2r} = \frac{\sum_{j=p}^{D-1} t_j^{2r} + \sum_{k=1}^{\tau-1}\sum_{j=0}^{p-1} t_{kn+j}^{2r}}{\sum_{j=0}^{D-1} t_j^{2r}}.\end{aligned}$$

■

Finally, we show that the "second descent" of the weighted min-norm estimator in the overparameterized regime achieves a lower risk than in the underparameterized regime, provided $q \ge r \ge 1$. In other words, it is where "over- is better than under-parameterization".

Theorem 4.4. ***(Lowest risk)*** *In the setting of Theorems 3.1 and 4.1, if $q \geq r \geq 1$, then*
(a) In the underparameterized regime ($p \leq n$), the risk is monotonically decreasing in p and the lowest risk in this regime is $\text{risk}^*_{under} = 2c_r \sum_{j=n}^{D-1} t_j^{2r}$.
(b) The lowest risk in the overparameterized regime ($p > n$) is strictly less than the lowest possible risk in the underparameterized regime.

Remark 4.5. While the above theorem holds for any q satisfying $q \geq r \geq 1$, our experiments suggest that $q = r$ is an appropriate choice for any $r \geq 0$, corresponding to the case where the assumed smoothness q employed in the weighted optimization matches the true underlying smoothness r. For a range of choices for r and q, the plots of the theoretical extended risk curves (fixed n, varying p) can be found in the appendix.

Proof of Theorem 4.4. (a) We show that, for fixed n and $r \geq 1$, the risk in the underparameterized setting is monotonically decreasing in p. To this end, we set $f(p) = \sum_{j=p}^{D-1} t_j^{2r} + \sum_{k=1}^{\tau-1} \sum_{j=0}^{p-1} t_{kn+j}^{2r}$, where $t_j = (j+1)^{-1}$. Let $\Delta(p) = f(p+1) - f(p)$ be the increment. Then $\Delta(p) = -t_p^{2r} + \sum_{k=1}^{\tau-1} t_{kn+p}^{2r}$. The goal is to show $\Delta(p) \leq 0$ for all $p \in [n]$. Since $2r > 1$, by Lemma 3.7

$$\Delta(p) \leq -(p+1)^{-2r} + \frac{1}{n(2r-1)}[(p+1)^{-2r+1} - (n(\tau-1)+p+1)^{-2r+1}] =: \Delta^+(p).$$

Hence, it suffices to show that $\Delta^+(p) \leq 0$ for $p \in [n]$. Its derivative w.r.t. p is

$$\begin{aligned}\Delta^{+\prime}(p) &= 2r(p+1)^{-2r-1} + \frac{1}{n}[(n(\tau-1)+p+1)^{-2r} - (p+1)^{-2r}] \\ &= (p+1)^{-2r}\left(2r(p+1)^{-1} - \frac{1}{n}\right) + \frac{1}{n}(n(\tau-1)+p+1)^{-2r} > 0\end{aligned}$$

since $2nr > p+1$ and hence all the terms are positive. Because Δ^+ is increasing, it suffices to check if the end point, $\Delta^+(n-1)$, is non-positive in order to ensure $\Delta^+ \leq 0$. Indeed,

$$\begin{aligned}\Delta^+(n-1) &= -n^{-2r} + \frac{1}{n(1-2r)}[(n\tau)^{1-2r} - n^{1-2r}] = -n^{-2r} + \frac{n^{-2r}}{2r-1}[1-\tau^{1-2r}] \\ &= -n^{-2r}\left(1 - \frac{1}{2r-1}\right) - \frac{n^{-2r}\tau^{1-2r}}{2r-1} < 0\end{aligned}$$

because $2r-1 > 1$. Therefore, the lowest risk in the underparameterized regime is $\text{risk}^*_{under} = 2c_r \sum_{j=n}^{D-1} t_j^{2r}$.

(b) Next, we consider two cases in the overparameterized regime: $p = n$ and $p = D$.
When $p = n$ (i.e., $l = 1$), the risk can be written as

$$\text{risk}_q(p=n) = 1 - c_r \sum_{k=0}^{n-1} t_k^{2r} + c_r \sum_{k=0}^{n-1} \sum_{\nu=l}^{\tau-1} t_{k+n\nu}^{2r} = 1 - c_r \sum_{k=0}^{n-1} \underbrace{\left(t_k^{2r} - \sum_{\nu=1}^{\tau-1} t_{k+n\nu}^{2r}\right)}_{=:b_k}.$$

When $p = D$ (i.e., $l = \tau$), we have $\mathcal{Q}_q = 0$ so that

$$\mathrm{risk}_q(p = D) = 1 - c_r \sum_{k=0}^{n-1} \frac{\sum_{\nu=0}^{l-1} t_{k+n\nu}^{2q+2r}}{\sum_{\nu=0}^{l-1} t_{k+n\nu}^{2q}} = 1 - c_r \sum_{k=0}^{n-1} \underbrace{\frac{\sum_{\nu=0}^{\tau-1} t_k^{2q+2r}}{t_k^{2q} + \sum_{\nu=1}^{\tau-1} t_{k+n\nu}^{2q}}}_{=:d_k}.$$

The quotient d_k/b_k satisfies

$$\frac{d_k}{b_k} = \frac{t_k^{2q+2r} + \sum_{\nu=1}^{\tau-1} t_{k+n\nu}^{2q+2r}}{\left(t_k^{2q} + \sum_{\nu=1}^{\tau-1} t_{k+n\nu}^{2q}\right)\left(t_k^{2q} - \sum_{v=1}^{\tau-1} t_{k+n\nu}^{2q}\right)} = \frac{t_k^{2q+2r} + \sum_{\nu=1}^{\tau-1} t_{k+n\nu}^{2q+2r}}{t_k^{4q} - \left(\sum_{v=1}^{\tau-1} t_{k+n\nu}^{2q}\right)^2}.$$

If $q \geq r$ then $d_k/b_k > 1$. Hence, if $\tau > 1$, then $\mathrm{risk}_q(p = D) < \mathrm{risk}_q(p = n)$. In other words, the lowest risk in the over- regime is strictly less than that in the underparameterized regime. ■

5. Experiments.

Discrete Fourier Models. In this experiment, we use Fourier series models $\boldsymbol{F} \in \mathbb{C}^{n\times D}, D = 1024, n = 64$ with r-decaying multivariate Gaussian coefficients ($r = 0.3, 0.5, 1.0$). $\boldsymbol{F}_T \in \mathbb{C}^{n\times p}$ is the observation matrix with $p < n$ in underparameterized regime; and $p = ln, l = 1, 2, \ldots, \tau$, in the overparameterized regime. The weighted min-norm estimator uses Σ^q, $q \geq 0$ to define the weighted ℓ_2-norm. The theoretical curves are the risks calculated according to Theorem 3.1 and 4.1. The empirical mean curves and the 80% confidence intervals (CI) are estimated by 100 runs of independently sampled feature vectors $\boldsymbol{\theta}$.

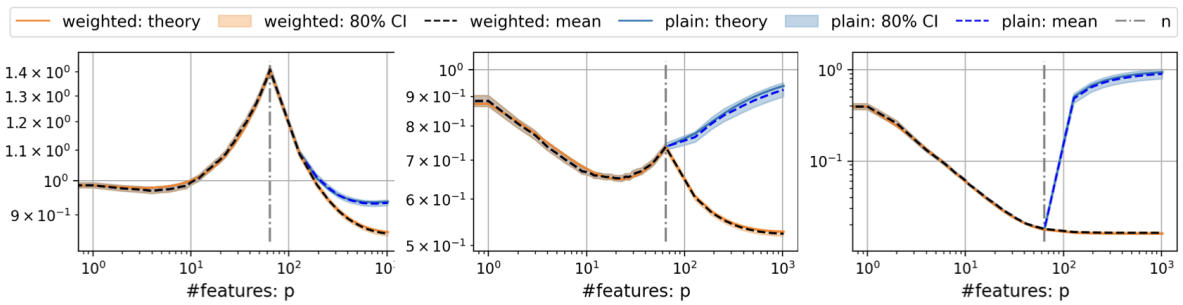

Figure 2. *Theoretical and empirical risks ($\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^2$) of plain and weighted min-norm estimators in log-log scale. Left to right: $r = q = 0.3, 0.5, 1.0$.*

Figure 2 shows that the empirical mean risks match the theoretical risks $\mathbb{E}[\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^2]$ of Theorems 3.1 and 4.1 very accurately. Figure 2 validates that weighted optimization results in better generalization in the overparameterized regime (Theorem 4.4), and shows non-degenerated double descent curves when $r = q = 0.5$.

Function Interpolation. In this experiment, we interpolate the functions at n equispaced points x_j on $[-1, 1]$. The observed n samples are $(x_j, y_j)_{j=1}^n$, where $y_j = f(x_j)$ with Fourier series $f(x) = \sum_{k=-\infty}^{\infty} \theta_k \exp(\pi i k x)$. We fit training samples to a hypothesis class of p-truncated Fourier series: $f_{\hat{\boldsymbol{\theta}}}(x) = \sum_{k=-m}^{m} \hat{\theta}_k \exp(\pi i k x)$ with $p = 2m + 1$ via least squares in underparameterized case; via plain and weighted min-norm estimator in overparameterized case with $\boldsymbol{F}\boldsymbol{\theta} = \mathbf{y}$, where $\boldsymbol{F} \in \mathbb{C}^{n\times p}$ with $F_{j,k} = e^{2\pi i j k/n}$. In the experiment, we use $D = 1000$, $n = 15$, $q = 1.5$ for $f_1(x)$ and $q = 2$ for $f_2(x)$.

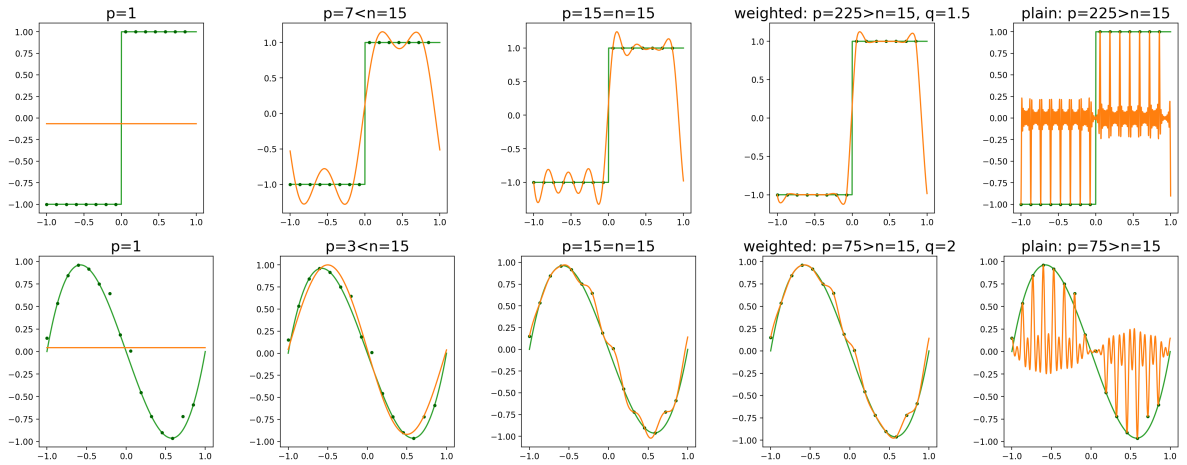

Figure 3. *Interpolation (in orange) of stage function and smooth function. Up:* $f_1(x) = 1, \forall x \in [-1, 0); f_1(x) = 1, \forall x \in [0, 1]$. *Down:* $f_2(x) = 2.5(x^3 - x)$ *with noise. From left to right: least square with* $p = 1$*; least square with* $p < n$*;* $p = n$*; interpolation by weighted min-norm estimator* $(p > n)$*; interpolation by plain min-norm estimator* $(p > n)$.

Figure 3 presents the interpolation (in orange) using different estimators with the same set of equispaced samples (dark green). The overparameterization with the plain min-norm estimator is useless, while the weighted min-norm estimator has the best performance in both noiseless and noise cases. It also shows the benefit of the weighted optimization's regularization towards smoother interpolants.

6. Discussion on Necessity of Randomness. In order to illustrate the necessity of randomness, we provide an example of a nontrivial function f which has r-decaying Fourier coefficients and vanishes at all the sample points x_j in Figure 4. According to the algorithm, the estimation of those signals will be identically zero, which is clearly incorrect. The example is generated numerically by applying gradient descent to the loss function $L(\boldsymbol{\phi}) := \sum_{j=0}^{n-1} \left| \sum_{k=0}^{D-1} (k+1)^{-r} e^{i\phi_k} e^{\frac{2\pi ijk}{n}} \right|^2$. Once we find $\boldsymbol{\phi}^*$ such that $L(\boldsymbol{\phi}^*) = 0$, we set $f(x) = \sum_{k=0}^{D-1} (k+1)^{-r} e^{i\phi_k} e^{ikx}$, which vanished on all x_j by construction.

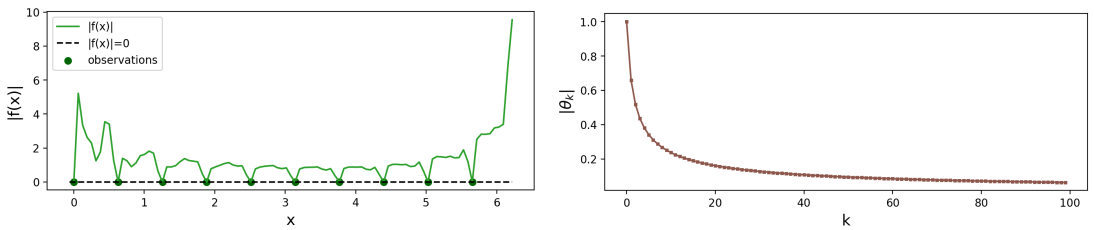

Figure 4. *In this example we see that requiring* θ_k *decaying at rate* r *is insufficient, because it is possible that* f *vanishes on all the sample points, and hence impossible to recover. Here* $D = 100$, $n = 10$, $r = 0.6$.

Besides the numerical experiment, we also prove the existence of such counterexamples. For $n = 1$, finding $\boldsymbol{\phi}^* s.t.\ L(\boldsymbol{\phi}^*) = 0$ is reduced to constructing a D-polygon in $\mathbb{R}^2$ (equivalent to $\mathbb{C}$) with edges of length $(k+1)^{-r}$ for $k \in [D]$. For $r \leq 1$, since any edge is shorter than the sum of other edges, i.e., $(k'+1)^{-r} < \sum_{k \in [D] \setminus \{k'\}} (k+1)^{-r}, \forall k' \in [D]$, such polygon always exists, which can be proved by induction and triangle inequality. Thus, there exists a

nontrivial function which has r-decaying Fourier coefficients and vanishes at the origin (x_0).

7. Conclusion and Outlook. This paper answers an open question on how and when the weighted minimal ℓ_2 norm trigonometric interpolation achieves low generalization error in the overparameterized regime. From our non-asymptotic expressions for the risk, we quantify how the bias towards smooth interpolations can be exploited to reduce the risk in the overparameterized setting and show that this risk is strictly better than the lowest possible risk in the underparameterized regime under certain conditions. In this way, our work also contributes to the understanding of the "double descent" curve. Extending our theoretical results to general bounded orthonormal systems and neural networks is an exciting direction for future research. It is also interesting to choose random sampling points instead of equidistant sampling points. First numerical experiments show similar behavior except around the pole $p = n$, where the risk for random samples blows up. The theoretical investigation of the generalization error in such context is left for future work.

Acknowledgments. R. Ward and Y. Xie were supported in part by AFOSR 2018 MURI Award "Verifiable, Control-Oriented Learning On The Fly". H.H Chou and H. Rauhut were supported in part by the DAAD grant 57417829 "Understanding stochastic gradient descent in deep learning" and by the Excellence Initiative of the German federal and state governments.

Appendix A. Visualization of Theoretical Risks.

A.1. Heat Maps of Theoretical Risks. We show the heat maps of the theoretical risks of weighted and plain min-norm estimators in Figure 5, which are calculated by Theorem 3.1 and 4.1. Here, we use Fourier series models with $D = 1024$, varying n, and $q = r$. The x-axis is r of the r-decaying coefficients (from 0 to 2 with 0.1 as the step), the y-axis is p (where $p < n$ in the underparameterized regime and $p = ln, l \in \mathbb{N}_+$ in the overparameterized regime), and the risks are in log scale. We can see the trends of the risks: the top three plots show that when $q = r > 1$ the risk monotonically decreases as p increases in the underparameterized regime and the lowest risk lies in the overparameterized regime; while the bottom three plots show that after $p > n$, the risks of plain min-norm estimator ($q = 0$) increase suddenly and they are higher (i.e., the light color block in each heat map) than the risks in then underparameterized regime when $r > 1$. Hence, these plots also verify Theorem 4.4.

A.2. Theoretical Risks with varying r and q. Figure 6 shows the plots of the theoretical extended risk curves (fixed n) with a range of choices for r and q as mentioned in Remark 4.5, from which we can see the trends and patterns of the risks. In this experiment, we use Fourier series models with $D = 1024$, $n = 8, 128$ to 1024, $p < n$ in the underparameterized regime and $p = ln, l \in \mathbb{N}_+$ in the overparameterized regime. We investigate on $r = 0, 0.3, 0.5, 0.7, 1.0, 1.5, 2.0, 3.0$ and p with the same range but not necessarily equal to r. The curves with $q = 0$ correspond to the risks of the plain min-norm estimator. Some observations of the plots are as follows.

1. For varying n, the trends with the same r and q are similar along with different transition points ($p = n$), except for the case $r = 0$ and $n \geq D/2$ (as it states in Remark 4.2, when $n < D/2$, the risk increases with p in the underparameterized regime while for $n \geq D/2$ it goes to the other direction.)

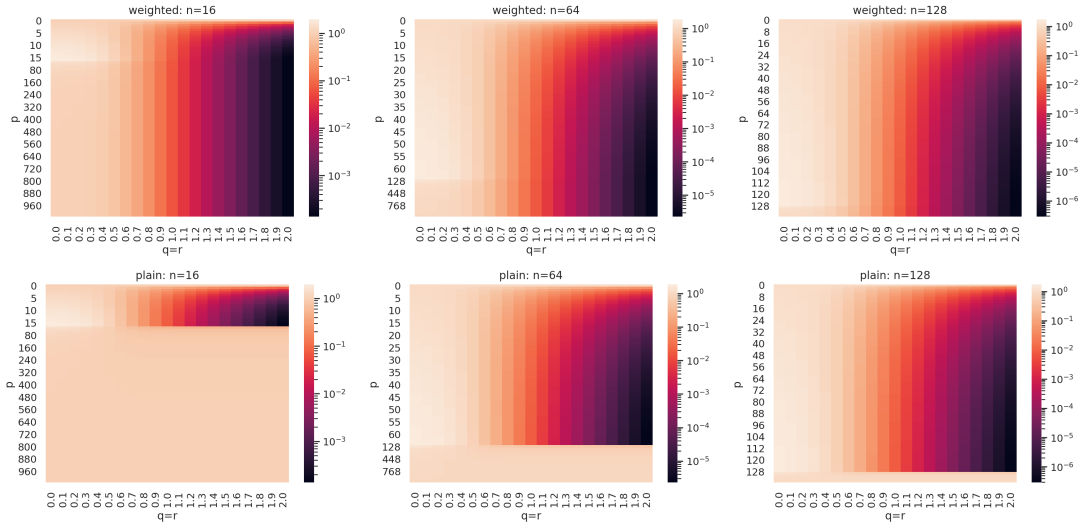

Figure 5. *Heat maps of theoretical risks of weighted (up) and plain min-norm (down) estimators of r-decaying features. $D = 1024, q = r$, and $n = 16, 64, 128$. (Note the these heat maps on the right are not corrupted: there are light color blocks since the risks of the plain min-norm estimators (i.e., $q = 0$ in Figure 6) changes to around 1 after $p > n$, and the color bar is in log scale. This transition also occurs with the weighted estimator, where the faint horizontal line takes place ($p = n$). It corresponds to a peak in risk, as in Figure 6.)*

2. In the underparameterized regime, when $r = 0$ and $n < D/2$, the risk increases with p until $p = n$. The phase transition from $r = 0$ to $r \geq 1$ validates Theorem 4.4.
3. In the overparameterized regime, the risk of the plain min-norm estimator is almost above the weighted min-norm estimator when $r \geq 0.5$. Even if the weight matrix does not match the covariance matrix exactly for r-decaying coefficients, the weighted min-norm estimator usually achieves lower risks than the plain min-norm estimator.
4. As stated in the proof of Theorem 4.4, the plots also show that when $q \geq r$, the risk at $p = D$ is strictly lower than that at $p = n$, and $r \geq 1$ is a sufficient condition to assure the monotonic decrease when $p < n$ and the lowest risk in the over- is strictly lower than that in the under-parameterized regime.

REFERENCES

[1] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, *Benign overfitting in linear regression*, Proceedings of the National Academy of Sciences, (2020).

[2] M. Belkin, D. 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 (2019), pp. 15849–15854.

[3] M. Belkin, D. Hsu, and J. Xu, *Two models of double descent for weak features*, arXiv preprint arXiv:1903.07571, (2019).

[4] M. Belkin, A. Rakhlin, and A. B. Tsybakov, *Does data interpolation contradict statistical optimality?*, arXiv preprint arXiv:1806.09471, (2018).

[5] A. Canziani, A. Paszke, and E. Culurciello, *An analysis of deep neural network models for practical applications*, arXiv preprint arXiv:1605.07678, (2016).

[6] M. Geiger, A. Jacot, S. Spigler, F. Gabriel, L. Sagun, S. d'Ascoli, G. Biroli, C. Hongler, and M. Wyart, *Scaling description of generalization with number of parameters in deep learning*, Journal of Statistical Mechanics: Theory and Experiment, 2020 (2020), p. 023401.

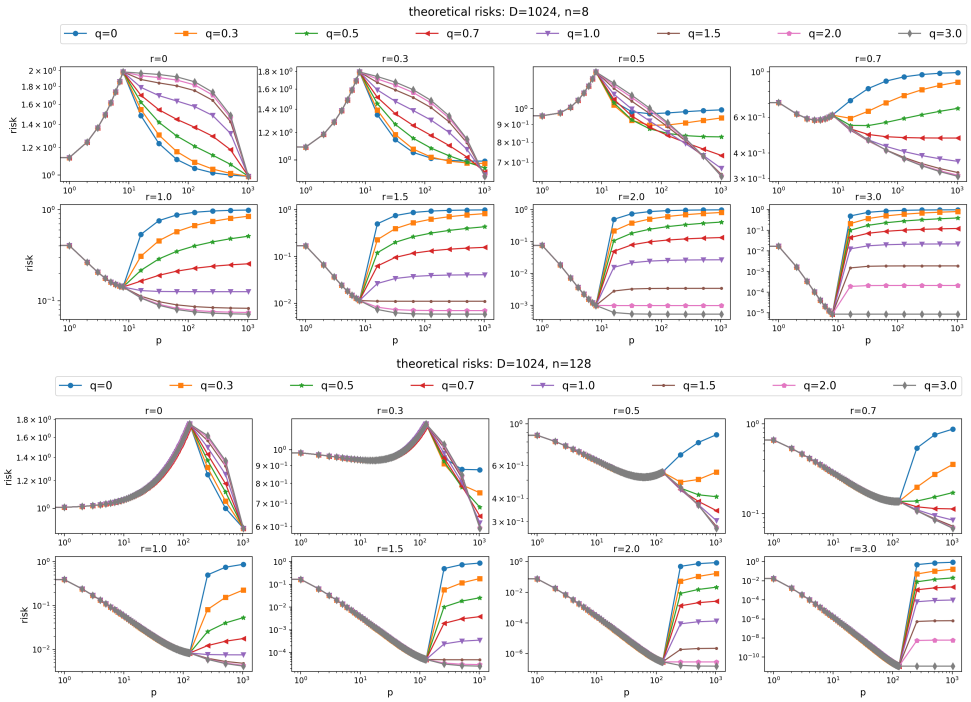

Figure 6. *Theoretical risks of r-decaying features and varying q with $D = 1024$, up: $n = 8$, down: $n = 128$.*

[7] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari, *Linearized two-layers neural networks in high dimension*, arXiv preprint arXiv:1904.12191, (2019).

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

[9] K. He, X. Zhang, S. Ren, and J. Sun, *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification*, in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026–1034.

[10] W. Li, *Generalization error of minimum weighted norm and kernel interpolation*, arXiv preprint arXiv:2008.03365, (2020).

[11] T. Liang and A. Rakhlin, *Just interpolate: Kernel" ridgeless" regression can generalize*, arXiv preprint arXiv:1808.00387, (2018).

[12] S. Mei and A. Montanari, *The generalization error of random features regression: Precise asymptotics and double descent curve*, arXiv preprint arXiv:1908.05355, (2019).

[13] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, *Deep double descent: Where bigger models and more data hurt*, arXiv preprint arXiv:1912.02292, (2019).

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

[15] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, arXiv preprint arXiv:1409.1556, (2014).

[16] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, *Understanding deep learning requires rethinking generalization*, arXiv preprint arXiv:1611.03530, (2016).